An Approach to Generate Association Rules in Assisted Medical Diagnosis Using Bayesian Network and Three-Way Decisions (BNTWD)

A Thesis
Presented to the
Department of Computer Science
Institute of Information and Computing Sciences
University of Santo Tomas

In Partial Fulfilment
of the Requirements for the Degree
Bachelor of Science in Computer Science

By:
Bitera, Clyde Ravi, R.
Crisostomo, Brian Paul, V.
Teodoro, Jamil Kristian, V.
Tumulak, Rachel Monique, K.

Adviser:
Catubag, Joseph Richard, G.

Approval Sheet

Thesis Title: An Approach to Generate Association Rules in Assisted Medical Diagnosis Using Bayesian Network and Three-Way Decisions

Researchers

1. **Clyde Ravi R. Bitera**
2. **Brian Paul V. Crisostomo**
3. **Jamil Kristian V. Teodoro**
4. **Rachel Monique K. Tumulak**

In partial fulfillment of the requirements for the degree of **Bachelor of Science in Computer Science,** the thesis mentioned above, has been adequately prepared and submitted by above mentioned proponents. This thesis was duly defended in an oral examination before a duly constituted tribunal with a grade of

Ms. Ria A. Sagum
Panel Member

Asst. Prof. Jonathan B. Cabero
Panel Member

Asst. Prof. Donata D. Acula
Panel Member

Asst. Prof. Cecil Jose A. Delfinado
Thesis Coordinator

Mr. Richard Joseph G. Catubag
Thesis Adviser

Accepted in partial fulfillment of the requirements for the degree of Bachelor of Science in Computer Science.

Assoc. Prof. Jose L. Seño
Chairperson
Department of Computer Science
Institute of Information and Computing Sciences

University of Santo Tomas
Institute of Information and Computing Sciences
Department of Information Computer Studies

Certificate of Authenticity and Originality

We, the authors of this thesis, "A Proposed Enhancement of Association and Classification Algorithm And Applied to Market Basket and Assisted Medical Diagnosis", hereby certify and vouch that the contents of this research work is solely our own original work; that no part of this work has been copied nor taken without due permission or proper acknowledgment and citation of the respective authors; that we are upholding academic professionalism by integrating intellectual property rights laws in research and projects as requirements of our program.

If found and proven that there is an attempt or committed an infringement of copyright ownership, we are liable for any legal course of action sanctioned by the University and the Philippine laws.

Clyde Ravi R. Bitera — **Brian Paul V. Crisostomo**

Jamil Kristian V. Teodoro — **Rachel Monique K. Tumulak**

Program: BS Computer Science

Date: January 10 2018

ACKNOWLEDGEMENTS

The researchers would like to send their deepest gratitude to the following people for their contributions that lead to the success of this study.

To Almighty God for the wisdom and fortitude throughout the study. To our families and friends for their support and love which held us together.

To Mr. Joseph Richard Catubag, for guiding us and sharing with us his knowledge needed for this study to be even possible. For his time and patience during our consultation hours. Also for keeping the team together at its lowest.

To Asst. Prof. Cecil Jose Delfinado, for his weekly efforts on reminding us what to pass and when to pass the needed requirements. Also in doing his best for the betterment of the research paper and his patience as well.

Table of Contents

List of Figures

List of Tables

ABSTRACT

Bayesian network and Three way decisions is an algorithm that aims to create a new approach of generating association rules in assisted medical diagnosis. The main objective of the study is to provide a way to determine frequent combinations and generate association rules of diagnosis. Results were compared to Apriori through three doctors output. The researchers were able to reach the objective in applying the algorithm to assisted medical diagnosis in providing the possible frequent combinations illness.

Chapter 1 The Problem and Its Background

A. Introduction

One of the recent practices of medium and large-scale companies involves the use of data analysis to obtain useful information from collected data. This involves the collection and storage of huge amounts of data, collected throughout the span of years. However, according to Manpreet Kaur and Shivani Kang (2016), not all information is useful to users when various techniques or methods of data extraction or *data mining* processes are applied. Examples of data mining techniques include the following: association, classification, prediction, clustering, outlier analysis.

Association Analysis is the task of uncovering correlation relationships among large set of data (Raghu, et. al., 2012). It is used in market basket data, bioinformatics, medical diagnosis, Web mining and scientific data analysis. Market basket analysis is used to determine purchasing behavior of customers, a factor that is considered in cross selling programs (Tan, et. al., 2005). The Apriori *algorithm* and weka tool are some examples of the usual tools used for association analysis. (Kaur and Kang. 2016).

In Nikam's (2015) comparative study, he defined *classification* as a data mining technique that is applied when one is trying to find out in which group an item belongs or is related. The item is then distributed in different classes according to the constraint given and the technique follows three approaches namely Statistical, Machine Learning, and Neural Network for classification.

Since data mining is very helpful to businesses, the market basket analysis, also known as association rule mining, is used by companies to mimic the performance of the

market. This is often referred as how the market combines the products or services to increase sales (Investopedia Staff , 2007).

According to Rajendra Akerkar and Priti Srinivas Sajja (2016), the most accepted algorithm for finding frequent sets is the *Apriori* algorithm. It uses the downward and closure property; moreover, the algorithm implements a bottom-up search. This algorithm prunes as many sets that are unlikely to be frequent as it can before it reads the database at every level.

One of the limitations of the Apriori Algorithm is that it takes time to run. A phenomenon usually observed when data sets becomes too large since the best candidate is often found in the *bottleneck*. Considerations are needed when computing because of the requirement to scan every level or n+1 scans, where n is the length of longest pattern. Therefore, improving the algorithm efficiency incorporates methods using transaction reduction and/or sampling: mining of a subset of data or determining *completeness* (Investopedia Staff , 2007).

According to Raghu, et. al., (2012), "Association rules of events/nodes can be regarded as probability rules due to their co-occurrence". *Cross-selling programs* are one of the real-life examples that implement the idea. Raghu, et. al. (2012), used Bayesian Network(BN) to developed an automated data mining technique, focusing on the association rules using Associated Rules Binary Symmetric Matrix using K2 (ARBSM-K2) to generate Hierarchical Association Rule (HAR). In doing so, Raghu, et. al.(2012), proved that the algorithm has improved in comparison to the Apriori algorithm specifically on the number and size of patterns identified, and the searching time reduced. On the other hand, Jia, et. al. (2016), enhanced the traditional method of

classifying - two way decisions to three way decisions that is able to reduce the minimum cost attribute. For that reason, the proponents propose to apply Hierarchical Association Rule through a Bayesian Network approach of association and Jia's, et. al. (2016), approach of classification to resolve on the issue of transaction reduction and sampling of the Apriori algorithm.

B. Background of the Study

According to Mark Casson and John S. Lee (2011); "The market has played a critical role in many of the economic 'revolutions' that have been identified by economic historians, from the credit revolution of the fourteenth century, through the commercial revolution and the agricultural revolution of the seventeenth and eighteenth centuries, down to the Industrial Revolution and beyond." This slow evolving revolution eventually led to what markets are today: Supermarkets, wet and dry markets and even online.

Eventually, certain researchers by the names of Rakesh Agrawal and Ramakrishnan Srikant proposed the Apriori Algorithm in 1994. It is a rule based, association method where it identifies items to a larger set of items that often appear in the database. But that requires large amount of space to hold large amount of associations since the computation requires to scan up to the at least the length of the longest pattern made. With that, certain modules require for apriori algorithm to be enhanced to improve efficiency such as faster tree creation since the creation of the tree uses up much needed time, or improving the process of searching a tree by implementing certain algorithms like Bayesian Networks. (Han et. al., 2012)

Speaking of which BN is a model that uses probabilistic approach and can provide a smooth, consistent and flexible decision making more applicable in complex domains while association rule relies on confidence percentage value to prove effectiveness (Raghu, et al., 2012). Therefore using a different approach in traversing the tree, to produce output/s.

Two-way decisions focus on acceptance and rejection of an information. Hence, no supporting of insufficient evidence. An addition of deference property on classification in three-way decisions allow further examination. Where in it uses decision threshold and conditional probability. Decision threshold was based on the game theory from Jia, et al. researches - Herbert and Yao proposed approach on governing the modification of cost functions in order to adjust decision thresholds. As for the conditional probability, according to Jia et. al. (2016) most researches apply equivalence classes to describe objects in rough set theory however applying naive Bayesian theory alone to rough set theory that was done by Yao and Zhou mentioned by Jia et. al. the conditional probability was not well calibrated because naive Bayes apply an independence assumption between attributes.

The idea of applying algorithms in the medical field, particularly in the field of medical diagnosis, has been around for awhile. One of methods involves the utilization of Data mining to determine patterns in patient data. A study conducted by Awodun, M., & Adedara, R. (2017) utilizes the Apriori Algorithm, a popular association rule mining algorithm, in determining the most likely symptoms of infectious diseases such as malaria, influenza, typhoid etc. based on collected patient records. In doing so, Medical

professionals would have a reference to giving a diagnosis to future patients, preventing misdiagnosis. However, the nature of the system would only be applicable as a point of reference to the medical professional and would only be used to compare the symptoms exhibited by the patient based on the given diagnosis.

Computer-assisted Medical Diagnosis or Medical Diagnosis Decision Support (MDDS) tackles the idea of a computer being able to perform a medical diagnosis. Using Bayesian Networks, Zagorecki, A., Orzechowski, P., and Hołownia, K. (2013) were able to create a general medical diagnostic tool that is intended for self-diagnosis of what the user may be experiencing or suffering from. The diagnosis would start by asking the user his/her measurements, sex, and age alongside the initial symptom and parts where they would feel discomfort. The system would then generate an initial list of probable diseases then prepare the next question to be asked to the user. A score based on the value of information and cross-entropy is used and the most informative observation based on this score is determined. The system will repeat the process until the probability of a the most likely disease exceeds 70%. Depending on the diagnosis, the list will include 1 to 3 probable diseases you would likely have.

C. Theoretical Framework

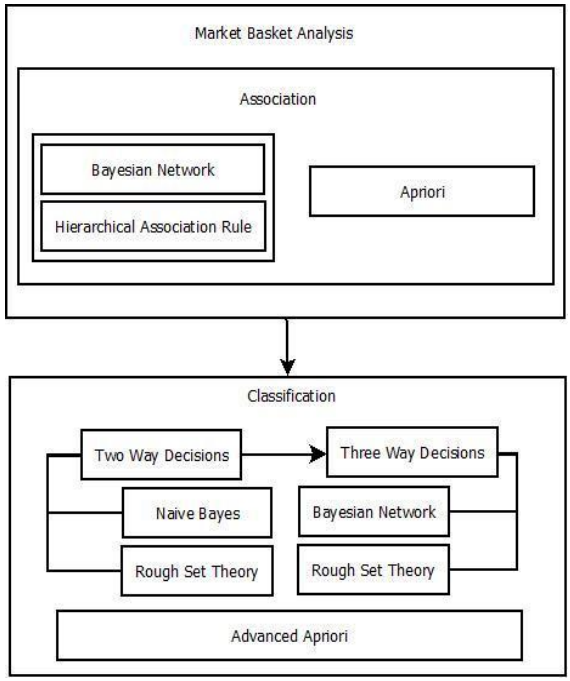


Figure I.C-1: Theoretical Framework of the Study

Association Analysis is the task of uncovering correlation relationships among large set of data,(Raghu, et al., 2012) it uses a set of transactions to discover rules that indicate the likely occurrence of an item based on the occurrences of other items in the transaction. Classification is a data mining technique that is applied when one is trying to find out in which group it belongs or is related. It is distributed in different classes according to the constraint given. (Nikam's 2015) Bayesian Networks is also known as belief networks that belongs under the category of probabilistic *graphical models* (GMs) more specifically as a *directed acyclic graph* (DAG). It is a combined principles from graph theory, probabilistic theory, computer science and statistics. Each node in the graph is represented by variables and the edge connecting the nodes represent the probabilistic dependencies that is computed by estimating the known statistical and

computational method. Which enable effective representation and computation of *joint probability distribution* over the set of random variables. (F, R.,2007)

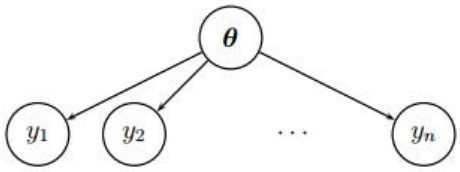


Figure I.C-2: Non-hierarchical model from Probabilistic Models in the Study of Language by Levy (2012)

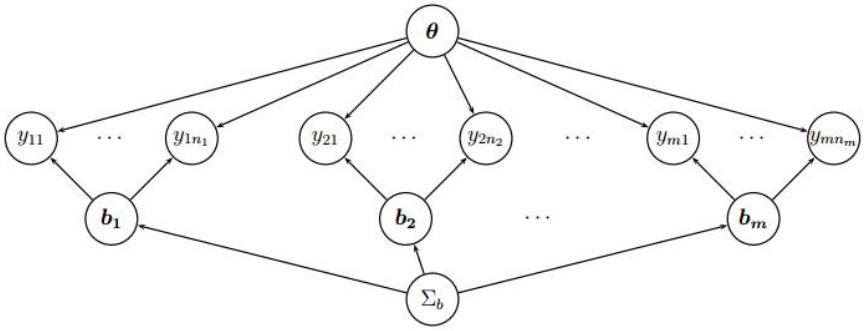


Figure I.C-3: Hierarchical model from Probabilistic Models in the Study of Language by Levy (2012)

Hierarchical model is demonstrated in Figure I.C-2 illustrate a model family has parameters θ, which determine a probability distribution over outcomes, and as a collection of independent draws from this distribution a set of observations y arises. On the other hand Figure I.C-3 depicts a straightforward type of hierarchical model where scrutiny fall into a number of clusters and the distribution over outcomes is resolute jointly by

a) Parameters θ shared across the clusters, and

b) Parameters b which are shared among the scrutiny within a cluster, maybe different across clusters.

Pressingly, parameter Σb is the second probability distribution over the cluster-specific parameters b.(Levy, 2012)

Classification is from the word itself, it classifies models into classes. If the task of the data analysis is to classify the models, a classifier is constructed to predict class or categorical labels such as for medical data “treatment A”, “treatment B”, or “treatment C”; “safe” or “risky” for loan application data, and so on. These categorical labels can be represented by discrete values, where ordering is no importance. (Han et. al., 2012)

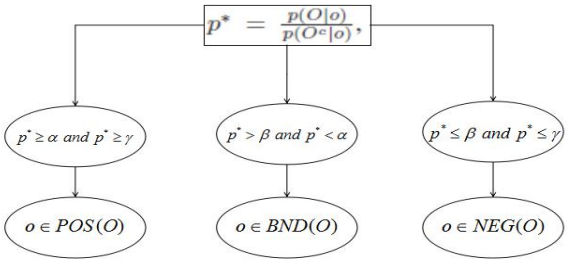


Figure I.C-4: Xiuyi Jia et. al. Function based three-way decisions based Bayesian network 2016

Three-way decision theory is an extension of the two-way decision. In two-way decisions methods usually by comparing the decision threshold α and the conditional probability p(O|o). Moreover, it is based on the combination of Bayesian decision theory and *rough set theory*, Ω = {O,O^c} is the set of two states indicating that an object is in category O or not in category O, respectively. Whenever an information is either acceptance nor rejection, three-way gives a third support which is allowing the information to be further examined. As for solving the α and β Bayesian Decision procedure provides a good rough set model that adapts a decision threshold learning method. As to estimate the conditional probability, Bayesian Network was applied to extract the exact value of the conditional probability. Jia et. al. aimed to minimize cost

attribute reduct, the purpose of this is to find minimal subset of attributes that satisfies the criteria compared if it will use the entire set of attributes. (Jia, et al., 2016)

A rough set theory is the approximation of lower and upper boundary of a set. The lower approximation is said to be the part where the objects are definitely the interest subset. As for the upper approximation, it is the part that where objects are probably the interest subset. Over the years representing uncertain knowledge availability of information with respect to consistency and presence of data patterns are some functions of rough data set (Rissino and Lambert-Torres, 2009).

D. Conceptual Framework

a.) Existing Framework

The study is to propose an enhancement in association and classification by applying it to market basket analysis and assisted medical diagnosis through the implementing a combination of two Bayesian concept algorithm in formating the association rule and in Three-Way Decisions. Allowing multiple algorithms to work side by side to each other may enhance the capability of one another. To prove the accuracy, performance of it was compared to Apriori.

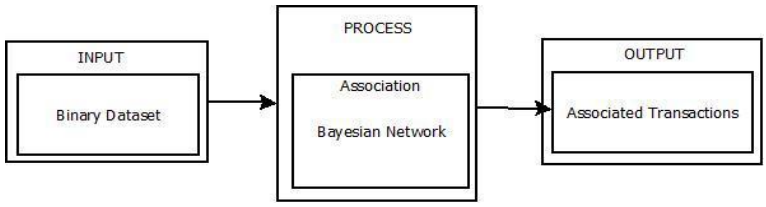


Figure I.D-1: Raghu et al conceptual framework (2012)

Antecedent and consequent are the two basis of an association rule which have condition attributes and decision attributes respectively. Since each association rule uses confidence percentage value that shows the effectiveness of a rule in terms of probability,

the BN and association rule can be joined together in a probabilistic approach (Raghy et. al., 2012).

The dataset used by Raghu et. al. was first converted to binary dataset using their confidence percentage values file as dataset. Raghu and his proponents suggested the use of Association Rules Binary Symmetric Matrix using K2 (ARBSM-K2) technique for the generation of Hierarchical Association Rules (HAR). After maximizing the probability for each association rule which was represented as a node; the hierarchy showed more certainty of association rules and it becomes evident with the number and size of patterns identified and the searching time reduced

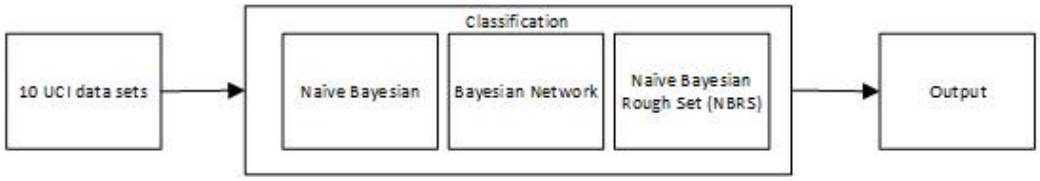


Figure I.D-2: Xiuyi Jia, Huaxiong Li, Lin Shang's Experimental Setting (2016)

Figure I.D-2 shows Xiuyi Jia, Huaxiong Li, Lin Shang's Experimental Setting which used 10 UCI data sets to be classified in three ways namely Naive Bayesian, Bayesian Network and Naive Bayesian Rough Set (NBRS). For each data sets a 10 different group of cost functions were randomly generated. The values are (0,1) with the following constraint conditions - $0 = \lambda PP \leq \lambda BP \leq \lambda NP$ and $0 = \lambda NN \leq \lambda BN \leq \lambda PN$. For the comparison criteria of the length of reduction both the accuracy and miscalculation cost were used. The misclassification cost is defined as:

$$TC = nPN \cdot \lambda PN + nBN \cdot \lambda BN + nBP \cdot \lambda BP + nNP \cdot \lambda NP, \quad (1)$$

where "nPN means the number of objects which are classified into the positive region when it belongs to the negative region, nBN means the number of objects which are classified into the boundary region when it belongs to the negative region, nBP means the

number of objects which are classified into the boundary region when it belongs to the positive region, and nNP means the number of objects which are classified into the negative region when it belongs to the positive region, respectively." This classification schema can be transformed into an optimization problem, the minimum cost attribute can be seen as a process of finding a minimal attribute set which induce minimum cost (Jia, et al., 2016).

b.) Proposed Framework

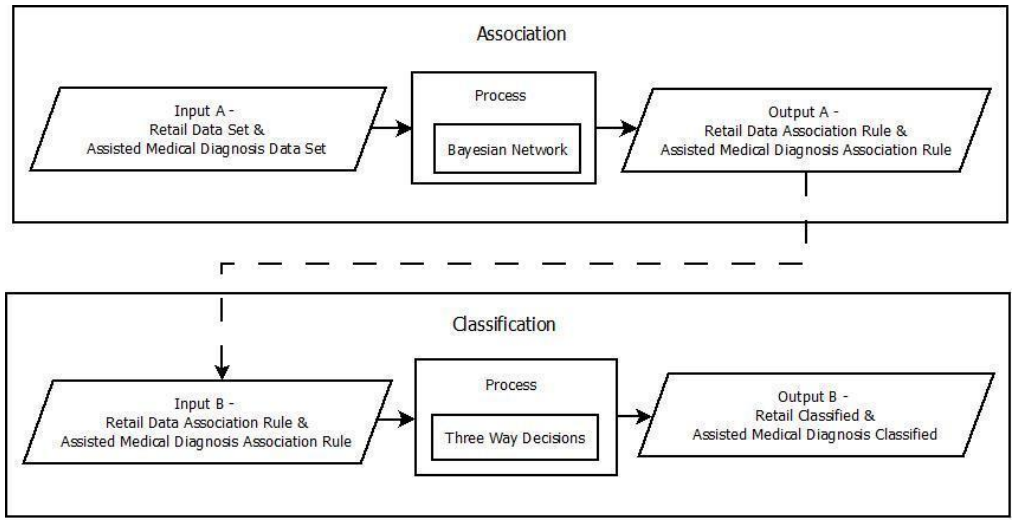


Figure I.D-3: Proposed Conceptual Framework of the Study

Figure I.D-3 shows the proposed conceptual framework of the proponents. The proponents have pre-processed the dataset. Cleaning the data in a sense where only the attributes that would be needed was retained. The proponents proposed to create an *associative classification* by integrate association and classification with the offer to adopt the algorithms used by Raghu, et al. (2012) to be used for association and Three-Way decisions based on Bayesian Network by Jia, et al. (2016) to optimize classification.

