iaes
Institute of Advanced Engineering and Science

Using Naïve Bayes and Bayesian Network for Prediction of Potential Problematic Cases in Tuberculosis

Awad Ali*, Moawia Elfaki, Dayang N.A. Jawawi***
* Departement of Software Engineering,UniversitiTeknologi Malaysia
**College of Computer Science& IT,King Faisal University (Saudi Arabia), University of Khartoum (Sudan).

Article Info

Article history:

Received Sept 10th, 2012
Accepted Oct 23th, 2012

Keyword:

Naïve Bayes
Bayesian Networks
Machine Learning
prediction model
Tuberculosis Treatments

ABSTRACT

Both Data Mining techniques and Machine Learning algorithms are tools that can be used to provide beneficial support in constructing models that could effectively assist medical practitioners in making comprehensive decisions regarding potential problematic cases in Tuberculosis (TB). This study introduces two machine learning techniques which are Naïve Bayes inductive learning technique and the state of the art Bayesian Networks. These two techniques can be used towards constructing a model that can be used for predicting potential problematic cases in Tuberculosis. To construct a model, this study made have use of data collected from an Epidemiology laboratory. The volume of data was collated and divided into two data sets which are the training dataset and the investigation dataset. The model constructed by this study has shown a high predictive capability strength compared to other models presented on similar studies.

Corresponding Author:

Awad Ali,
Departement of Software Engineering,
UniversitiTeknologi Malaysia
UniversitiTeknologi Malaysia, 81310Skudai, Johor, Malaysia.
Email: awad.uofkassala@gmail.