The input of the system was the pre-processed data applicable for associating. For the processing phase, the dataset was first associated to the bought

together products to produce the association DAG as for the assisted medical diagnosis association, each symptoms was associated to other symptoms if that symptoms goes with it. Afterwhich the association DAG was classified for market industry: how often or frequent the transactions were based on the confidence and support of the association rule for medical industry.

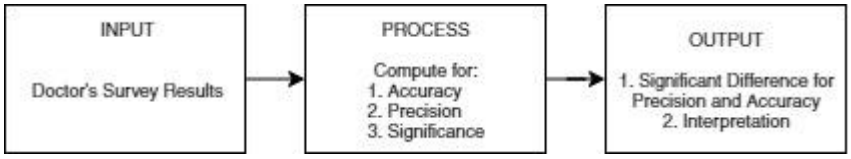


Figure I.D-4: Conceptual Framework for the Analysis

Along with the process of the proposed algorithm, Apriori ran using RapidMiner tool using the same dataset or input used in the proposed algorithm after which the association rules that was output by the Apriori. The test results were compared in terms of accuracy and precision. (Computations shown in Chapter III)

E. Statement of the Problem

The Apriori algorithm was created to determine associations between products within transactions to better understand which groups of products would be suited to be placed together. It can also be used to determine the causality percentage of symptoms, given available medical data. Although the algorithm wasn't intended for such an application, results are yet to be feasible. The study specifically addressed the following problems:

1. Will the proposed algorithm generate the correct association rules used in the assisted medical diagnosis?
2. Will the proposed algorithm provide correct combination of diagnosis and frequency to that of the Apriori algorithm?

3. Will the proposed algorithm be precise to that of the Apriori algorithm?

F. Objectives

The main objective of this study is to determine if the combined Bayesian Network Hierarchical Association Analysis and Three-way Decisions algorithm would provide a reliable, alternative method to perform Association and Classification as applied to assisted medical diagnosis.

The specific objectives of the study are the following:

1. To test if the proposed algorithm would generate the correct association rules used in the assisted medical diagnosis.
2. To test if the proposed algorithm would have more correct combination of diagnosis and frequency than the Apriori algorithm.
3. To test if the proposed algorithm would be more precise than the Apriori algorithm.

G. Scope and Limitations

The study focuses on the checking the feasibility of Bayesian Network in generating association rules in medical diagnosis. Note that the study is limited to the following:

1. The study focuses on the two algorithms that are based on Bayesian Networks which are Hierarchical Association Rule and Three-way decisions.
2. The study will be limited to the opinions of the doctors who will participate in the survey.

3. The study will focus on the infectious and parasitic diseases found in the dataset.
4. This study is specific to feasibility of generating association rules and frequent combinations.
5. The algorithm will only be compared to the Apriori algorithm.

H. Significance of the Study

This will benefit the following groups/industries:

1. Professionals working in the medical field through assisted diagnosis
    a) To help associate symptoms with other symptoms.
    b) To give guidance in the prognosis of the patient.

I. Definition of Terms

Algorithm. A procedure or formula in solving a problem.

Apriori. An algorithm for frequent item set mining and **association rule** learning over transactional databases. The algorithm the proponents is trying to match with the proposed algorithm. The most influential algorithm for association rules. It searches items by its frequency or how many times it shows up in the dataset.

Assisted Diagnosis. Systems that aid doctors in their diagnosis. One of the applications of our proposed algorithm.

Association rule. Is a model that identifies how data items are associated with one another. It associates different entities/transactions/item with other items it may be related to. This includes the Apriori algorithm.

Associative classification (AC). incorporates classification and association rule mining to model construction. The model the proponents are proposing.

Bayesian network, **Bayes network**, **belief network**. model or probabilistic directed acyclic graphical model is a probabilistic graphical model (a type of statistical model) that represents a set of random variables and their conditional dependencies via a directed acyclic graph (DAG). This is the alternative solution the proponents intends to use. An algorithm that works well with dataset that is large.

Bottleneck. a point of congestion or blockage, in particular. The main problem of Apriori, it bottlenecks when dataset is too large.

Classification. It is a data mining technique use to classify data in different classes. This is use as another data mining technique of the proponents.

Completeness. Refers to an indication of whether or not all the data necessary to meet the current and future business information demand are available in the data resource.

Cross-selling. Is the action or practice of selling an additional product or service to an existing customer. In practice, businesses define cross-selling in many different ways. Elements that might influence the definition might include the size of the business, the industry sector it operates within and the financial motivations of those required to define the term.

Data mining. is the analysis step of the "knowledge discovery in databases" process, or KDD. The term is a misnomer, because the goal is the extraction of patterns and knowledge from large amounts of **data**, not the extraction (**mining**) of **data** itself.

Deference. (also called **submission** or **passivity**) is the condition of submitting to the espoused, legitimate influence of one's superior or superiors.

Directed acyclic graph (**DAG**). is a finite directed graph with no directed cycles. That is, it consists of finitely many vertices and edges, with each edge directed from one vertex to another, such that there is no way to start at any vertex *v* and follow a consistently-directed sequence of edges that eventually loops back to *v* again.

Graphical model or probabilistic **graphical model** (PGM). is a probabilistic **model** for which a graph expresses the conditional dependence structure between random variables. They are commonly used in probability theory, statistics—particularly Bayesian statistics—and machine learning. In the study of probability, given at least two random variables *X*, *Y*, that are defined on a probability space, the **Joint probability distribution** for *X*, *Y*, is a probability distribution that gives the probability that each of *X*, *Y*, ... falls in any particular range or discrete set of values specified for that variable

Hierarchical Model is a data model in which the data is organized into a tree-like structure. The data is stored as records which are connected to one another through links. A record is a collection of fields, with each field containing only one value.

Market Basket Analysis. a type of data analysis used in marketing and retailing that determines what items are being purchased together. One of the areas of study that the proponents aim to improve.

Normal Form. a defined standard structure for relational databases in which a relation may not be nested within another relation. The data pre-processed to be used by the proponents.

Prognosis. the likely course of a disease or ailment. A part of the medical process the study is tackling.

Rough set theory. is one of the important methods for knowledge discovery. This method can analyze intact data, obtain uncertain knowledge and offer an effective tool by reasoning. This theory was used by previous studies to develop their system.

Three-way decisions are described in terms of a three item classification according to evaluations of a set of criteria. The model the proponents plan to adopt.

Chapter II Review of Related Literature

This chapter focuses on the previous studies and researches related to the use of Bayesian Network. A compilation of different studies were used to complete this section of paper. The thematic explanation of Hierarchical Association Model and Three-way decisions using the concept of Bayesian Network is rephrased and summarized properly.

A. Market Basket

Mining for frequent itemset leads to finding association and correlation among items in a database. Industries are interested in mining there patterns for analysis from their databases for discovering these patterns can help their business when it comes to decision-making processes such as catalog design, cross-marketing, and customer behavior (Han et al., 2012)

A common example of mining frequent itemset is the market basket analysis. This process observes and analyzes the customer behavior by finding association in the items that a customer buys per transaction. For further explanation, given a customer buying a milk, how likely are they to buy bread too in a transaction (Han et al., 2012).

For instance a universe is a set of items at the store, then each item is represented as present or absent in a Boolean variable. Each basket therefore can be illustrated as Boolean vector of values assigned to these variables. Utilizing these Boolean vectors the buying pattern can now be analyzed that are reflected in items that are frequently bought or associated together. That pattern is called the association rules.

Given an association rule:

computer ⇒ antivirus software [support = 2%,confidence = 60%] .(1)

Rule support and confidence are two measures of rule thought provoking. Respectively it reflects the usefulness and certainty of discovered rules. In the example a support of 2% means that all transactions showed that computer and antivirus software were bought together. A confidence of 60% denotes that 60% of the customers who purchased a computer also bought an antivirus software. Association rules are considered valid if it meets both the minimum support threshold and minimum confidence threshold (Han et al., 2012).

A comparative study on Market Basket Analysis and Apriori Association Technique done by Warnia Nengsih discussed the comparison between market basket analysis with Apriori algorithm and without using any algorithm in generating association rules. According to the research both methods resulted the same association rule (Nangsih, 2015).

B. Association

Let I = $\{I_1, I_2, \ldots I_m\}$ be an itemset. Let D be the database of where each transaction T is a nonempty itemset such that T is a subset of I. Let A be a set of items, a transaction T is said to contain A if A is a subset of T. Now an association rule is an implicaiton of the form A implies B, where A is not B, both A and B is not empty and A, B is part of I and and empty is the intersection of A and B. A rule has a support s, where s is the percentage that AB is contained in D which is equal to P(AB). furthermore it has a confidence c, where c is the

percentage that of transactions in D that contains A that also contains B which is equal to the conditional probability P(B|A) (Han et. al., 2012)

Both support and confidence play a big role in association. A rule with a low support may be uninteresting to a business for it may lead to a low profit or no profit. For that support helps eliminates uninteresting rules. While a rule with a low confidence show that the reliability of inference made by the rule is weak. Take for instance X -> Y, the higher the confidence means that Y is likely to be in X. As stated by the formula it is the probability of Y given X. (Tan et. al., 2005).

C. Apriori

Apriori is an algorithm proposed by R. Agrawal and R. Srikant in 1994 for mining frequent itemsets for Boolean association rules (Hans et. al., 2012). Apriori algorithm finds all frequent itemset using breadth first search approach. The first step is finding all frequent 1-itemsets, and then discovering 2-itemsets and so on finding increasingly larger itemsets (Hegland, 2005). This process of the Apriori algorithm is an iterative approach, most commonly called level-wise search. (Hans et. al., 2012).

To improve efficiency of Apriori, a property is used to reduce search space that is: *All nonempty subsets of a frequent itemset must be a frequent.* By definition, if an itemset I does not satisfy the minimum threshold, it states that I is not frequent. Furthermore if an item is added to I, for example AI, the result can not be more frequent than I (Hans et. al., 2012).

D. Classification

Classification is from the word itself, it classifies models into classes. If the task of the data analysis is to classify the models, a classifier is constructed to predict class or categorical labels such as for medical data “treatment A”, “treatment B”, or “treatment C”; “safe” or “risky” for loan application data, and so on. These categorical labels can be represented by discrete values, where ordering is no importance. (Han et. al., 2012)

Data classification works in two-step process, namely learning step and classification step

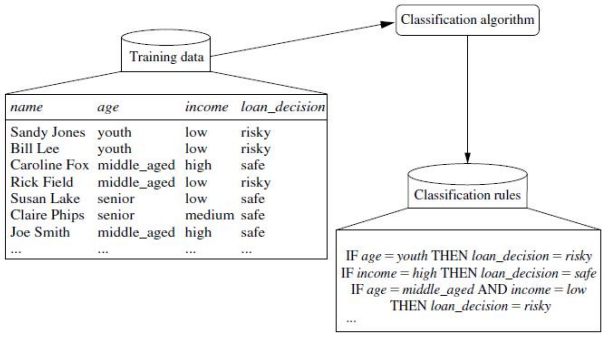


Figure II.D-1: Learning Process of Classification from Han et. al. (2012)

Figure II.D-1 shows the process of training data being analyzed by the classifier. The learned model is represented in the form of classification rules.

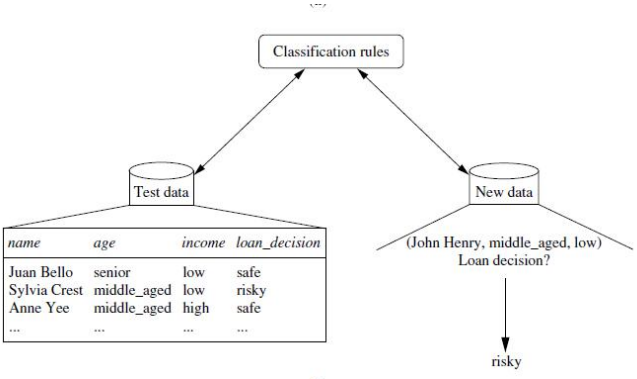


Figure II.D-2: Classification Process of Classification from Han et. al. (2012)

Figure II.D-2 shows the process of test data being estimated the classification rules, that where the classification is done based on the test data done (Han et. al.. 2012).

E. Applied Study - Analysis of Association Rule Mining using Bayesian Network (Raghu et. al., 2012) Applied Study - Minimum Cost Attribute Reduction In Three-Way Decisions Based Bayesian Network (Jia et. al., 2016)

Naive Bayes Algorithm is a classifier technique that is based on Bayesian theorem. This type of classifier is particular to high dimension of inputs since it is capable of calculating the “best” possible output. This considers the presence or absence of an attribute of a class even if the of a class is unrelated to the presence of absence of any other feature when a class variable is given. Even if some features may depend on each other and once there exist another feature of a class, a naive bayes classifier reflects all these properties to contribute independently. (Nikam, 2015)

According to Jia (2016) The three-way decisions based Bayesian network was better than the two-way decision. It could obtain a lower misclassification error. And can generate minimal decision cost. Because of this a new cost attribute reduct for the three-way decisions based Bayesian network is defined along with its heuristics. Knowing that, the only problem is what kind of classifier would perform better using three-way decisions.

id	*Data set*	*Condition attributes*	*Objects*
1	Bank Markting	16	4520
2	Credit Approval	15	653
3	Breast-cancer	9	572
4	Chronic Kidney Disease	24	400
5	Fertility	9	200
6	Haberman's Survival	3	612
7	The Monk's Problems	7	1713
8	SPECT heart	22	534
9	Australian Credit Approval	14	689
10	Punishing Websites	29	1054

Figure II.E-1: Brief Description of the Datasets

Based on their experiments, Figure II.E-1 shows the average length of different derived reducts based on the heuristic approach. Figure II.E-2 shows the misclassification costs of the three classifiers.

id	*Pawlak attribute reduct*	*Minimum cost attribute reduct*
1	11.2 ± 4.93	13.51 ± 2.82
2	10.5 ± 3.57	12.33 ± 3.92
3	3.7 ± 3.50	8.62 ± 6.30
4	7.8 ± 2.44	9.03 ± 1.26
5	6.4 ± 0.97	7.5 ± 1.53
6	1.88 ± 0.33	2.2 ± 3.55
7	2.4 ± 0.97	3.63 ± 2.21
8	15.6 ± 5.54	20.93 ± 4.67
9	7.78 ± 2.73	11.08 ± 3.13
10	10.9 ± 9.67	20.28 ± 5.85

Figure II.E-2: Average length of a reduct based on heuristic approach

id	*naive Bayesian*		*Bayesian Network*		*NBRS*	
	Pawlak reduct	*Minimum cost reduct*	*Pawlak reduct*	*Minimum cost reduct*	*Pawlak reduct*	*Minimum cost reduct*
1	136 ± 5.34	85.88 ± 8.35	109.39 ± 9.23	36.65 ± 8.82	89.52 ± 7.19	35.38 ± 5.92
2	13.89 ± 3.33	9.88 ± 3.74	15.28 ± 3.10	8.23 ± 2.69	12.89 ± 3.18	3.64 ± 2.71
3	57.50 ± 3.55	18.80 ± 8.61	29.68 ± 1.97	5.90 ± 1.17	38.24 ± 7.05	7.87 ± 2.90
4	15.81 ± 2.09	9.57 ± 3.92	11.28 ± 4.22	4.31 ± 2.51	8.66 ± 0.64	0.87 ± 0.91
5	8.31 ± 0.67	2.01 ± 1.23	4.15 ± 0.68	0.92 ± 0.64	5.73 ± 1.60	1.48 ± 1.14
6	61.44 ± 11.11	14.78 ± 3.73	59.35 ± 16.01	15.07 ± 3.53	34.94 ± 4.73	11.78 ± 4.29
7	58.11 ± 9.29	45.88 ± 3.21	47.24 ± 9.61	44.51 ± 5.23	34.28 ± 1.77	34.28 ± 4.37
8	45.80 ± 7.69	4.81 ± 4.81	34.55 ± 3.53	3.02 ± 3.02	25.67 ± 4.26	2.26 ± 2.23
9	35.77 ± 0.48	7.68 ± 1.65	32.44 ± 2.03	5.56 ± 1.62	25.25 ± 4.90	5.82 ± 1.27
10	90.16 ± 6.83	21.06 ± 4.11	81.64 ± 5.39	19.43 ± 3.17	78.75 ± 3.92	15.64 ± 2.43

Figure II.E-3: Misclassification costs of three classifiers based on different reducts

Figures II.E-1, II.E-2 and II.E-3 show the comparison accuracies of three classifiers based on the whole attributes and derived reducts.

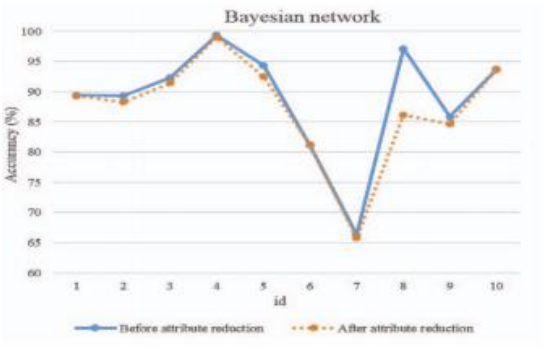


Figure II.E-4: Comparison accuracies of Bayesian network based on the whole attributes and the minimum cost attribute reduct

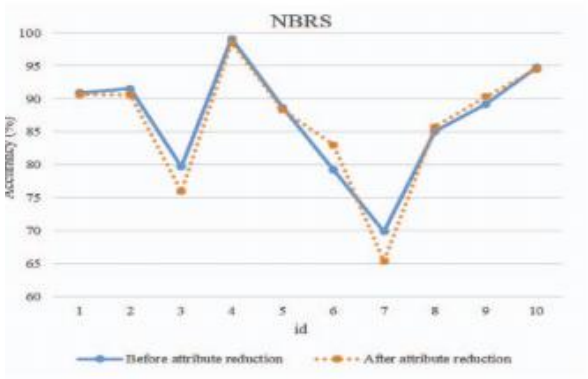


Figure II.E-5: Comparison accuracies of NBRS based on the whole attributes and the minimum cost attribute reduct.

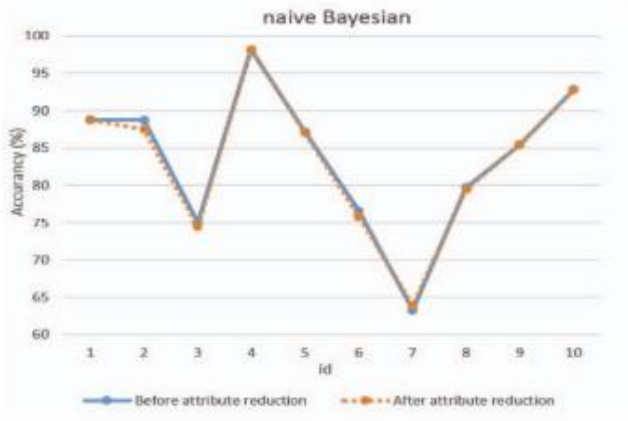


Figure II.E-6: Comparison accuracy of naive Bayesian based on the whole attributes and the minimum cost attribute reduct.

Based on their tests, different classifiers could induce less misclassification costs with varying classification accuracy based on their proposed reduct.

G. Naive Bayes

Naive Bayes is a probabilistic model based on Bayes theorem that goes well with for uncomplicated filtering. It works by building a probabilistic model that is used to predict an output based on the given sets of inputs by learning the conditional probabilities based on Bayes' rule of each input given a possible value taken by the output. Bayes' rule states that

$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)} \quad (2)$$

where P(A|B) is the probability of observing A given that B occurs. Basically P(A|B) is the posterior probability and P(B|A), P(A), and P(B) are prior probability. The probability of observing A given B is therefore can be found if when the individual probabilities of A and B are known as well as the probability of observing B given A is also known. (Akerkar and Sajja, 2016)

The naive bayes using Bayesian approach utilizes a set of training examples to classify instances. The class which will have the highest probability will be assigned to the instance that is observing each output class given the input attributes. The probability of the output attribute taking a value v_j when the given input attribute values are seen together is given by

$$P(v_j \mid a, b) \quad (3)$$

Applying the bayes theorem one will get

$$P(v_j \mid a,b) = \frac{P(a,b \mid v_j)\,P(v_j)}{P(a,b)} \quad = P(a,b \mid v_j)\,P(v_j)\,, \quad (4)$$

where $P(v_j)$ is the probability of observing v_j as the output value, and $P(a,b|v_j)$ is the probability of observing input attribute values a, b together when output value is v_j. However if the number of input attributes (a, b, c, d, ...) is large, then we likely will not have enough data to estimate the probability P(a, b, c, d, ...|v_j). The probability value P(a, b | v_j) can then be simplified as

$$P(a,b \mid v_j) = P(a \mid v_j)\,P(b \mid v_j)\,, (5)$$

Since naive bayes uses the conditional independence for all input attributes given the values for the output; it is assumed that all variables are independent of each other therefore the probability of observing an output value for the inputs can be obtained by multiplying the probabilities of individual inputs given the output value. where P(a | v_j) is the probability of observing the value a for an attribute when output value is v_j. Thus the probability of an output value v_j to be assigned for the given input attributes is

$$P(v_j \mid a,b) = P(v_j)\,P(a \mid v_j)\,P(b \mid v_j)\,. (6)$$

The naive Bayes algorithm involves all attributes in the instance to be discrete. Therefore if the value is continuous is has to be discretized before it can be utilized. Moreover missing values for an attribute are ignored, as they can be the cause of more problems in calculating the probability values for an attribute. Substituting is a default value is a general method to be used for replacing missing values. (Akerkar and Sajja, 2016)

H. Bayesian Network

A. Introduction

Bayesian Networks(BN) is also known as belief networks that belongs under the category of probabilistic graphical models (GMs) more specifically as a directed acyclic graph (DAG) defining a joint probability distribution over a set of variables. It is a combined principles from graph theory, probabilistic theory, computer science and statistics.

Each variable in a BN represented by a node in the graph and is independent of its parent node or ancestral node. With this property is utilize to characterized the JPD of variables since this reduces the number of parameters. Therefore BN demonstrate a conditional independence statement. Consider X, Y and Z variables and X is conditionally independent of Y given Z, if the probability distribution governing X is independent of the value of Y, given the value of Z.

$$(i,j,k)P(X=xi|Y=yj,Z=zk)=P(X=xi|Z=zk) \quad (7)$$

Commonly written as

$$P(X|Y,Z)=P(X|Z) \quad (8)$$

While node denotes a variable, edges denote dependencies or probabilistic dependencies. Where the conditional probability distribution (CPD) depends on the only on its parents.

For each node Xi describes

$$P(Xi|Pa(Xi)) \quad (9)$$

Where *Pa* is the immediate parents. Recalling the chain rule of probability

$$P(S,R,T,U,V) = P(S)P(R|S)P(T|R,S)P(U|T,R,S)P(V|U,T,R,S) \quad (10)$$

Put in the the Bayes Net

$$P(X1...Xn)=iP(Xi|Pa(Xi)) \quad (11)$$

For simplicity consider a person who might suffer from a back injury, an event represented by the variable Back. The injury can be caused by a backache represented by the variable Ache, might result from a wrong sport activity represented by the variable Sport or from new uncomfortable chairs installed at the person's office, represented by the variable Chair. For more cases, it was assumed that a coworker will suffer and report a similar backache syndrome, an event represented by the variable Worker. Variables are either true or false. The Conditional Probability Table of each node is listed besides the node. (Ruggeri et al, 2007)

Notice that the structure of the JPD modeled by a BN is called d-separation. It captures the conditional independence and how these affect the update. This is commonly implied to the Markov property which will determine the whether a set of nodes X is independent of another set Y, given a set of evidence nodes E. (Ruggeri, et al., 2007)

B. Inference

Inference in a Bayesian network is the task of computing the probability of each value of a node when all values are known. (Stephenson, 2000) For instance a BN was specified in JPD factored form, marginalize variables in order to compute all possible probability. According to Ruggeri, et al. Two types of inference can be done: (1) *Predictive Support* also known as top-down reasoning and, (2) diagnostic support also called bottom-up reasoning.

The diagnostic support is formulated as follows,

$$P(C=T \mid A=T) = P(C=T,A=T) / P(A=T) \quad (12)$$

Where

$$P(C=T,A=T) = S,W,B\{T,F\}\ P(C = T)P(S) \times P(W|C=T)P(B|S,C=T)P(A=T|B) \quad (13)$$

And

$$P(A = T) = S,W,B,C\in\{T,F\}P(C)P(S)P(W|C)P(B|S,C) \times P(A = T|B) \quad (14)$$

JPD running time however is O(2n)where n is the number of nodes. Hence takes exponential time this means to say that getting the summation of everything or exact inference falls under the NP-hard problem. Message passing algorithm and cycle-cutset conditioning and variable elimination are some examples of exact inference algorithm. Monte Carlo sampling that gives estimates to samples and more Markov chain Monte Carlo (MCMC) methods are samples of Approximate inference methods which can decrease the running time.

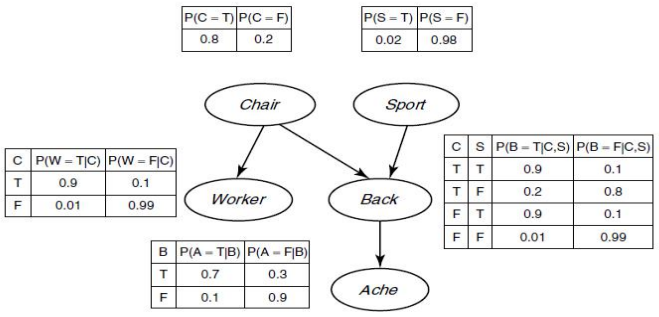


Figure II.H-1: Example of Bayesian Network Framework by Ruggeri F., Faltin F. & Kenett R.(2007)

Figure II.H-1 is an example given by Ruggeri F., Faltin F. and Kenett R. (2007) to illustrate the characteristic of Bayesian Network. It considers a person who might suffer from a back injury, an event represented by the variable Back (denoted by B). The injury can be caused by a backache represented by the variable Ache (denoted by A), might result from a wrong sport activity represented by the variable Sport (denoted by S) or from new uncomfortable chairs installed at the person's office, represented by the variable Chair (denoted by C). For more cases, it was assumed that a coworker will suffer and report a similar backache syndrome, an event represented by the variable Worker (denoted by W). Variables are either true (denoted by "T") or false (denoted by "F"). The Conditional Probability Table of each node is listed besides the node.

C. Learning

Values for BN is realistically unknown, that the be known it has to undergo incremental learning or training prior to information, combining expert knowledge or learning causal relationship. This is to have efficient representation and inference and be able to handle missing data This process is estimating the

network structure or graph topology and the parameters in JPD. Learning the parameters is considered easier than the structure. There are four main cases to learn task:

Table II.H-2: Four cases of BN Learning according to Ruggeri, et al. (2007)

CASE	BN STRUCTURE	OBSERVABILITY
1	Known	Full
2	Known	Partial
3	Unknown	Full
4	Unknown	Partial

The first case in Table II.H-1 is the simplest goal learning, for this is finding the values of the BN parameters in each conditional probability distribution that will maximize the likelihood training dataset. Second case only includes computing the estimate parameters. However in the third case, the goal would be learning the DAG that would explain the whole data since it is unknown, mentioned earlier in this paper that getting the running time through all the nodes of a DAG is exponentially high in N and is under the category of an NP-hard problem. One approach given by Ruggeri et al. (2007) is to assuming that the variables are conditionally independent given a class, which is represented by a single common parent node to all the variable nodes. This structure corresponds to the na¨ıve BN, which gave good results to practical problems. Lastly, the last case marginalizing the hidden nodes and the parameters since the DAG is partially observable and unknown.

I. Synthesis

Market basket analysis is the process of analyzing customer buying habits that customers buy. Studying these associations can help retailers to develop marketing strategies by gaining knowledge to what products are frequently bought

together by customers. The proponents propose to construct an associative classification that incorporates the two data mining techniques to create a better rule model for the market basket as well as for the medical industry.

The Apriori is an association rule base where it checks the frequency of an item or item set, which predicts the future trends, or predicting item/item set that would be bought the most in terms of owning a supermarket that sells wide array of items, which would fall under Market Basket.

To improve on Apriori, the proponents will be using Bayesian Network concept because two types of data mining procedure namely association and classification will be using Bayesian Network. To further explain, BN will be used for association in creating the probability to compute for the support, confidence and lift. Furthermore, it will play a role in three-way decisions together with the rough set theory for classification.

Chapter III Research Design and Methodology

A. Hypothesis

With reference to the problems stated in Chapter I, the proponents hypothesized that the usage of Bayesian Networks and Three-Way Decisions in the created solution could become a comparable alternative or better yet solve the limitations of Apriori.

B. Research Methods

1. Scientific method or Experimental Method

a. Problem

The Apriori algorithm operates on database that contains the transactions. It will then scan the dataset as many times needed and will create itemsets to produce the values for support and confidence.

This in turn will make the algorithm run very slow.

b. Experimentation

. *Gather the data*

The proponents had gathered the data to be experimented. Data that was gathered are related to retailing transactional data and assisted medical diagnosis data.

ii. Compare support, confidence, and lift for association

The proponents recorded the support, confidence and lift generated by the association rule of the proposed algorithm and Apriori. After which had tested the improvement of the means by t-test.

iii. Compute for accuracy and precision

The proponents computed for the following: accuracy and precision and was tested whether it made any significant difference.

iv. Test processing speed of the proposed algorithm versus Apriori

The proponents determined the processing speed of the proposed algorithm with respect to the size of the database. Then determined its peak efficiency.

C. Research Design

The study was focused on comparing the new system to Apriori algorithm. The study compared results for accuracy and actual running time. The proponents had undergo different phases in order to perform the study:

1. Data and Information Gathering

This step focused on gathering the information and data to be used in the study. The proponents had used different related materials that are available online like books, theses, journals, articles and other online resources relevant to association and classification. The data gathered had been used to test the proposed system.

2. Conceptualization and Data Preprocessing

After the proponents had gathered the related information, the proponents had analyzed the information and filter out the ones most relevant to the topic. As for the data gathered to be used for testing, the data had been pre-processed in a manner that will be accepted by the system that will be created.

3. Designing the Algorithm

The proponents had designed an algorithm implementing Bayesian Network in associating the dataset and Three-Way decisions for classifying the dataset. The inputs for the system had come from the data gathered. The output of the system had been compared to the output of the Apriori association and classification.

4. Creation and Implementation of the Proposed Solution

The approach conceptualized by the proponents had been translated into codes to create a system that will aim to satisfy the desired solution. Afterwhich test had been done to see if the solution improves the previous solution.

5. Assessment

After a series of tests and debugging, and when the system was ready to accept actual inputs of data. The proponents had generated results and outputs from the system, which had been compared to the outputs from Apriori.

6. Conclusion

The outputs of both systems had been compared and the testing of the differences had proven in a statistical manner to conclude whether the proposed system exhibited improvement in comparison to association and classification of Apriori. As well as, the validity of the output consultations with experts on the field will be needed.

D. Research Instruments

The proponents will use the following instruments to conduct the study:

1.) Hardware

The proponents will be using computers or laptops to test the software and prove the hypothesis. The devices should have the required specifications needed to run the software. The minimum specifications are as follows:

- Processor: Intel® Core™ i5
- Memory:8gb RAM
- Storage:20gb of storage space using HDD (better performance with SDDs)

These are the base specifics on what is needed for the hardware. Other parts such as the motherboard, graphics card, etc. will be abstracted.

The devices, aside from testing and proving the hypothesis, will also be used for researching, documentation, and implementation of the software

2.) Software

The proponents will be using tools for the Apriori as well as the proposed system, the results of Apriori and the results of the proposed system will be compared. For documentation purposes the proponents will be using Microsoft Word as well as Google Docs for collaboration purposes. Excel will also be used to view datasets and for tallying off results.Tools that computes the statistical treatment of the data will also be used such as Minitab moreover, Usage of database like oracle and sql is

also needed, since the input data will be stored in the database and the process will be producing different tables hence the need for database. Data Mining tools will be needed to finding correlations or patterns among dozens of fields in large relational databases. There are different data mining tools out there that will be readily used by the public such as Rapidminer. See table III-1 for tabulation.

Table III.D-1: Possible tools for each category

Data Mining Tools	1. Rapid Miner
Database	1. Oracle 2. PL/SQL
Documentation	1. Microsoft Word 2. Google Docs
Tabulation/Tallying of Results	1. Microsoft Excel 2. Google Sheets

3.) Identify sources of data:

The proponents will be gathering their input data from medical data and supermarket data. The data from these fields will be requested will the help of our relatives who are related to that field.

The study will only process data from retail industry and medical information.

E. Sampling and Data Gathering Procedure

The proponents will be following the ETL procedure, extract, transform and load procedure to how to process the data that is loaded from the source to the data warehousing. Below are the steps of the ETL procedure:

1. Extract

   Extracting includes getting the relevant data to be used for example given the data set the relevant data to be left are the transaction number and which items were bought by,

Table III.D-2: Dummy Transactions of a Company

Transaction Number	Items
1	1,3,4
2	2,3,5
3	1,2,3,5
4	2,5

given the sample dataset the ETL extract step will leave out the attributes to be used and disregard the other attributes.

2. Transform

   In this phase, the proponents will transform the data into associating rules data that will be followed by the classified data.

   a. During the transformation to associating rules, the proponents will apply the Bayesian Network algorithm to formulate it and thus generating support, confidence and lift per rule.

Table III.D-3: Strong Rules of the transaction

Strong Rules	Support	Confidence	Lift
A ->E	X	X	X
A ->BE	X	X	X
.	X	X	X
.	X	X	X
.	X	X	X
.	X	X	X
E ->ABC	X	X	X

b. During the transformation to classified data, the proponents will compute for the logistic regression per association rule for the means to classify the possibility of it to be bought and if it's under critical or not for market basket analysis and assisted medical diagnosis, respectively.

3. Load

Loading is creating a collection, that is summarizing and storing data which will the be used for ease of extraction for visuals if needed. The importance of aggregates is to improve performance of end-user queries.

F. Statistical Treatment of the Data

The treatment of the data will be gathered by the researchers subjected to computational procedure. Data or information that describes the accuracy, reliability and speed of the system in terms of associating the retailing purchases and medical symptoms and classifying the which associated purchases or symptoms is more likely to happen.

1. Accuracy and Precision

Accuracy is how close the results of the new algorithm and Apriori are. To get the accuracy of the new algorithm the formula was used

$$Accuracy = \frac{\text{Results of algorithm}}{\text{Doctor's Opinion}} x100 \quad (15)$$

For the precision, computing the true positives and false positives for the formula of precision. After which precision of both algorithm will be computed and compared if there is a significant difference. Formula for precision is as follow,

$$\text{Precision} = \frac{tp}{tp + fp} \quad (16)$$

3. Significance Testing

For significance testing, t test was used. The significance test has two hypotheses: a null and an alternative hypothesis. A null hypothesis states that there is no significant difference between the two variables while the alternative is either larger or smaller to the given value. For testing the accuracy of the one-sample t test was used to compare the mean of the population with the theoretical value. To compare the mean of a population to a specified theoretical mean,

$$t = \frac{m-\mu}{s/\sqrt{n}} \quad (17)$$

Let X represents a set of values with size n, with mean m and with standard deviation S. The comparison of the observed mean (m) of the population to a theoretical value μ is performed with the formula above.

For the run time of the system two independent sample t test was used to compare the mean run time of the new algorithm and Apriori. For independent t test formula

$$t = \frac{m_A - m_B}{\sqrt{\frac{S^2}{n_A} + \frac{S^2}{n_B}}} \quad (18)$$

Let A and B represent the two groups to compare, m_A and m_B represent the means of groups A and B, respectively and n_A and n_B represent the sizes of group A and B, respectively. S^2 is an estimator of the common variance of the two samples. It can be calculated as follow

$$S^2 = \frac{\Sigma(x-m_A)^2+\Sigma(x-m_B)^2}{n_A + n_B - 2} \quad (19)$$

Chapter IV Presentation and Analysis of Data

A. System Architecture

This study is focused on creating an improved market basket analysis using Apriori algorithm. To test the two different programs in terms of accuracy and speed, the researchers programmed the new system which uses Bayesian Network and Three-Way decisions and used rapid miner for Apriori algorithm.

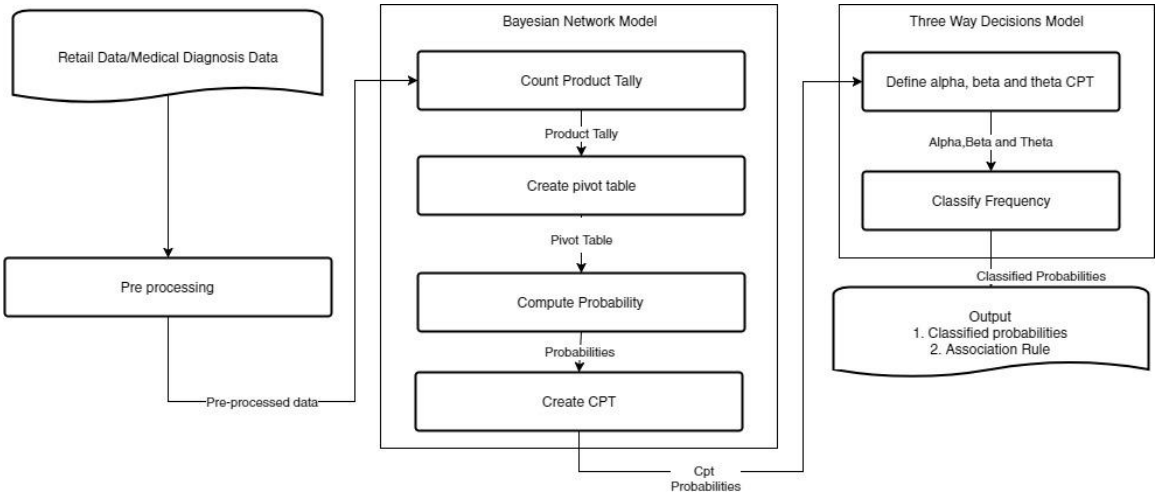


Figure IV.A-1 System Architecture

The system in Figure IV.A-1 has two major modules: Bayesian Network and Three-Way Decisions. The program starts with the input Medical Diagnosis Data. The pre processing of the data includes the extraction of columns that will be use to count the frequencies of each item in the Count Product Tally then the product tally is then used as input in Create Pivot Table to create a Pivot Table which will be used to generate probabilities using the Compute Probability module and the said probabilities will be then used to create the Conditional Probability Table (CPT) of each product.. The CPTs are then used to define the alpha, beta, and theta values. These are the threshold that will be used to classify the most frequent combinations and generate the association rule.

B. Description of the Modules and Interfaces

B.1. Pre- processing Module

This module accepts the raw data then extracts the data from specific columns needed in the whole process, patient-id and symptoms.

B.2. Bayesian Network Model

The Bayesian Network Model, composed of four functions, is used to generate the Conditional Probability Table (CPT) of each disease.

B.2.1. Input

The input for the Bayesian Network Model is the output of the preprocessed data of the preprocessing module which is the combined data of both old and new removing all the same transactions or patient.

B.2.2. Process

1. Count Product Tally - The module will count how many times the disease was diagnosed, tallying it per patient. It will then creating a topology of what diseases influence the others. This results an update to Dimension Table on what disease is the parent or children of what disease, basically creating a network of diseases.

2. Create Pivot Table - The module creates a pivot table that would show in which patient the disease was present or bought.

3. Compute Probability - The module computes the base probability of each disease.

4. Create CPT - After the base probability is computed, the module creates the “Conditional Probability Table/s” to be used as inputs for the Three Way Decisions Model.

B.2.3. Output

The module will produce Conditional Probability Table (CPT) of each disease. The probabilities on the table will be used to determine which disease are likely to be together.

B.3. Three Way Decisions Model

It is the classifier of the program which uses the alpha, beta, and theta as ranges. These alpha, beta, and theta values came from the Bayesian model using the CPT probabilities.

B.3.1. Input

The input for Three Way Decisions Model is the output of the Bayesian Model which is the CPT probabilities that will be used in determining the alpha, beta, and theta.

B.3.2. Process

1. Define alpha, beta, and theta - Determining the alpha, beta and theta of the probabilities involves getting the maximum, median and minimum of all probabilities produced excluding the zero.

2. Classify Frequent - After defining the maximum, median and minimum of the probabilities computed, the system will classify all transactions creating a table for frequent, neither frequent or non-frequent and non-frequent.

B.3.3. Output

The module will produce an output of Classified probabilities distributed into three tables, frequent, neither frequent or non-frequent and non-frequent, displaying its support, lift and confidence.

C. Sample System Simulation of Test Data

Here is a sample system simulation of sample raw data using Bayesian Networks and Three-Way Decisions (BNTWD) Simulation, and RapidMiner MarketBasket.

C.1. Bayesian Network and Three Way Decisions Simulation

Table IV.C-1 Sample_Raw Table

Invoice_no	product_code	product
675	141	A41
819	141	A41
857	141	A41
857	116	A16
1573	109	A09
1743	116	A16
14496	109	A09
16176	149	A49
16502	141	A41
17068	234	B34
18836	118	A18
21187	109	A09
21762	109	A09
21951	109	A09
22527	109	A09
23152	109	A09
23198	235	B35

Table IV.C-1 shows the sample raw data used for simulation, containing the extracted columns and no repeated disease per patient no. The system will tally the diseases and create a topology resulting to Table IV.C-2, IV.C-3, IV.C-4, and IV.C-5.

Table IV.C-3: Count Table

PRODUCT_CODE	COUNT_PRO...
117	13
103	14
106	16
215	16
101	17
102	19
104	23
202	24
201	25
216	25
119	27
237	41
115	50
210	59
149	61

Table IV-3 is the tally of per products. Using the tally, the system will assign which is the parent node and the children node.

Table IV.C-5: Children Table

PRODUCT_...	COUNT_PRODUCT	CHILDREN_ID
278	1	1
294	1	1
205	1	1
206	1	1
135	1	1
282	1	1
254	1	1
220	1	1
146	2	2
236	2	2
225	2	2
127	2	2
285	2	2

Table IV.C-4: Parent Table

PRODUCT...	COUNT_P...	PARENT_ID
127	2	1
285	2	1
258	2	1
233	2	1
265	2	1
180	2	1
107	3	2
207	3	2
192	3	2
131	4	3
189	4	3
137	4	3
186	5	4
224	5	4
219	6	5
148	6	5
163	6	5
235	6	5
249	6	5
152	6	5
245	6	5
217	7	6
138	8	7
200	9	8
299	11	9
230	12	10

Table IV.C-5 and IV.C-4 is the look up table for children table and parent table respectively. It tells which disease is the diseas assigned to children/parent ID.

Table IV.C-5: Dimension Table

PRODU...	PRODUCT	PARENT_ID
101	A01	101
102	A02	102
103	A03	103
104	A04	104
106	A06	106
107	A07	107
108	A08	108
109	A09	109
115	A15	115
116	A16	116
117	A17	117
118	A18	118
119	A19	119

Table IV.C-5 is another lookup table called dimension table. It is a general lookup table that also a general representation of the network itself.

Table IV.C-6: Probabilities Table

PRODUCT_...	PROBABILITY
227	0.000227376080036380172805820827648931332424
140	0.000227376080036380172805820827648931332424
226	0.000227376080036380172805820827648931332424
150	0.000227376080036380172805820827648931332424
259	0.000227376080036380172805820827648931332424
187	0.000227376080036380172805820827648931332424
243	0.000227376080036380172805820827648931332424
244	0.000227376080036380172805820827648931332424
166	0.000227376080036380172805820827648931332424
278	0.000227376080036380172805820827648931332424
294	0.000227376080036380172805820827648931332424
205	0.000227376080036380172805820827648931332424
206	0.000227376080036380172805820827648931332424
135	0.000227376080036380172805820827648931332424
282	0.000227376080036380172805820827648931332424

After creating the base network of the data, the system will compute the probabilities of each disease resulting to Table IV.C-6.

Table IV.C-7: Pivot Table

PRODUCT_CODE	PRODUCT	N1	N2	N3	N4	N17
102	A02	0	0	0	0	0
104	A04	0	0	0	0	0
106	A06	0	0	0	0	0
109	A09	0	0	0	0	0
115	A15	0	0	0	0	0
116	A16	0	0	0	0	0
117	A17	0	0	0	0	0
118	A18	0	0	0	0	0
119	A19	0	0	0	0	0
141	A41	0	0	0	0	0
149	A49	0	0	0	0	0
189	A89	0	0	0	0	0
190	A90	0	0	0	0	0
202	B02	0	0	0	0	0

As shown in Table IV.C-7, it is the result of the system creating a pivot table displaying which disease was present in which transactions.

The pivot table will be used to compute for the Conditional Probability Table (CPT) of each product. If the disease does not have any parent, the parent/s column is replaced by its own. The value each parents hold is an indication; if it is present, it will be represented by 1. If it isn't present, it will be represented by 0. It will then check if the probability that such combination will exist based on the pivot table.

During the computation of the CPT of each products, the products included together alongside the probability will be added to another table named CPT_P excluding all 0 probabilities computed. This is the input for the three-way decision, instead of using all CPTs generated.

Table IV.C-8 Three-way Table

ALPHA	BETA	THETA
0.30496453900709219858156028368794326241 13	0.02135718891038039974210186976144422952 88	0.00709219858156028368794326241134751773049 6

Using the CPT_P as basis, the system will compute for the maximum, median and minimum probabilities denoted by alpha, beta, and theta. It will be stored in the table called three-way table.

Table IV.C-9: Frequent Combinations

Combination
A09, A02
A09, A03
A09, A90
A09, B09
A09, B24
A09, B34

Table IV.C-10: Association Rule

PREMISE	CONCLUSION	CONFIDENCE
A09,	A02	1
A09,	A03	1
A09,	A17	0.333333333
A09,	A90	0.5
A09,	A91	0.5
A09,	B09	1
A09,	B24	1
A09,	B34	0.692307692
A09,	B99	0.6
A09, A16,	A17	0.333333333
A09, A16, A18,	A17	0.333333333
A09, A18,	A17	0.333333333
A09, B34,	A90	0.25

After attaining the alpha, beta and theta values, it will now split the data in CPT_P, it will get all frequent combinations and rules created.

C.2. RapidMiner Apriori Simulation

Table IV.C-11: Raw Table

[A16'	'A41']		
[A16'	'A31'	'B49'	'B96']
[A49'	'B95']		
[A16'	'A41']		
[A06'	'A09']		
[A09'	'B30']		
[A09'	'B18']		

[A41'	'B95']		
[A41'	'B18']		
[A16'	'A18'	'A19'	'B02']
[A41'	'A49']		

This is the input data for RapidMiner's Analysis. It is the same one that the proposed system used, with only the changed column names for the program to work properly in RapidMiner. "Invoice" replacing invoice_no from sample_raw, "Product 1" replacing product_no from sample_raw.

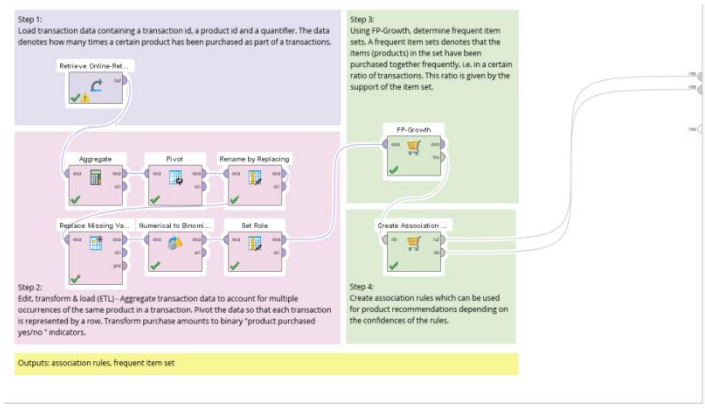


Figure IV.C-1: RapidMiner Analysis Process

This is the process for Analysis using RapidMiner.

Table IV.C-12: Frequency Table

Combination	Support
(A09)	0.362
(A89)	0.014
(A41)	0.433
(A52)	0.014
(B16)	0.021
(A19)	0.043
(B99)	0.028
(B58)	0.014
(A91)	0.014
(B24)	0.014
(A02)	0.014
(A04)	0.043
(A49)	0.099
(A90)	0.028

This table shows the frequent item sets with its support value.

Table IV.C-13: Association Rules Table

PREMISE	CONCLUSIC	CONFIDENCE
A09,	A02	1
A09,	A03	1
A09,	A17	0.333333333
A09,	A90	0.5
A09,	A91	0.5
A09,	B09	1
A09,	B24	1
A09,	B34	0.692307692
A09,	B99	0.6
A09, A16,	A17	0.333333333
A09, A16, A18,	A17	0.333333333
A09, A18,	A17	0.333333333
A09, B34,	A90	0.25
A15,	A89	0.5
A15,	B17	0.25
A15,	B58	0.5
A15, B37,	A89	0.5
A15, B37,	B58	0.5

This table shows the value for support, confidence, and lift.

D. Test Results

Here is the results of the old algorithm and new algorithm

D.1. Correctness of Association Rules

Table IV.D-1: Doctor's Opinion Regarding the Correctness of Association Rules

Doctor #	Apriori	Bayes Net
1	72.2%	41.66%
2	39.79%	79.59%
3	46.29%	34.69%

Table IV.D-1 shows the percentage of agreement of the doctors in terms of the association rule generated by both algorithms. We can see from the results that doctor 1 gave most disagreement to the generated results of the Bayesian Networks and Three-Way Decisions (BNTWD) algorithm as compared to doctor 2 who gave a higher percentage of agreement to the generated results of BNTWD as compared to the

generated results of the Apriori algorithm. However, the last doctor gave a below 50% correctness to both algorithms

D.2. Correctness of Frequent Combinations

Table IV.D-2: Doctor's Opinion Regarding the Correctness of Frequent Combinations

Doctor #	Apriori	Bayes Net
1	69.44%	50%
2	44.44%	100%
3	44.44%	45.83%

Both algorithms generated frequent combinations, together with its probability of chance to happen, Table IV.D-2 shows the result of doctor's opinion regarding the correctness of the combinations and its frequency. Same with the result to the association rules, doctor 1 gave a higher agreement to the Apriori algorithm's results as compared to the Bayesian Networks and Three-Way Decisions (BNTWD) algorithm. Doctor 2 agreed to all of the combinations generated by BNTWD, and less than half of the combinations of the Apiori algorithm were disagreed. As for the last doctor, she gave both algorithms below 50% correctness in terms of frequency.

D.3. Precision

Table IV.D-3: Tree Label for Doctor 1

Doctor 1				
Association Rule - Apriori			Association Rule - Bayes Net	
	1	0	1	0
1	39	0	21	0
0	15	0	28	0
Frequency - Apriori			Frequency - Bayes Net	
1	25	0	39	0

0	11	0	15	0

Table IV.D-3 shows the tree label for doctor 1 and how the precision was computed.

Table IV.D-4: Tree Label for Doctor 2

Doctor 2				
Association Rule - Apriori			Association Rule - Bayes Net	
	1	0	1	0
1	20	0	39	0
0	34	0	10	0
Frequency - Apriori			Frequency - Bayes Net	
1	16	0	25	0
0	20	0	0	0

Table IV.D-3 shows the tree label for doctor 2 and how the precision was computed.

Table IV.D-5: Tree Label for Doctor 3

Doctor 3				
Association Rule - Apriori			Association Rule - Bayes Net	
	1	0	1	0
1	25	0	17	0
0	29	0	32	0
Frequency - Apriori			Frequency - Bayes Net	
1	16	0	11	0
0	20	0	13	0

Table IV.D-3 shows the tree label for doctor 3 and how the precision was computed.

Table IV.D-6: Precision of Results

	ASSOCIATION	FREQUENT	ASSOCIATION	FREQUENT
	Apriori		Bayes Net	
1	72.22%	69.44%	42.85%	50%
2	37.04%	44.44%	79.59%	100%

3	46.29%	44.44%	34.69%	45.83%

The precision of the results are equal to the fraction of relevant instances among the retrieved instances. Table IV.D-2 shows the precision score of both algorithm in terms of association and frequency.

For first doctor results, 72.22% is greater than 42.85% which means to say that association rule of the Apriori algorithm to doctor 1 is significantly better. As for the frequency, 69.44% is greater than 50% the Apriori algorithm is still significantly better.

For the second doctor results, 79.59% is greater than 39.04% which means to say that association rule of the Bayesian Networks and Three-Way Decisions (BNTWD) algorithm to doctor 2 is significantly better. As for the frequency, 100% is greater than 44.44% BNTWD algorithm is still significantly better.

Lastly for the third doctor results, 46.29% is greater than 34.69% which means to say that association rule of the Apriori algorithm is significantly better. As for the frequency, 45.83% is greater than 44.44%. BNTWD algorithm is significantly better.

F. Analysis and Interpretation of the Results

Three doctors were asked to answer two questionnaires containing fifty-five and fifty sets of association rules generated by the Apriori algorithm and the Bayesian Networks and Three-Way Decisions (BNTWD) algorithm respectively, verifying if said association rules were correct according to each doctor’s expertise and experience in the doctor’s respective medical field. The results from the questionnaires were used to as a

point of comparison, determining the accuracy of the association rules and the accuracy of the frequency combinations of each algorithm

According to the results of the first doctor's questionnaire, BNTWD would generate the correct combination of sickness for every association rule. However, the probabilities associated to each result would either be abnormality high or low. The Apriori algorithm would perform better than BNTWD, both in terms of the accuracy of the association rules and the accuracy of the frequency combinations. It should be taken into consideration that the doctor's specialization in pediatrics may influence the doctor's opinion on both algorithm's association rules.

In the perspective of the second doctor, the probabilities associated for every association rule should only be 100% or 50%. A diagnosis should be delivered with the respective probabilities to prevent confusion and unnecessary panic. In doing so, the results would be influenced by fitting the probability output of both algorithms according to the doctor's perception. According to the answers of the second doctor, BNTWD proved to be better both in the accuracy of the association rules and the accuracy of the frequency combinations. The doctor specializes in EENT (eyes, ears, nose, and tongue), which may influence the doctor's opinion on both algorithm's association rules.

The third doctor, on the other hand, expressed that there was always a possibility for a patient to have two diagnoses, regardless of its relatedness. Should the third doctor answer the questionnaires based solely on their experience and expertise, all of the association rules would be deemed correct. According to the answers of the second doctor, the Apriori algorithm was better in terms of the accuracy of the association rules,

while on the other hand, BNTWD was better in terms of the accuracy of the frequency combinations.

Table IV.D-7: Summary of Results in Correctness

	T-value	P-value
Association Rule Correctness	0.0456	.482908
Frequency Correctness	-0.64799	.276155

The Summary of results is shown in table IV.D-7, it says that both correctness shows no significance at significant level 0.01. Meaning there is no enough evidence to show that the BNTWD algorithm is better at medical diagnosis, compared to the Apriori algorithm. From a medical standpoint, diagnosis depends on a plethora of factors, alongside patients' history or medical records. A diagnosis may be affected based on a disease a patient may have contracted years before the patient's current state. Both algorithms use a probabilistic approach, generating probabilities to determine which would be the most appropriate diagnosis based on the association rules generated by both algorithms. Referencing diagnosis of a patient in a case to case standpoint deviates from the medical method, which focuses more on the factors affecting the patient themselves to make a diagnosis.

Table IV.D-8: Count of Significant Better

Algorithm	Count	
	Association	Frequency
Apriori	2	1
Bayes Net	1	2

Table IV.D-8 shows how many times that the Apriori algoritm and the Bayesian Networks and Three-Way Decisions (BNTWD) algorithm was significantly better. Results show than the Apriori algorithm won in association rule and BNTWD won in showing correct frequent combinations.

Based on the results, two out of the three doctors agreed that the Apriori algorithm proved to be significantly better in correctness of association rule. A factor in

the result is the way the Apriori algorithm traverses all the transaction present in the database. The Apriori algorithm is more likely to correctly determine the cause and effect for each diagnosis. On the other hand, two out of the three doctors agreed that BNTWD was significantly better in correctness of combination frequency. Which proves to be better in determining the correct relationships between signs and symptoms.

Chapter V Summary, Conclusions, and Recommendations

A. Summary

The researchers aim to develop an approach in generating association rules using the Bayesian Networks and Three-Way Decisions (BNTWD) algorithm in assisted medical diagnosis. The study made use of medical diagnosis dataset to be run using the old and the new algorithm.

The system accepts dataset inputs that are arranged by patient id with the symptoms or sickness the patient has. The dataset was specific to sickness in certain infectious and parasitic disease. With total of 141 patients and 198 diseases and infectious. It first counts the number of times the illness occurred to the patient list as basis for the topology of the network. After which the network was build together with its conditional probability table. The support build in the table is used to classify whether the association rule is frequent or not.

The algorithm was compared by first generating the association rule and frequent combinations of both algorithms. The proponents surveyed three doctors to verify the correctness of the output of both algorithms. Results was both algorithm showed correct combinations and probabilities vary from doctor to doctor's opinions, knowledge on the disease, and specialization of the doctor. Two out of three doctors resulted that Apriori algorithm is significantly better in correctness of association rule. While two out of three doctors resulted that Bayes Net algorithm is significantly better in correctness of combination frequency.

B. Conclusions

The researchers conclude that Bayesian Networks and Three-Way Decisions (BNTWD) algorithm has shown its success in the frequency of correct combinations compared to the Apriori algorithm.

According to the doctors' survey BNTWD association rules is not better than the Apriori algorithm.

Both algorithms' output neither correct nor incorrect since not enough evidence can support that one is better than the other in diagnosis. Diagnosis is best accommodated by patient's medical examination results and results are covered in doctor's opinions and knowledge of the disease.

C. Recommendations

The researchers recommends that future researchers, willing to improve the study can provide the following for improvement:

1. Improve the association by making use of dataset that has patient's physical examination than using patient's medical history.
2. Survey specific specialization of doctors with respect to the classification of disease to see if there is a relationship of accuracy with respect to the specialization of the doctor to the disease associated.
3. Allow the algorithm to accept more inputs and larger dataset for better association rule generation.
4. Integrate machine learning to help system store and makes use of previous input for better association rule generation.

References

Akerkar, R., Sajja, P. S. (2016). Intelligent Techniques for Data Science. Springer International Publishing AG.

Awodun, M., & Adedara, R. (2017) *An Efficient Rule-Mining for Medical Diagnosis: A Market-Basket Approach.* IJRAR- International Journal of Research and Analytical Reviews, 96-107

Ruggeri, F. (2007). Bayesian Networks. Encyclopedia of Statistics in Quality & Reliability, Wiley & Sons.

Investopedia Staff (2007). Market basket. In. Retrieved from http://www.investopedia.com/terms/m/market_basket.asp

Han, J., et. al. *Data Mining: Concepts and Techniques,* 3rd Edition, Morgan Kaufmann Publishers, Elsevier Inc., 2012

Hegland, M. *The Apriori Algorithm- a Tutorial.* CMA, Australian National University, March 2005

Jia, X. (2016). *Minimum Cost Attribute Reduction in Three-Way Decisions Bayesian Network.* Nanjing,China: Nanjing University of Science and Technology.

Kaur, M., & Kang, S. (2016). *Market Basket Analysis: Identify the changing trends of market data using association rule mining.* Prodedia Computer Science, 74-85.

Lee, M. C. (2011). *The Origin and Development of Markets:A Business History Perspective*. Retrieved:http://www.hbs.edu/businesshistory/Documents/origin-and-development-of-markets.pdf

Levy, R. (2012). *Probabilistic Models in the Study of Language.* University of California.

Mark Casson & John S. Lee (2011). *The Origin and Development of Markets: A Business History Perspective.*

Nengsih, W. *A Comparative Study on Market Basket Analysis and Apriori Association Technique*. 3rd International Conference on Information and Communication Technology (ICoICT), 2015

Nikam, S. *A Comparative Study of Classification Techniques in Data Mining Algorithms.* Oriental Journal of Computer Science & Technology. Vol. 8. 13-19. April 2015.

Raghu, D., P. Jagadeesh, and CH Raja Jacob. *Analysis Of Association Rule Mining Using Bayesian Network*. 1st ed. Dept. of CSE, NOVA College of Engineering & Tech., WG, AP, India: N.p., 2017. Web. 19 Feb. 2017.

Silvia Rissino and Germano Lambert-Torres (2009). Rough Set Theory — Fundamental Concepts, Principals, Data Extraction, and Applications, Data Mining and Knowledge Discovery in Real Life Applications, Julio Ponce and Adem Karahoca (Ed.), ISBN: 978-3-902613-53-0, InTech, Available from: http://www.intechopen.com/books/data_mining_and_knowledge_discovery_in_real_life_applications/rough_set _theory_-_fundamental_concepts__principals__data_extraction__and_applications

Stephenson, T. A. (2000). *An Introduction to Bayesian Network Theory and Usage.* Switzerland: Dalle Molle Institute for Perceptual Artificial Intelligence.

Tan, P., Strinbach, M., Kumar, V. (2005) *Introduction to Data Mining*. Pearson Education

Zagorecki, A., Orzechowski, P., & Hołownia, K. (2013) *A System for Automated General Medical Diagnosis using Bayesian Networks.* Studies in health technology and informatics. 192. 461-5. 10.3233/978-1-61499-289-9-461.

Appendices

A. Test Results

A.1. Apriori Results

Table AA-1: Frequent Combinations Generated by Apriori

Combination	Support
(A09)	0.362
(A89)	0.014
(A41)	0.433
(A52)	0.014
(B16)	0.021
(A19)	0.043
(B99)	0.028
(B58)	0.014
s(A91)	0.014
(B24)	0.014
(A02)	0.014
(A04)	0.043
(A49)	0.099
(A90)	0.028
(B09)	0.028
(B18)	0.057
(A16)	0.27
(A15)	0.085
(B95)	0.043
(B96)	0.106
(B02)	0.028
(B49)	0.035
(A17)	0.021
(B17)	0.028
(A18)	0.05

(B37)	0.071
(B34)	0.092
(A31)	0.021
(A03)	0.014
(B96, A41)	0.078
(A16, A18)	0.035
(A49, A41)	0.064
(B49, A41)	0.021
(A19, A18)	0.014
(A16, A41)	0.099
(A09, A90)	0.014
(B34, A09)	0.064
(A09, A16)	0.064
(B02, A41)	0.014
(A16, B02)	0.014
(B96, A49)	0.014
(A09, A49)	0.021
(B34, A90)	0.014
(B18, A41)	0.014
(B96, A16)	0.014
(A18, A41)	0.014
(B37, A41)	0.043
(A09, A04)	0.014
(B95, A49)	0.014
(A19, A41)	0.014
(A09, A41)	0.057
(A02, A09)	0.014
(B24, A09)	0.014
(B16, B18)	0.014
(A16, B18)	0.014

(A15, A16)	0.035
(A09, B99)	0.014
(A04, A41)	0.021
(B09, A09)	0.028
(A16, A49)	0.014
(B99, A41)	0.014
(A03, A09)	0.014
(A31, A16)	0.014
(B96, A09)	0.021
(A15, A41)	0.021
(B34, A16)	0.014
(B96, A49, A41)	0.014

Table AA-2: Association Rule Generated by Apriori

Premise	Conclusion	Confidence
(A41)	(B96)	0.18
(B96)	(A41)	0.733
(A18)	(A16)	0.714
(A16)	(A18)	0.132
(A41)	(A49)	0.148
(A49)	(A41)	0.643
(B49)	(A41)	0.6
(A18)	(A19)	0.286
(A19)	(A18)	0.333
(A41)	(A16)	0.23
(A16)	(A41)	0.368
(A90)	(A09)	0.5
(A09)	(B34)	0.176
(B34)	(A09)	0.692
(A16)	(A09)	0.237
(A09)	(A16)	0.176

(B02)	(A41)	0.5
(B02)	(A16)	0.5
(A49)	(B96)	0.143
(B96)	(A49)	0.133
(A49)	(A09)	0.214
(A90)	(B34)	0.5
(B34)	(A90)	0.154
(B18)	(A41)	0.25
(B96)	(A16)	0.133
(A18)	(A41)	0.286
(B37)	(A41)	0.6
(A04)	(A09)	0.333
(A49)	(B95)	0.143
(B95)	(A49)	0.333
(A19)	(A41)	0.333
(A41)	(A09)	0.131
(A09)	(A41)	0.157
(A02)	(A09)	1
(B24)	(A09)	1
(B18)	(B16)	0.25
(B16)	(B18)	0.667
(B18)	(A16)	0.25
(A16)	(A15)	0.132
(A15)	(A16)	0.417
(B99)	(A09)	0.5
(A04)	(A41)	0.5
(B09)	(A09)	1
(A49)	(A16)	0.143
(B99)	(A41)	0.5
(A03)	(A09)	1

(A31)	(A16)	0.667
(B96)	(A09)	0.2
(A15)	(A41)	0.25
(B34)	(A16)	0.154
(A41,A49)	(B96)	0.222
(A49)	(A41)	0.143
(B96,A41)	(A49)	0.182
(B96)	(A41)	0.133
(B96,A49)	(A41)	1

A.2. Bayes Net Results

Table AA-3: Association Rule Generated by Bayes Net

PREMISE	CONCLUSION	CONFIDENCE
A09,	A02	1
A09,	A03	1
A09,	A17	0.333333333
A09,	A90	0.5
A09,	A91	0.5
A09,	B09	1
A09,	B24	1
A09,	B34	0.692307692
A09,	B99	0.6
A09, A16,	A17	0.333333333
A09, A16, A18,	A17	0.333333333
A09, A18,	A17	0.333333333
A09, B34,	A90	0.25
A15,	A89	0.5
A15,	B17	0.25
A15,	B58	0.5
A15, B37,	A89	0.5
A15, B37,	B58	0.5

A16,	A15	0.416666667
A16,	A17	0.333333333
A16,	A18	0.714285714
A16,	A31	0.666666667
A16,	B02	0.5
A16,	B18	0.25
A16, A18,	A17	0.333333333
A18,	A17	0.333333333
A18,	A19	0.333333333
A19,	A17	0.333333333
A19,	A52	0.5
A19,	B58	0.5
A41,	A04	0.333333333
A41,	A09	0.156862745
A41,	A16	0.368421053
A41,	A19	0.333333333
A41,	A49	0.571428571
A41,	A52	0.5
A41,	A89	0.5
A41,	B02	0.5
A41,	B18	0.25
A41,	B37	0.6
A41,	B49	0.6
A41,	B96	0.733333333
A41, A16,	B02	0.25
A41, A18,	A19	0.166666667
A49,	B95	0.333333333
B18,	B16	0.666666667
B18,	B17	0.25
B34,	A90	0.5

B37,	A89	0.5
B37,	B58	0.5
B95,	A91	0.5
	A04	0.5
	A09	0.843137255
	A15	0.5
	A16	0.631578947
	A17	0.333333333
	A18	0.285714286
	A19	0.5
	A31	0.333333333
	A49	0.357142857
	A90	0.25
	B02	0.25
	B16	0.333333333
	B17	0.5
	B18	0.5
	B34	0.307692308
	B37	0.4
	B49	0.4
	B95	0.666666667
	B96	0.266666667
	B99	0.4

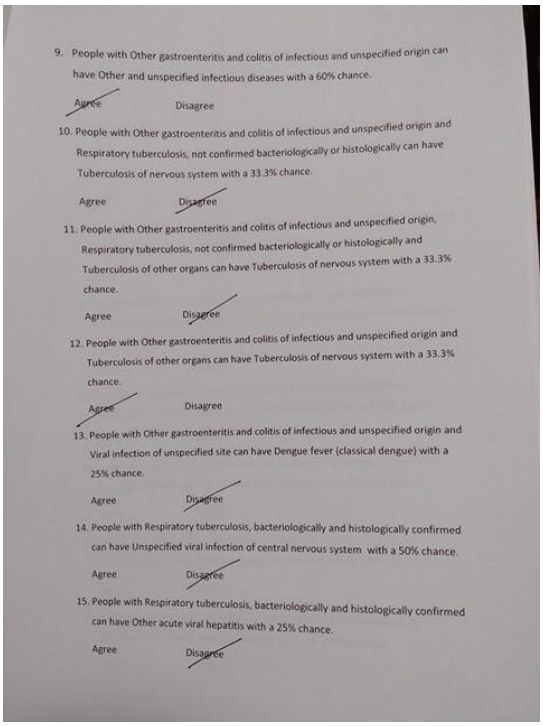
Table AA-4: Frequent Combinations Generated by Bayes Net

Combination	SUPPORT
A09, A02	0.014184397
A09, A03	0.014184397
A09, A90	0.014184397
A09, B09	0.028368794
A09, B24	0.014184397

A09, B34	0.063829787
A09, B99	0.021276596
A16, A15	0.035460993
A16, A18	0.035460993
A16, A31	0.014184397
A16, B02	0.014184397
A16, B18	0.014184397
A18, A19	0.014184397
A41, A04	0.014184397
A41, A09	0.056737589
A41, A16	0.09929078
A41, A19	0.014184397
A41, A49	0.056737589
A41, B02	0.014184397
A41, B18	0.014184397
A41, B37	0.042553191
A41, B49	0.021276596
A41, B96	0.078014184
A49, B95	0.014184397
B18, B16	0.014184397
B34, A90	0.014184397
A04	0.021276596
A09	0.304964539
A15	0.042553191
A16	0.170212766
A18	0.014184397
A19	0.021276596
A49	0.035460993
B17	0.014184397
B18	0.028368794

B34	0.028368794
B37	0.028368794
B49	0.014184397
B95	0.028368794
B96	0.028368794
B99	0.014184397

A.3. Doctors Survey



9. People with Other gastroenteritis and colitis of infectious and unspecified origin can have Other and unspecified infectious diseases with a 60% chance.

Agree Disagree

10. People with Other gastroenteritis and colitis of infectious and unspecified origin and Respiratory tuberculosis, not confirmed bacteriologically or histologically can have Tuberculosis of nervous system with a 33.3% chance.

Agree Disagree

11. People with Other gastroenteritis and colitis of infectious and unspecified origin, Respiratory tuberculosis, not confirmed bacteriologically or histologically and Tuberculosis of other organs can have Tuberculosis of nervous system with a 33.3% chance.

Agree Disagree

12. People with Other gastroenteritis and colitis of infectious and unspecified origin and Tuberculosis of other organs can have Tuberculosis of nervous system with a 33.3% chance.

Agree Disagree

13. People with Other gastroenteritis and colitis of infectious and unspecified origin and Viral infection of unspecified site can have Dengue fever (classical dengue) with a 25% chance.

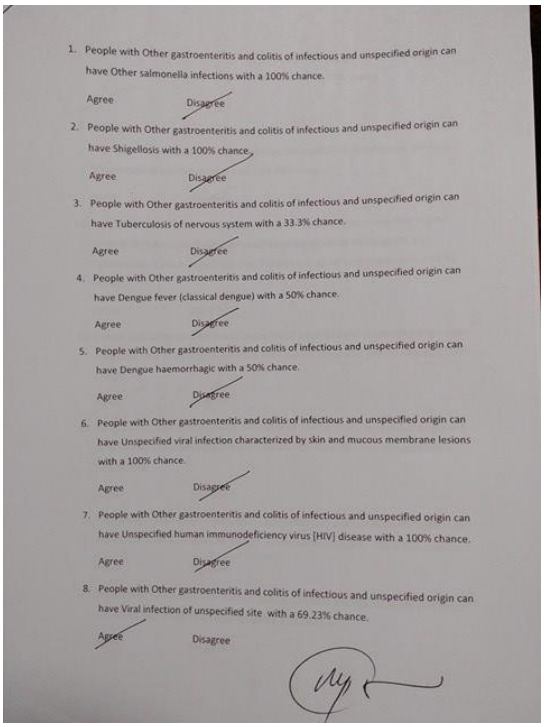
Agree Disagree

14. People with Respiratory tuberculosis, bacteriologically and histologically confirmed can have Unspecified viral infection of central nervous system with a 50% chance.

Agree Disagree

15. People with Respiratory tuberculosis, bacteriologically and histologically confirmed can have Other acute viral hepatitis with a 25% chance.

Agree Disagree



1. People with Other gastroenteritis and colitis of infectious and unspecified origin can have Other salmonella infections with a 100% chance.

Agree Disagree

2. People with Other gastroenteritis and colitis of infectious and unspecified origin can have Shigellosis with a 100% chance.

Agree Disagree

3. People with Other gastroenteritis and colitis of infectious and unspecified origin can have Tuberculosis of nervous system with a 33.3% chance.

Agree Disagree

4. People with Other gastroenteritis and colitis of infectious and unspecified origin can have Dengue fever (classical dengue) with a 50% chance.

Agree Disagree

5. People with Other gastroenteritis and colitis of infectious and unspecified origin can have Dengue haemorrhagic with a 50% chance.

Agree Disagree

6. People with Other gastroenteritis and colitis of infectious and unspecified origin can have Unspecified viral infection characterized by skin and mucous membrane lesions with a 100% chance.

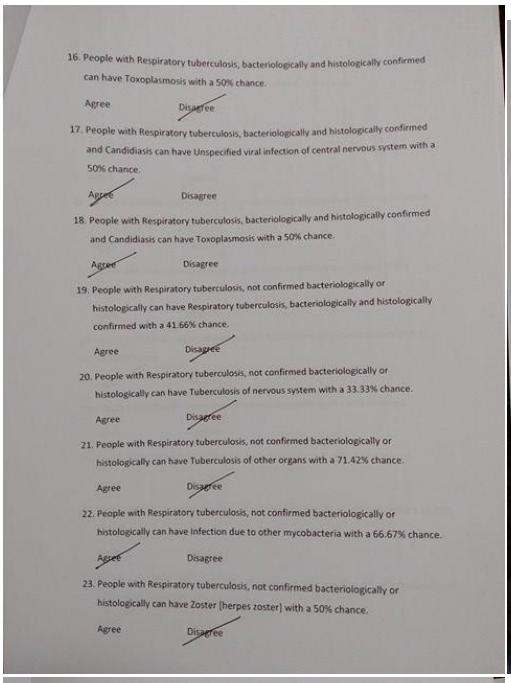
Agree Disagree

7. People with Other gastroenteritis and colitis of infectious and unspecified origin can have Unspecified human immunodeficiency virus [HIV] disease with a 100% chance.

Agree Disagree

8. People with Other gastroenteritis and colitis of infectious and unspecified origin can have Viral infection of unspecified site with a 69.23% chance.

Agree Disagree



16. People with Respiratory tuberculosis, bacteriologically and histologically confirmed can have Toxoplasmosis with a 50% chance.

Agree Disagree

17. People with Respiratory tuberculosis, bacteriologically and histologically confirmed and Candidiasis can have Unspecified viral infection of central nervous system with a 50% chance.

Agree Disagree

18. People with Respiratory tuberculosis, bacteriologically and histologically confirmed and Candidiasis can have Toxoplasmosis with a 50% chance.

Agree Disagree

19. People with Respiratory tuberculosis, not confirmed bacteriologically or histologically can have Respiratory tuberculosis, bacteriologically and histologically confirmed with a 41.66% chance.

Agree Disagree

20. People with Respiratory tuberculosis, not confirmed bacteriologically or histologically can have Tuberculosis of nervous system with a 33.33% chance.

Agree Disagree

21. People with Respiratory tuberculosis, not confirmed bacteriologically or histologically can have Tuberculosis of other organs with a 71.42% chance.

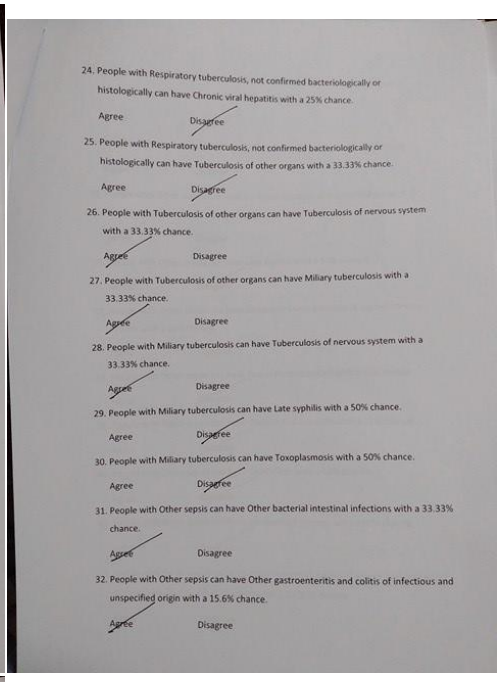
Agree Disagree

22. People with Respiratory tuberculosis, not confirmed bacteriologically or histologically can have Infection due to other mycobacteria with a 66.67% chance.

Agree Disagree

23. People with Respiratory tuberculosis, not confirmed bacteriologically or histologically can have Zoster [herpes zoster] with a 50% chance.

Agree Disagree



24. People with Respiratory tuberculosis, not confirmed bacteriologically or histologically can have Chronic viral hepatitis with a 25% chance.

Agree Disagree

25. People with Respiratory tuberculosis, not confirmed bacteriologically or histologically can have Tuberculosis of other organs with a 33.33% chance.

Agree Disagree

26. People with Tuberculosis of other organs can have Tuberculosis of nervous system with a 33.33% chance.

Agree Disagree

27. People with Tuberculosis of other organs can have Miliary tuberculosis with a 33.33% chance.

Agree Disagree

28. People with Miliary tuberculosis can have Tuberculosis of nervous system with a 33.33% chance.

Agree Disagree

29. People with Miliary tuberculosis can have Late syphilis with a 50% chance.

Agree Disagree

30. People with Miliary tuberculosis can have Toxoplasmosis with a 50% chance.

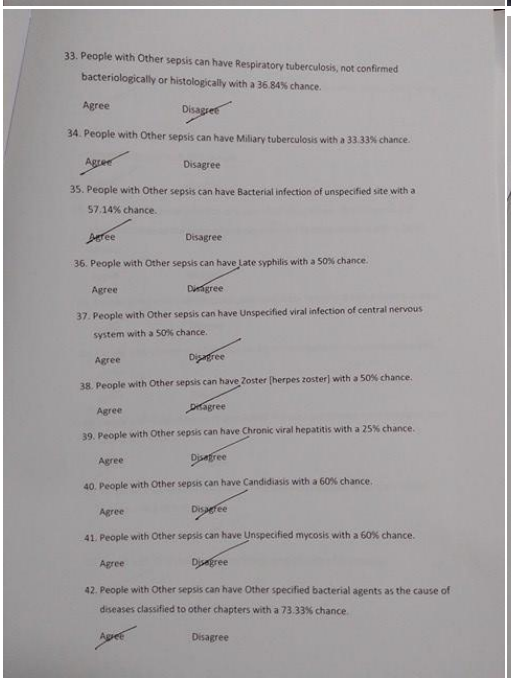
Agree Disagree

31. People with Other sepsis can have Other bacterial intestinal infections with a 33.33% chance.

Agree Disagree

32. People with Other sepsis can have Other gastroenteritis and colitis of infectious and unspecified origin with a 15.6% chance.

Agree Disagree



33. People with Other sepsis can have Respiratory tuberculosis, not confirmed bacteriologically or histologically with a 36.84% chance.

Agree Disagree

34. People with Other sepsis can have Miliary tuberculosis with a 33.33% chance.

Agree Disagree

35. People with Other sepsis can have Bacterial infection of unspecified site with a 57.14% chance.

Agree Disagree

36. People with Other sepsis can have Late syphilis with a 50% chance.

Agree Disagree

37. People with Other sepsis can have Unspecified viral infection of central nervous system with a 50% chance.

Agree Disagree

38. People with Other sepsis can have Zoster [herpes zoster] with a 50% chance.

Agree Disagree

39. People with Other sepsis can have Chronic viral hepatitis with a 25% chance.

Agree Disagree

40. People with Other sepsis can have Candidiasis with a 60% chance.

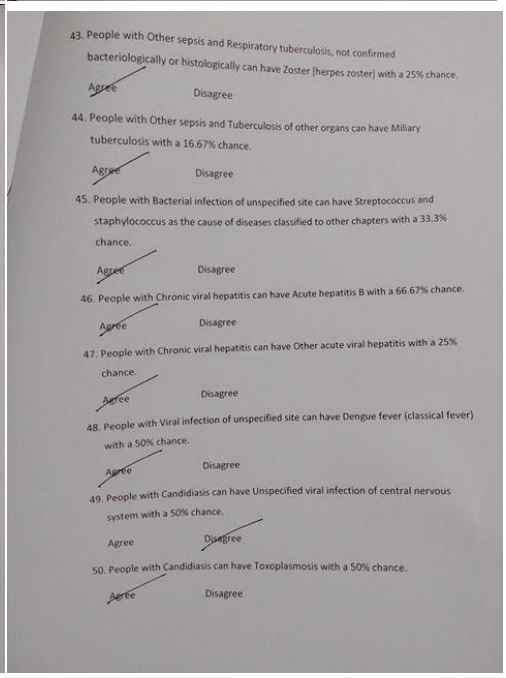
Agree Disagree

41. People with Other sepsis can have Unspecified mycosis with a 60% chance.

Agree Disagree

42. People with Other sepsis can have Other specified bacterial agents as the cause of diseases classified to other chapters with a 73.33% chance.

Agree Disagree



43. People with Other sepsis and Respiratory tuberculosis, not confirmed bacteriologically or histologically can have Zoster [herpes zoster] with a 25% chance.

Agree Disagree

44. People with Other sepsis and Tuberculosis of other organs can have Miliary tuberculosis with a 16.67% chance.

Agree Disagree

45. People with Bacterial infection of unspecified site can have Streptococcus and staphylococcus as the cause of diseases classified to other chapters with a 33.3% chance.

Agree Disagree

46. People with Chronic viral hepatitis can have Acute hepatitis B with a 66.67% chance.

Agree Disagree

47. People with Chronic viral hepatitis can have Other acute viral hepatitis with a 25% chance.

Agree Disagree

48. People with Viral infection of unspecified site can have Dengue fever (classical fever) with a 50% chance.

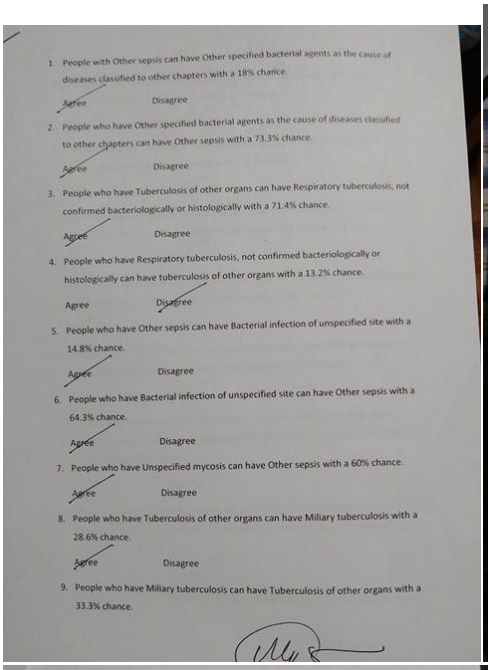
Agree Disagree

49. People with Candidiasis can have Unspecified viral infection of central nervous system with a 50% chance.

Agree Disagree

50. People with Candidiasis can have Toxoplasmosis with a 50% chance.

Agree Disagree



1. People with Other sepsis can have Other specified bacterial agents as the cause of diseases classified to other chapters with a 18% chance.

Agree Disagree

2. People who have Other specified bacterial agents as the cause of diseases classified to other chapters can have Other sepsis with a 73.3% chance.

Agree Disagree

3. People who have Tuberculosis of other organs can have Respiratory tuberculosis, not confirmed bacteriologically or histologically with a 71.4% chance.

Agree Disagree

4. People who have Respiratory tuberculosis, not confirmed bacteriologically or histologically can have tuberculosis of other organs with a 13.2% chance.

Agree Disagree

5. People who have Other sepsis can have Bacterial infection of unspecified site with a 14.8% chance.

Agree Disagree

6. People who have Bacterial infection of unspecified site can have Other sepsis with a 64.3% chance.

Agree Disagree

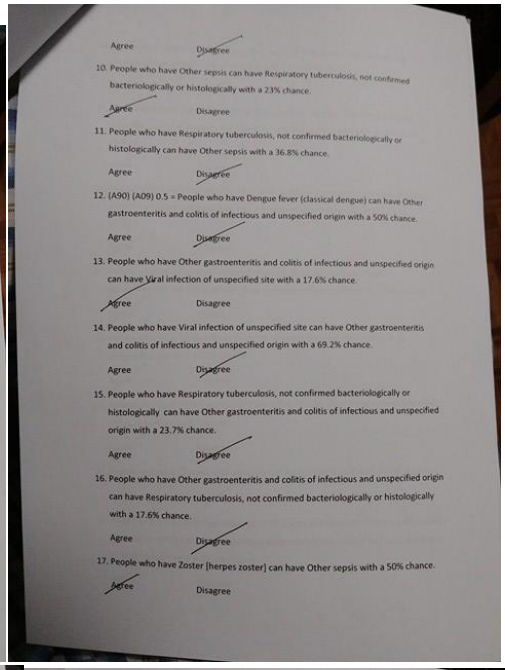
7. People who have Unspecified mycosis can have Other sepsis with a 60% chance.

Agree Disagree

8. People who have Tuberculosis of other organs can have Miliary tuberculosis with a 28.6% chance.

Agree Disagree

9. People who have Miliary tuberculosis can have Tuberculosis of other organs with a 33.3% chance.



Agree Disagree

10. People who have Other sepsis can have Respiratory tuberculosis, not confirmed bacteriologically or histologically with a 23% chance.

Agree Disagree

11. People who have Respiratory tuberculosis, not confirmed bacteriologically or histologically can have Other sepsis with a 36.8% chance.

Agree Disagree

12. (A90) (A09) 0.5 = People who have Dengue fever (classical dengue) can have Other gastroenteritis and colitis of infectious and unspecified origin with a 50% chance.

Agree Disagree

13. People who have Other gastroenteritis and colitis of infectious and unspecified origin can have Viral infection of unspecified site with a 17.6% chance.

Agree Disagree

14. People who have Viral infection of unspecified site can have Other gastroenteritis and colitis of infectious and unspecified origin with a 69.2% chance.

Agree Disagree

15. People who have Respiratory tuberculosis, not confirmed bacteriologically or histologically can have Other gastroenteritis and colitis of infectious and unspecified origin with a 23.7% chance.

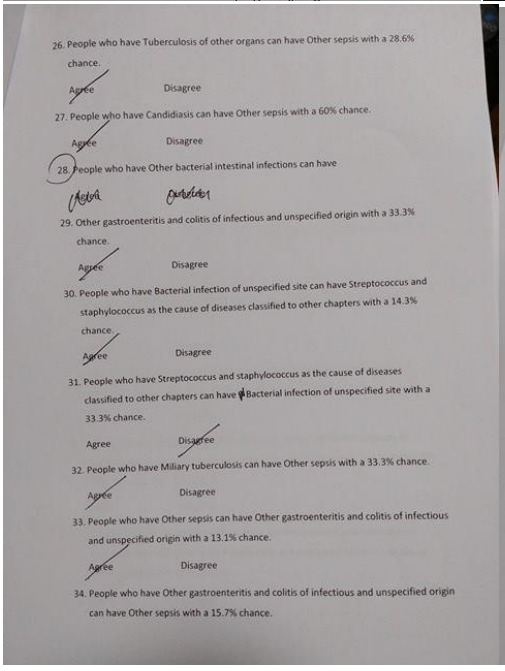
Agree Disagree

16. People who have Other gastroenteritis and colitis of infectious and unspecified origin can have Respiratory tuberculosis, not confirmed bacteriologically or histologically with a 17.6% chance.

Agree Disagree

17. People who have Zoster [herpes zoster] can have Other sepsis with a 50% chance.

Agree Disagree



26. People who have Tuberculosis of other organs can have Other sepsis with a 28.6% chance.

Agree Disagree

27. People who have Candidiasis can have Other sepsis with a 60% chance.

Agree Disagree

28. People who have Other bacterial intestinal infections can have

Agree Disagree

29. Other gastroenteritis and colitis of infectious and unspecified origin with a 33.3% chance.

Agree Disagree

30. People who have Bacterial infection of unspecified site can have Streptococcus and staphylococcus as the cause of diseases classified to other chapters with a 14.3% chance.

Agree Disagree

31. People who have Streptococcus and staphylococcus as the cause of diseases classified to other chapters can have Bacterial infection of unspecified site with a 33.3% chance.

Agree Disagree

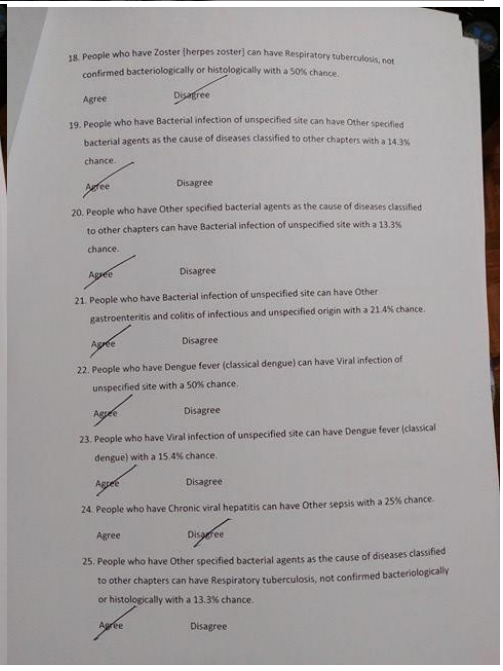
32. People who have Miliary tuberculosis can have Other sepsis with a 33.3% chance.

Agree Disagree

33. People who have Other sepsis can have Other gastroenteritis and colitis of infectious and unspecified origin with a 13.1% chance.

Agree Disagree

34. People who have Other gastroenteritis and colitis of infectious and unspecified origin can have Other sepsis with a 15.7% chance.



18. People who have Zoster [herpes zoster] can have Respiratory tuberculosis, not confirmed bacteriologically or histologically with a 50% chance.

Agree Disagree

19. People who have Bacterial infection of unspecified site can have Other specified bacterial agents as the cause of diseases classified to other chapters with a 14.3% chance.

Agree Disagree

20. People who have Other specified bacterial agents as the cause of diseases classified to other chapters can have Bacterial infection of unspecified site with a 13.3% chance.

Agree Disagree

21. People who have Bacterial infection of unspecified site can have Other gastroenteritis and colitis of infectious and unspecified origin with a 21.4% chance.

Agree Disagree

22. People who have Dengue fever (classical dengue) can have Viral infection of unspecified site with a 50% chance.

Agree Disagree

23. People who have Viral infection of unspecified site can have Dengue fever [classical dengue) with a 15.4% chance.

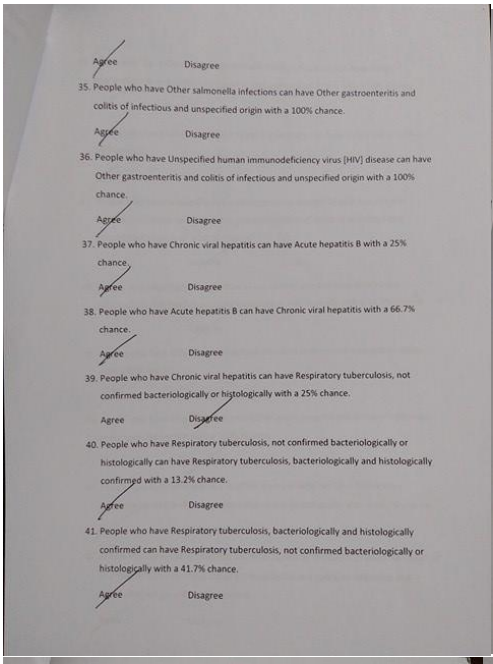
Agree Disagree

24. People who have Chronic viral hepatitis can have Other sepsis with a 25% chance.

Agree Disagree

25. People who have Other specified bacterial agents as the cause of diseases classified to other chapters can have Respiratory tuberculosis, not confirmed bacteriologically or histologically with a 13.3% chance.

Agree Disagree



Agree Disagree

35. People who have Other salmonella infections can have Other gastroenteritis and colitis of infectious and unspecified origin with a 100% chance.

Agree Disagree

36. People who have Unspecified human immunodeficiency virus [HIV] disease can have Other gastroenteritis and colitis of infectious and unspecified origin with a 100% chance.

Agree Disagree

37. People who have Chronic viral hepatitis can have Acute hepatitis B with a 25% chance.

Agree Disagree

38. People who have Acute hepatitis B can have Chronic viral hepatitis with a 66.7% chance.

Agree Disagree

39. People who have Chronic viral hepatitis can have Respiratory tuberculosis, not confirmed bacteriologically or histologically with a 25% chance.

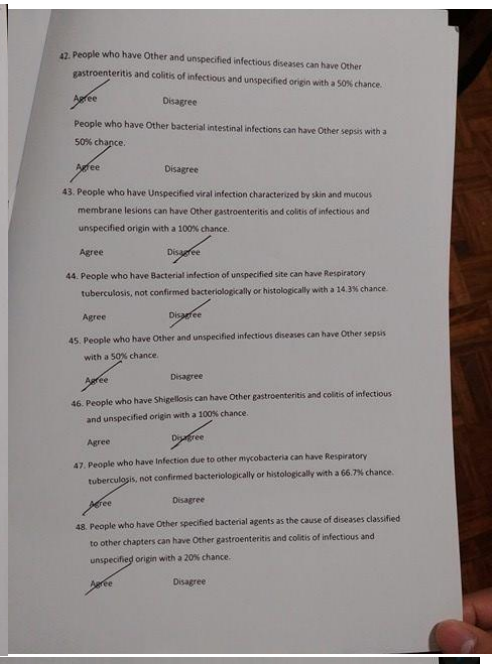
Agree Disagree

40. People who have Respiratory tuberculosis, not confirmed bacteriologically or histologically can have Respiratory tuberculosis, bacteriologically and histologically confirmed with a 13.2% chance.

Agree Disagree

41. People who have Respiratory tuberculosis, bacteriologically and histologically confirmed can have Respiratory tuberculosis, not confirmed bacteriologically or histologically with a 41.7% chance.

Agree Disagree



42. People who have Other and unspecified infectious diseases can have Other gastroenteritis and colitis of infectious and unspecified origin with a 50% chance.

Agree Disagree

People who have Other bacterial intestinal infections can have Other sepsis with a 50% chance.

Agree Disagree

43. People who have Unspecified viral infection characterized by skin and mucous membrane lesions can have Other gastroenteritis and colitis of infectious and unspecified origin with a 100% chance.

Agree Disagree

44. People who have Bacterial infection of unspecified site can have Respiratory tuberculosis, not confirmed bacteriologically or histologically with a 14.3% chance.

Agree Disagree

45. People who have Other and unspecified infectious diseases can have Other sepsis with a 50% chance.

Agree Disagree

46. People who have Shigellosis can have Other gastroenteritis and colitis of infectious and unspecified origin with a 100% chance.

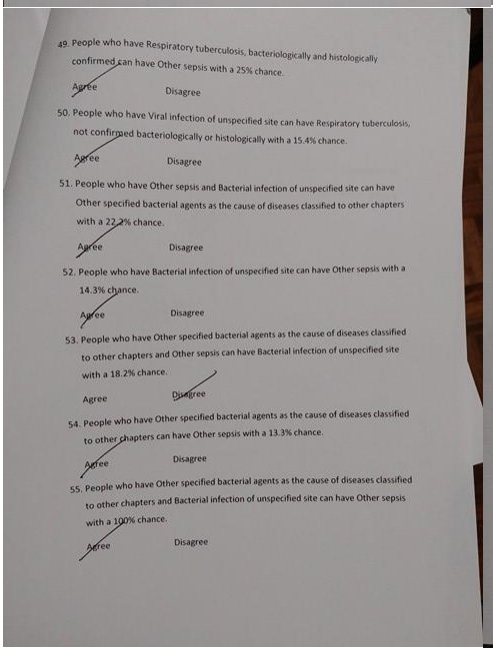
Agree Disagree

47. People who have Infection due to other mycobacteria can have Respiratory tuberculosis, not confirmed bacteriologically or histologically with a 66.7% chance.

Agree Disagree

48. People who have Other specified bacterial agents as the cause of diseases classified to other chapters can have Other gastroenteritis and colitis of infectious and unspecified origin with a 20% chance.

Agree Disagree



49. People who have Respiratory tuberculosis, bacteriologically and histologically confirmed can have Other sepsis with a 25% chance.

Agree Disagree

50. People who have Viral infection of unspecified site can have Respiratory tuberculosis, not confirmed bacteriologically or histologically with a 15.4% chance.

Agree Disagree

51. People who have Other sepsis and Bacterial infection of unspecified site can have Other specified bacterial agents as the cause of diseases classified to other chapters with a 22.2% chance.

Agree Disagree

52. People who have Bacterial infection of unspecified site can have Other sepsis with a 14.3% chance.

Agree Disagree

53. People who have Other specified bacterial agents as the cause of diseases classified to other chapters and Other sepsis can have Bacterial infection of unspecified site with a 18.2% chance.

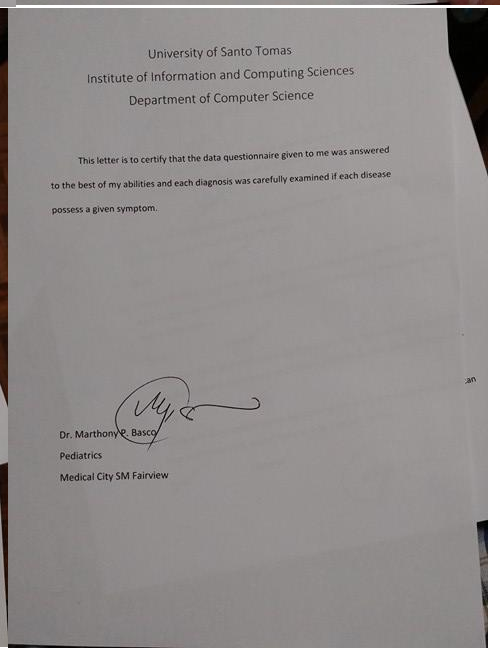
Agree Disagree

54. People who have Other specified bacterial agents as the cause of diseases classified to other chapters can have Other sepsis with a 13.3% chance.

Agree Disagree

55. People who have Other specified bacterial agents as the cause of diseases classified to other chapters and Bacterial infection of unspecified site can have Other sepsis with a 100% chance.

Agree Disagree



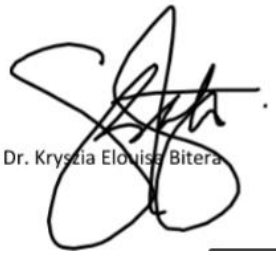
University of Santo Tomas
Institute of Information and Computing Sciences
Department of Computer Science

This letter is to certify that the data questionnaire given to me was answered to the best of my abilities and each diagnosis was carefully examined if each disease possess a given symptom.

Dr. Marthony P. Basco
Pediatrics
Medical City SM Fairview

University of Santo Tomas
Institute of Information and Computing Sciences
Department of Computer Science

This letter is to certify that the data questionnaire given to me was answered to the best of my abilities and each diagnosis was carefully examined if each disease possess a given symptom.



Dr. Kryszia Eloïsa Bitera

1. (A41) (B96) 0.18 = People with Other sepsis can have Other specified bacterial agents as the cause of diseases classified to other chapters with a 18% chance.
   Agree Disagree
2. (B96) (A41) 0.733 = People who have Other specified bacterial agents as the cause of diseases classified to other chapters can have Other sepsis with a 73.3% chance.
   Agree Disagree
3. (A18) (A16) 0.714 = People who have Tuberculosis of other organs can have Respiratory tuberculosis, not confirmed bacteriologically or histologically with a 71.4% chance.
   Agree Disagree
4. (A16) (A18) 0.132 = People who have Respiratory tuberculosis, not confirmed bacteriologically or histologically can have tuberculosis of other organs with a 13.2% chance.
   Agree Disagree
5. (A41) (A49) 0.148 = People who have Other sepsis can have Bacterial infection of unspecified site with a 14.8% chance.
   Agree Disagree
6. (A49) (A41) 0.643 = People who have Bacterial infection of unspecified site can have Other sepsis with a 64.3% chance.
   Agree Disagree
7. (B49) (A41) 0.6 = People who have Unspecified mycosis can have Other sepsis with a 60% chance.
   Agree Disagree
8. (A18) (A19) 0.286 = People who have Tuberculosis of other organs can have Miliary tuberculosis with a 28.6% chance.
   Agree Disagree
9. (A19) (A18) 0.333 = People who have Miliary tuberculosis can have Tuberculosis of other organs with a 33.3% chance.
   Agree Disagree
10. (A41) (A16) 0.23 = People who have Other sepsis can have Respiratory tuberculosis, not confirmed bacteriologically or histologically with a 23% chance.
    Agree Disagree
11. (A16) (A41) 0.368 = People who have Respiratory tuberculosis, not confirmed bacteriologically or histologically can have Other sepsis with a 36.8% chance.
    Agree Disagree
12. (A90) (A09) 0.5 = People who have Dengue fever (classical dengue) can have Other gastroenteritis and colitis of infectious and unspecified origin with a 50% chance.
    Agree Disagree
13. (A09) (B34) 0.176 = People who have Other gastroenteritis and colitis of infectious and unspecified origin can have Viral infection of unspecified site with a 17.6% chance.
    Agree Disagree
14. (B34) (A09) 0.692 = People who have Viral infection of unspecified site can have Other gastroenteritis and colitis of infectious and unspecified origin with a 69.2% chance.
    Agree Disagree
15. (A16) (A09) 0.237 = People who have Respiratory tuberculosis, not confirmed bacteriologically or histologically can have Other gastroenteritis and colitis of infectious and unspecified origin with a 23.7% chance.
    Agree Disagree

16. (A09) (A16) 0.176 = People who have Other gastroenteritis and colitis of infectious and unspecified origin can have Respiratory tuberculosis, not confirmed bacteriologically or histologically with a 17.6% chance.

Agree Disagree

17. (B02) (A41) 0.5 = People who have Zoster [herpes zoster] can have Other sepsis with a 50% chance.

Agree Disagree

18. (B02) (A16) 0.5 = People who have Zoster [herpes zoster] can have Respiratory tuberculosis, not confirmed bacteriologically or histologically with a 50% chance.

Agree Disagree

19. (A49) (B96) 0.143 = People who have Bacterial infection of unspecified site can have Other specified bacterial agents as the cause of diseases classified to other chapters with a 14.3% chance.

Agree Disagree

20. (B96) (A49) 0.133 = People who have Other specified bacterial agents as the cause of diseases classified to other chapters can have Bacterial infection of unspecified site with a 13.3% chance.

Agree Disagree

21. (A49) (A09) 0.214 = People who have Bacterial infection of unspecified site can have Other gastroenteritis and colitis of infectious and unspecified origin with a 21.4% chance.

Agree Disagree

22. (A90) (B34) 0.5 = People who have Dengue fever (classical dengue) can have Viral infection of unspecified site with a 50% chance.

Agree Disagree

23. (B34) (A90) 0.154 = People who have Viral infection of unspecified site can have Dengue fever (classical dengue) with a 15.4% chance.

Agree Disagree

24. (B18) (A41) 0.25 = People who have Chronic viral hepatitis can have Other sepsis with a 25% chance.

Agree Disagree

25. (B96) (A16) 0.133 = People who have Other specified bacterial agents as the cause of diseases classified to other chapters can have Respiratory tuberculosis, not confirmed bacteriologically or histologically with a 13.3% chance.

Agree Disagree

26. (A18) (A41) 0.286 = People who have Tuberculosis of other organs can have Other sepsis with a 28.6% chance.

Agree Disagree

27. (B37) (A41) 0.6 = People who have Candidiasis can have Other sepsis with a 60% chance.

Agree Disagree

28. (A04) (A09) 0.333 = People who have Other bacterial intestinal infections can have

Agree Disagree

29. Other gastroenteritis and colitis of infectious and unspecified origin with a 33.3% chance.

Agree Disagree

30. (A49) (B95) 0.143 = People who have Bacterial infection of unspecified site can have Streptococcus and staphylococcus as the cause of diseases classified to other chapters with a 14.3% chance.

Agree Disagree

31. (B95) (A49) 0.333 = People who have Streptococcus and staphylococcus as the cause of diseases classified to other chapters can have 9 Bacterial infection of unspecified site with a 33.3% chance.

Agree Disagree

32. (A19) (A41) 0.333 = People who have Miliary tuberculosis can have Other sepsis with a 33.3% chance.

Agree Disagree

33. (A41) (A09) 0.131 = People who have Other sepsis can have Other gastroenteritis and colitis of infectious and unspecified origin with a 13.1% chance.

Agree Disagree

34. (A09) (A41) 0.157 = People who have Other gastroenteritis and colitis of infectious and unspecified origin can have Other sepsis with a 15.7% chance.

Agree Disagree

35. (A02) (A09) 1 = People who have Other salmonella infections can have Other gastroenteritis and colitis of infectious and unspecified origin with a 100% chance.

Agree Disagree

36. (B24) (A09) 1 = People who have Unspecified human immunodeficiency virus [HIV] disease can have Other gastroenteritis and colitis of infectious and unspecified origin with a 100% chance.

Agree Disagree

37. (B18) (B16) 0.25 = People who have Chronic viral hepatitis can have Acute hepatitis B with a 25% chance.

Agree Disagree

38. (B16) (B18) 0.667 = People who have Acute hepatitis B can have Chronic viral hepatitis with a 66.7% chance.

Agree Disagree

39. (B18) (A16) 0.25 = People who have Chronic viral hepatitis can have Respiratory tuberculosis, not confirmed bacteriologically or histologically with a 25% chance.

Agree Disagreeq1

40. (A16) (A15) 0.132 = People who have Respiratory tuberculosis, not confirmed bacteriologically or histologically can have Respiratory tuberculosis, bacteriologically and histologically confirmed with a 13.2% chance.

Agree Disagree

41. (A15) (A16) 0.417 = People who have Respiratory tuberculosis, bacteriologically and histologically confirmed can have Respiratory tuberculosis, not confirmed bacteriologically or histologically with a 41.7% chance.

Agree Disagree

42. (B99) (A09) 0.5 = People who have Other and unspecified infectious diseases can have Other gastroenteritis and colitis of infectious and unspecified origin with a 50% chance.

Agree Disagree

43. (A04) (A41) 0.5 = People who have Other bacterial intestinal infections can have Other sepsis with a 50% chance.

Agree Disagree

44. (B09) (A09) 1 = People who have Unspecified viral infection characterized by skin and mucous membrane lesions can have Other gastroenteritis and colitis of infectious and unspecified origin with a 100% chance.

Agree Disagree

45. (A49) (A16) 0.143 = People who have Bacterial infection of unspecified site can have Respiratory tuberculosis, not confirmed bacteriologically or histologically with a 14.3% chance.

Agree Disagree

46. (B99) (A41) 0.5 = People who have Other and unspecified infectious diseases can have Other sepsis with a 50% chance.

Agree Disagree

47. (A03) (A09) 1 = People who have Shigellosis can have Other gastroenteritis and colitis of infectious and unspecified origin with a 100% chance.

Agree Disagree

48. (A31) (A16) 0.667 = People who have Infection due to other mycobacteria can have Respiratory tuberculosis, not confirmed bacteriologically or histologically with a 66.7% chance.

Agree Disagree

49. (B96) (A09) 0.2 = People who have Other specified bacterial agents as the cause of diseases classified to other chapters can have Other gastroenteritis and colitis of infectious and unspecified origin with a 20% chance.

Agree Disagree

50. (A15) (A41) 0.25 = People who have Respiratory tuberculosis, bacteriologically and histologically confirmed can have Other sepsis with a 25% chance.

Agree Disagree

51. (B34) (A16) 0.154 = People who have Viral infection of unspecified site can have Respiratory tuberculosis, not confirmed bacteriologically or histologically with a 15.4% chance.

Agree Disagree

52. (A41,A49) (B96) 0.222 = People who have Other sepsis and Bacterial infection of unspecified site can have Other specified bacterial agents as the cause of diseases classified to other chapters with a 22.2% chance.

Agree Disagree

53. (A49) (A41) 0.143 = People who have Bacterial infection of unspecified site can have Other sepsis with a 14.3% chance.

Agree Disagree

54. (B96,A41) (A49) 0.182 = People who have Other specified bacterial agents as the cause of diseases classified to other chapters and Other sepsis can have Bacterial infection of unspecified site with a 18.2% chance.

Agree Disagree

55. (B96) (A41) 0.133 = People who have Other specified bacterial agents as the cause of diseases classified to other chapters can have Other sepsis with a 13.3% chance.

Agree Disagree

56. (B96,A49) (A41) 1 = People who have Other specified bacterial agents as the cause of diseases classified to other chapters and Bacterial infection of unspecified site can have Other sepsis with a 100% chance.

Agree Disagree

1. A09, A02 1 = People with Other gastroenteritis and colitis of infectious and unspecified origin can have Other salmonella infections with a 100% chance.

   Agree Disagree

2. A09, A03 1 = People with Other gastroenteritis and colitis of infectious and unspecified origin can have Shigellosis with a 100% chance.

   Agree Disagree

3. A09, A17 0.333333333 = People with Other gastroenteritis and colitis of infectious and unspecified origin can have Tuberculosis of nervous system with a 33.3% chance.

   Agree Disagree

4. A09, A90 0.5 = People with Other gastroenteritis and colitis of infectious and unspecified origin can have Dengue fever (classical dengue) with a 50% chance.

   Agree Disagree

5. A09, A91 0.5 = People with Other gastroenteritis and colitis of infectious and unspecified origin can have Dengue haemorrhagic with a 50% chance.

   Agree Disagree

6. A09, B09 1 = People with Other gastroenteritis and colitis of infectious and unspecified origin can have Unspecified viral infection characterized by skin and mucous membrane lesions with a 100% chance.

   Agree Disagree

7. A09, B24 1 = People with Other gastroenteritis and colitis of infectious and unspecified origin can have Unspecified human immunodeficiency virus [HIV] disease with a 100% chance.

   Agree Disagree

8. A09, B34 0.692307692 = People with Other gastroenteritis and colitis of infectious and unspecified origin can have Viral infection of unspecified site with a 69.23% chance.

   Agree Disagree

9. A09, B99 0.6 = People with Other gastroenteritis and colitis of infectious and unspecified origin can have Other and unspecified infectious diseases with a 60% chance.

   Agree Disagree

10. A09, A16, A17 0.333333333 = People with Other gastroenteritis and colitis of infectious and unspecified origin and Respiratory tuberculosis, not confirmed bacteriologically or histologically can have Tuberculosis of nervous system with a 33.3% chance.

    Agree Disagree

11. A09, A16, A18, A17 0.333333333 = People with Other gastroenteritis and colitis of infectious and unspecified origin, Respiratory tuberculosis, not confirmed bacteriologically or histologically and Tuberculosis of other organs can have Tuberculosis of nervous system with a 33.3% chance.

    Agree Disagree

12. A09, A18, A17 0.333333333 = People with Other gastroenteritis and colitis of infectious and unspecified origin and Tuberculosis of other organs can have Tuberculosis of nervous system with a 33.3% chance.

    Agree Disagree

13. A09, B34, A90 0.25 = People with Other gastroenteritis and colitis of infectious and unspecified origin and Viral infection of unspecified site can have Dengue fever (classical dengue) with a 25% chance.

    Agree Disagree

14. A15, A89 0.5 = People with Respiratory tuberculosis, bacteriologically and histologically confirmed can have Unspecified viral infection of central nervous system with a 50% chance.

Agree Disagree

15. A15, B17 0.25 = People with Respiratory tuberculosis, bacteriologically and histologically confirmed can have Other acute viral hepatitis with a 25% chance.

Agree Disagree

16. A15, B58 0.5 = People with Respiratory tuberculosis, bacteriologically and histologically confirmed can have Toxoplasmosis with a 50% chance.

Agree Disagree

17. A15, B37, A89 0.5 = People with Respiratory tuberculosis, bacteriologically and histologically confirmed and Candidiasis can have Unspecified viral infection of central nervous system with a 50% chance.

Agree Disagree

18. A15, B37, B58 0.5 = People with Respiratory tuberculosis, bacteriologically and histologically confirmed and Candidiasis can have Toxoplasmosis with a 50% chance.

Agree Disagree

19. A16, A15 0.416666667 = People with Respiratory tuberculosis, not confirmed bacteriologically or histologically can have Respiratory tuberculosis, bacteriologically and histologically confirmed with a 41.66% chance.

Agree Disagree

20. A16, A17 0.333333333 = People with Respiratory tuberculosis, not confirmed bacteriologically or histologically can have Tuberculosis of nervous system with a 33.33% chance.

Agree Disagree

21. A16, A18 0.714285714 = People with Respiratory tuberculosis, not confirmed bacteriologically or histologically can have Tuberculosis of other organs with a 71.42% chance.

Agree Disagree

22. A16, A31 0.666666667 = People with Respiratory tuberculosis, not confirmed bacteriologically or histologically can have Infection due to other mycobacteria with a 66.67% chance.

Agree Disagree

23. A16, B02 0.5 = People with Respiratory tuberculosis, not confirmed bacteriologically or histologically can have Zoster [herpes zoster] with a 50% chance.

Agree Disagree

24. A16, B18 0.25 = People with Respiratory tuberculosis, not confirmed bacteriologically or histologically can have Chronic viral hepatitis with a 25% chance.

Agree Disagree

25. A16, A18, A17 0.333333333 = People with Respiratory tuberculosis, not confirmed bacteriologically or histologically can have Tuberculosis of other organs with a 33.33% chance.

Agree Disagree

26. A18, A17 0.333333333 = People with Tuberculosis of other organs can have Tuberculosis of nervous system with a 33.33% chance.

Agree Disagree

27. A18, A19 0.333333333 = People with Tuberculosis of other organs can have Miliary tuberculosis with a 33.33% chance.

Agree Disagree

28. A19, A17 0.333333333 = People with Miliary tuberculosis can have Tuberculosis of nervous system with a 33.33% chance.

Agree Disagree

29. A19, A52 0.5 = People with Miliary tuberculosis can have Late syphilis with a 50% chance.

Agree Disagree

30. A19, B58 0.5 = People with Miliary tuberculosis can have Toxoplasmosis with a 50% chance.

Agree Disagree

31. A41, A04 0.333333333 = People with Other sepsis can have Other bacterial intestinal infections with a 33.33% chance.

Agree Disagree

32. A41, A09 0.156862745 = People with Other sepsis can have Other gastroenteritis and colitis of infectious and unspecified origin with a 15.6% chance.

Agree Disagree

33. A41, A16 0.368421053 = People with Other sepsis can have Respiratory tuberculosis, not confirmed bacteriologically or histologically with a 36.84% chance.

Agree Disagree

34. A41, A19 0.333333333 = People with Other sepsis can have Miliary tuberculosis with a 33.33% chance.

Agree Disagree

35. A41, A49 0.571428571 = People with Other sepsis can have Bacterial infection of unspecified site with a 57.14% chance.

Agree Disagree

36. A41, A52 0.5 = People with Other sepsis can have Late syphilis with a 50% chance.

Agree Disagree

37. A41, A89 0.5 = People with Other sepsis can have Unspecified viral infection of central nervous system with a 50% chance.

Agree Disagree

38. A41, B02 0.5 = People with Other sepsis can have Zoster [herpes zoster] with a 50% chance.

Agree Disagree

39. A41, B18 0.25 = People with Other sepsis can have Chronic viral hepatitis with a 25% chance.

Agree Disagree

40. A41, B37 0.6 = People with Other sepsis can have Candidiasis with a 60% chance.

Agree Disagree

41. A41, B49 0.6 = People with Other sepsis can have Unspecified mycosis with a 60% chance.

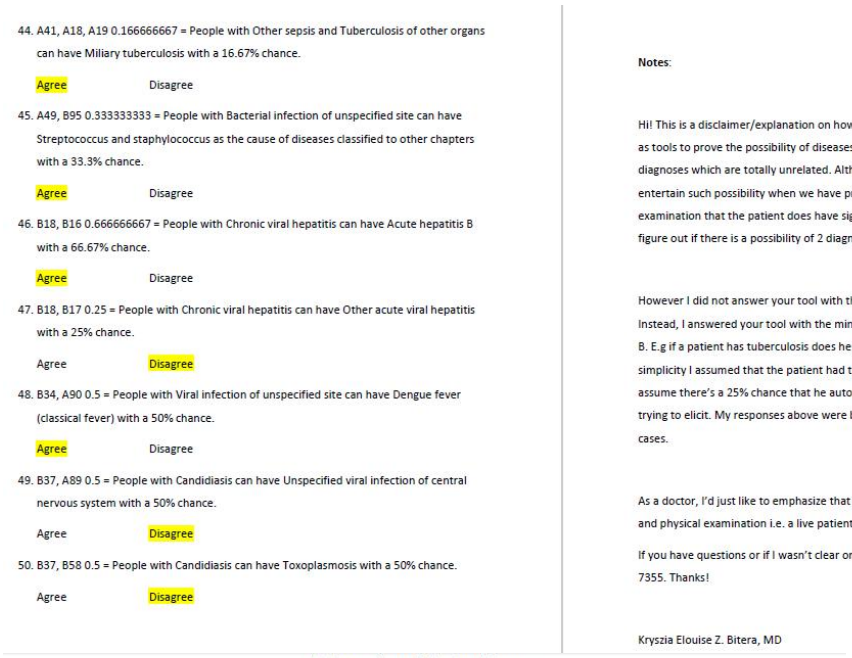
Agree Disagree

42. A41, B96 0.733333333 = People with Other sepsis can have Other specified bacterial agents as the cause of diseases classified to other chapters with a 73.33% chance.

Agree Disagree

43. A41, A16, B02 0.25 = People with Other sepsis and Respiratory tuberculosis, not confirmed bacteriologically or histologically can have Zoster [herpes zoster] with a 25% chance.

Agree Disagree



44. A41, A18, A19 0.166666667 = People with Other sepsis and Tuberculosis of other organs can have Miliary tuberculosis with a 16.67% chance.

Agree Disagree

45. A49, B95 0.333333333 = People with Bacterial infection of unspecified site can have Streptococcus and staphylococcus as the cause of diseases classified to other chapters with a 33.3% chance.

Agree Disagree

46. B18, B16 0.666666667 = People with Chronic viral hepatitis can have Acute hepatitis B with a 66.67% chance.

Agree Disagree

47. B18, B17 0.25 = People with Chronic viral hepatitis can have Other acute viral hepatitis with a 25% chance.

Agree Disagree

48. B34, A90 0.5 = People with Viral infection of unspecified site can have Dengue fever (classical fever) with a 50% chance.

Agree Disagree

49. B37, A89 0.5 = People with Candidiasis can have Unspecified viral infection of central nervous system with a 50% chance.

Agree Disagree

50. B37, B58 0.5 = People with Candidiasis can have Toxoplasmosis with a 50% chance.

Agree Disagree

Notes:

Hi! This is a disclaimer/explanation on how I accomplished the tool. I understood the questions as tools to prove the possibility of diseases coexisting. As a general rule, patients can have two diagnoses which are totally unrelated. Although unrelated, e.g tuberculosis and hepatitis, we entertain such possibility when we have proven from the patient's history and physical examination that the patient does have signs and symptoms. Therefore, if your tool is trying to figure out if there is a possibility of 2 diagnoses for patients, then there will always be a chance.

However I did not answer your tool with that mindset or else my responses will all be "Agree". Instead, I answered your tool with the mindset of Diagnosis A ***automatically*** entailing Diagnosis B. E.g if a patient has tuberculosis does he ***automatically*** have hepatitis with a 25% chance. For simplicity I assumed that the patient had tuberculosis symptoms only, and I am asked if I am to assume there's a 25% chance that he automatically has hepatitis. I hope that's what you're trying to elicit. My responses above were based on my personal experience and knowledge of cases.

As a doctor, I'd just like to emphasize that proper diagnosis is best achieved by history taking and physical examination i.e. a live patient-doctor encounter ☺

If you have questions or if I wasn't clear on making my statement you can call me 0949 412 7355. Thanks!

Kryszia Elouise Z. Bitera, MD

University of Santo Tomas

Institute of Information and Computing Sciences

Department of Computer Science

This letter is to certify that the data questionnaire given to me was answered to the best of my abilities and each diagnosis was carefully examined if each disease possess a given symptom.

Dr. Marisol R. Reynon, MD

A.4. Summary of Results of Doctor's Opinions

Table AA-5: BayesNet Association with Doctors' Opinion

PREMISE	CONCLUSION	CONFIDENCE	DOCTOR1	DOCTOR2	DOCTOR3
A09,	A02	1	0	1	0
A09,	A03	1	0	1	0

A09,	A17	0.333333333	0	0	0
A09,	A90	0.5	0	1	0
A09,	A91	0.5	0	1	0
A09,	B09	1	0	1	0
A09,	B24	1	0	1	0
A09,	B34	0.692307692	1	1	1
A09,	B99	0.6	1	1	1
A09, A16,	A17	0.333333333	0	1	0
A09, A16, A18,	A17	0.333333333	0	1	0
A09, A18,	A17	0.333333333	1	1	0
A09, B34,	A90	0.25	0	1	1
A15,	A89	0.5	0	1	0
A15,	B17	0.25	0	1	0
A15,	B58	0.5	0	0	0
A15, B37,	A89	0.5	1	1	0
A15, B37,	B58	0.5	1	0	0
A16,	A15	0.416666667	0	0	1
A16,	A17	0.333333333	0	0	1
A16,	A18	0.714285714	0	1	0
A16,	A31	0.666666667	1	1	1
A16,	B02	0.5	0	1	0
A16,	B18	0.25	0	0	0
A16, A18,	A17	0.333333333	0	0	1
A18,	A17	0.333333333	1	0	1
A18,	A19	0.333333333	1	1	1
A19,	A17	0.333333333	1	1	0
A19,	A52	0.5	0	0	0
A19,	B58	0.5	0	1	0
A41,	A04	0.333333333	1	1	1
A41,	A09	0.156862745	1	1	1

A41,	A16	0.368421053	0	1	0
A41,	A19	0.333333333	1	1	0
A41,	A49	0.571428571	1	1	1
A41,	A52	0.5	0	0	0
A41,	A89	0.5	0	1	0
A41,	B02	0.5	0	1	0
A41,	B18	0.25	0	1	0
A41,	B37	0.6	0	1	0
A41,	B49	0.6	0	1	0
A41,	B96	0.733333333	1	1	1
A41, A16,	B02	0.25	1	1	0
A41, A18,	A19	0.166666667	1	1	1
A49,	B95	0.333333333	1	1	1
B18,	B16	0.666666667	1	1	1
B18,	B17	0.25	1	1	0
B34,	A90	0.5	1	1	1
B37,	A89	0.5	0	1	0

Table AA-6: Apriori Association Rules with Doctors' Opinions

Premise	Conclusion	Confidence	DOCTOR 1	DOCTOR2	DOCTOR3
(A41)	(B96)	0.18	1	0	1
(B96)	(A41)	0.733	1	0	0
(A18)	(A16)	0.714	1	0	1
(A16)	(A18)	0.132	0	0	1
(A41)	(A49)	0.148	1	0	1
(A49)	(A41)	0.643	1	0	1
(B49)	(A41)	0.6	1	0	0
(A18)	(A19)	0.286	1	0	1
(A19)	(A18)	0.333	0	0	0
(A41)	(A16)	0.23	1	0	0
(A16)	(A41)	0.368	0	0	0

(A90)	(A09)	0.5	0	1	0
(A09)	(B34)	0.176	1	0	1
(B34)	(A09)	0.692	0	1	0
(A16)	(A09)	0.237	0	0	0
(A09)	(A16)	0.176	0	1	0
(B02)	(A41)	0.5	1	1	0
(B02)	(A16)	0.5	0	0	0
(A49)	(B96)	0.143	1	0	1
(B96)	(A49)	0.133	1	0	1
(A49)	(A09)	0.214	1	0	0
(A90)	(B34)	0.5	1	1	1
(B34)	(A90)	0.154	1	0	1
(B18)	(A41)	0.25	0	1	0
(B96)	(A16)	0.133	1	0	1
(A18)	(A41)	0.286	1	0	0
(B37)	(A41)	0.6	1	1	0
(A49)	(B95)	0.143	1	0	1
(B95)	(A49)	0.333	1	0	1
(A19)	(A41)	0.333	0	0	1
(A41)	(A09)	0.131	1	0	1
(A09)	(A41)	0.157	1	0	0
(A02)	(A09)	1	1	1	0
(B24)	(A09)	1	1	1	1
(B18)	(B16)	0.25	1	0	0
(B16)	(B18)	0.667	1	0	0
(B18)	(A16)	0.25	1	0	1
(A16)	(A15)	0.132	0	0	0
(A15)	(A16)	0.417	1	0	0
(B99)	(A09)	0.5	1	1	1
(A04)	(A41)	0.5	1	1	1

(B09)	(A09)	1	1	1	0
(A49)	(A16)	0.143	0	0	0
(B99)	(A41)	0.5	0	1	1
(A03)	(A09)	1	1	1	0
(A31)	(A16)	0.667	0	0	0
(B96)	(A09)	0.2	1	1	1
(A15)	(A41)	0.25	1	1	1
(B34)	(A16)	0.154	1	1	0
(A41,A49)	(B96)	0.222	1	1	0
(A49)	(A41)	0.143	1	1	1
(B96,A41)	(A49)	0.182	1	0	1
(B96)	(A41)	0.133	0	1	0
(B96,A49)	(A41)	1	1	0	1

Table AA-7: Apriori Frequency with Doctors' Opinions

Combination	DOCTOR1	DOCTOR2	DOCTOR3
(B96, A41)	1	0	0
(A16, A18)	1	0	1
(A49, A41)	1	0	1
(B49, A41)	1	0	0
(A19, A18)	1	0	1
(A16, A41)	1	0	0
(A09, A90)	0	1	0
(B34, A09)	1	0	1
(A09, A16)	0	0	0
(B02, A41)	1	1	0
(A16, B02)	0	0	0
(B96, A49)	1	0	1
(A09, A49)	1	0	0
(B34, A90)	1	1	1
(B18, A41)	0	1	0

(B96, A16)	1	0	1
(A18, A41)	1	0	0
(B37, A41)	1	1	0
(B95, A49)	1	0	1
(A19, A41)	0	0	1
(A09, A41)	1	0	1
(A02, A09)	1	1	0
(B24, A09)	1	1	1
(B16, B18)	1	0	0
(A16, B18)	1	0	0
(A15, A16)	0	0	0
(A09, B99)	1	1	1
(A04, A41)	1	1	1
(B09, A09)	1	1	0
(A16, A49)	0	0	0
(B99, A41)	0	1	1
(A03, A09)	1	1	0
(A31, A16)	0	0	0
(B96, A09)	1	1	1
(A15, A41)	1	1	1
(B34, A16)	1	1	0
(B96, A49, A41)	1	1	0

Table AA-8: BayesNet Frequency with Doctors' Opinions

Combination	DOCTOR1	DOCTOR2	DOCTOR3
A09, A02	0	1	0
A09, A03	0	1	0
A09, A90	0	1	0
A09, B09	0	1	0

A09, B24	0	1	0
A09, B34	1	1	1
A09, B99	1	1	1
A16, A15	0	1	0
A16, B18	0	1	0
A16, A31	1	1	1
A16, B02	0	1	0
A18, A19	1	1	1
A41, A04	1	1	1
A41, A09	1	1	1
A41, A16	0	1	0
A41, A19	1	1	0
A41, A49	1	1	1
A41, B02	0	1	0
A41, B18	0	1	0
A41, B37	0	1	0
A41, B49	0	1	0
A41, B96	1	1	1
A49, B95	1	1	1
B18, B16	1	1	1
B34, A90	1	1	1

CLYDE RAVI R. BITERA

9 Dapdap St., Ceris 1 Subdivision, Brgy. Canlubang, Calamba City, Laguna
09399255891
ravibitera@gmail.com

OBJECTIVE

To obtain an intern position in the field of Computer Science related to Data Science

EDUCATION

Bachelor of Science in Computer Science
major in Data Science
Institute of Information and Computing Sciences
University of Santo Tomas, España, Manila, Philippines
Aug 2014- present

Highschool
Santa Rosa Science and Technology School
2010 - 2014

EXPERIENCE

Natural Intelligence Solutions Pte. Ltd. - Lloopp
July 3 - August 11, 2017
Quality Assurance and Testing Department – Internship

Responsibilities

- ☒Created Test Scenarios for each module of the developed product
- ☒Test different scenarios to ensure each module would work as intended

ACADEMIC PROJECT

An Approach to Generate Association Rules in Assisted Medical Diagnosis Using Bayesian Network and Three-Way Decisions (BNTWD)
2018 Jan
Thesis-2

EXTRA-CURRICULAR ACTIVITIES

Coder's Development Circle, University of Santo Tomas
Creatives Staff, A.Y. 2016-2017

Institute of Information and Computing Sciences Student Council (ICSSC), University of Santo Tomas
Creatives Staff, A.Y. 2017-2018

SKILLS AND ABILITIES

- ☒Programming
  - Skilled in programming in Java, Struts2, Hibernate
  - Experienced in Python, PL/SQL
  - Knowledgable in C, Bootstrap, DB2, SQL, HTML5, CSS, php, R-programming
- ☒Computer
  - Experienced in using MS Office, Adobe CC Suite

BRIAN PAUL V. CRISOSTOMO

73 Dona Carmen St. Don Jose Heights, Quezon City
09177830190
bvcrisostomo2@gmail.com

OBJECTIVE To obtain an intern position in the field of Computer Science related to Data Science

EDUCATION

Bachelor of Science in Computer Science major in Data Science
Institute of Information and Computing Sciences
University of Santo Tomas, España, Manila, Philippines
Aug 2014- present

Highschool
Sisters of Mount Carmel Catholic School
2002-2014

CO-CURRICULAR ACTIVITIES

Coder's Development Circle, University of Santo Tomas
Director for Logistics and Events, August 2017- present

- Headed the logistic team of the organization

EXTRA-CURRICULAR ACTIVITIES

U-Hac, Coder's Development Circle, Institute of Information and Computing Sciences, University of Santo Tomas
Usher, December 2016

- Assisted on the event by monitoring the pitching and the presentation of summary video event

AWARDS RECEIVED

Honorable Mention/ Consistent Honor Student
Sisters of Mount Carmel Catholic School
March 2014

9th Placer GDG Codelabs Overall
GDG DevFest
October 2015

COMMUNITY ENGAGEMENT

NSTP, University of Santo Tomas
2015-2016
Volunteer – helped in the feeding program for the community

ACADEMIC PROJECTS

myPavic Registration System
2016-2017
Quality Assurance

An Approach to Generate Association Rules in Assisted Medical Diagnosis Using Bayesian Network and Three-Way Decisions (BNTWD)
2018 Jan
Thesis-2

TRAININGS/SEMINARS ATTENDED

DCX Summer Bootcamp - Analytics
10F, Two Cyberpod Centris
July 22 & 29, 2017

Nokia Univesity Coderetreat
Coder's Development Circle
Roque Ruano Building, UST, rm 49
January 2017

DevCon Summit Developer Future Forward
DevCon Summit
SMX
November 2016

SKILLS & ABILITIES

- Programming:
  - Skilled in programming in Java, Struts2, Hibernate
  - Knowledgable in C, Bootstrap, DB2, SQL, CSS, HTML5
- Computer
  - Experienced in using MS Office, Adobe Photoshop, Adobe Illustrator
- Photography
  - Adobe Lightroom
- Leadership
  - Led and worked in groups of volunteers in organization and class

JAMIL KRISTIAN V. TEODORO

39 Eleuterio Cruz Street, Cruzville Subdivision, Zabarte Road, Novaliches, Quezon City
+63 905 207 9816
eodorojamil@yahoo.com

EDUCATION

Bachelor of Science in Computer Science
major in Data Science
Institute of Information and Computing Sciences
University of Santo Tomas, España, Manila, Philippines
Aug 2014- present

Preschool to Highschool
Mater Carmeli School of Novaliches
June2003 — Mar2014

EXPERIENCE

UST Electrical Engineering Laboratory
Aug2015 — April2016
Working Scholar
UST Pharmacy Laboratory
Aug 2015 — Dec 2015
Working Scholar
Intern at Digital 2wist Inc.
June 2017 — Aug 2017
Mobile App Developer for Android and iOS

ACADEMIC PROJECT

An Approach to Generate Association Rules in Assisted Medical Diagnosis Using Bayesian Network and Three-Way Decisions (BNTWD)
2018 Jan
Thesis-2

AWARDS RECEIVED

San Pedro Calungsod Awardee (High School)

EXTRA-CURRICULAR ACTIVITIES

Coder's Development Circle
Team Lead for Logistics and Events
AY 2017-2018

System Analyst for RedHourGlass
Developed MyPAVIC for PAVIC (Parent's Association for Visually Impaired Children)

SKILLS AND ABILITIES

- HTML, CSS, Bootstrap, MVC (model 1 & 2)
- Photoshop
- Sony Vegas
- Java
- DB2
- C
- SQL
- Microsoft Office (Word, Excel, PowerPoint)

RACHEL MONIQUE K. TUMULAK

130 Biak na Bato St. Quezon City
09271308021 / 09193625912
rmktumulak@gmail.com

OBJECTIVE

To obtain an intern position in the field of Computer Science related to Data Science

EDUCATION

Massive Open Online Courses (MOOC)
DataCamp-Intro to Python for Data Science (2017)
- Developed foundations of data science applied to Python such as the use of Numpy package

Udacity-Programming Foundations with Python (2017)
- Experienced familiarity with Python by developing mini-projects related to the course

Udemy-PL/SQL Fundamenental I (2017)
- Developed fundamenals of PL/SQLby developing thesis system/algorithm

Bachelor of Science in Computer Science
major in Data Science
Institute of Information and Computing Sciences
University of Santo Tomas, España, Manila, Philippines
Aug 2014- present

Preschool to Highschool
Saint Jude Catholic School
327 Ycaza, San Miguel, Manila, 1005 Metro Manila
2001 - 2014

EXPERIENCE

UnionBank of the Philippines
June 23 - August 8, 2017
Technology Management System Department - Internship
- Documented systems and recommended system improve
- Included in the Student Mentoring Program for interns

TRAININGS/SEMINARS ATTENDED

DCX Summer Bootcamp - Analytics
10F, Two Cyberpod Centris
July 22 & 29, 2017

SOCC LTS
SOCC UST
July 2016

ACADEMIC PROJECT

An Approach to Generate Association Rules in Assisted Medical Diagnosis Using Bayesian Network and Three-Way Decisions (BNTWD)
2018 Jan
Thesis-2

AWARDS RECEIVED

Dean's Lister
Univesity of Santo Tomas, Manila
2014-2015, 1st Semester

Top 10 - Student Mentoring Program
UnionBank of the Philippines
2017 July 30

EXTRA-CURRICULAR ACTIVITIES

Block 3CSF, University of Santo Tomas
President, A.Y. 2016-2017, 2nd Semester
- Leads the class

Coder's Development Circle, University of Santo Tomas
Director for Community Development, A.Y. 2016-2017
- Headed the community engagement of the organization

Nokia University Coderetreat, Coder's Development Circle, Institute of Information and Computing Sciences, University of Santo Tomas
Co-project head, January 2017
- In charged of the design and presentation of the certificates and assisted in production which includes venue and program flow

UnionBank's Hackathon and Convention (U-HAC), Coder's Development Circle, Institute of Information and Computing Sciences, University of Santo Tomas
Usher, December 2016
- Designed and presented the certificates
- Helped in documenting the program

SKILLS AND ABILITIES

- Programming
  - Skilled in programming in Java, Struts2, Hibernate
  - Experienced in Python, PL/SQL
  - Knowledgable in C, Bootstrap, DB2, SQL, HTML5, CSS, php, R-programming
- Computer
  - Experienced in using MS Office, Adobe Photoshop, Adobe Illustrator, Adobe Flash
- Leadership
  - Led and worked in groups of volunteers in organization and class
- Language
  - Basic speaking, reading, and writing skills in Chinese Mandarin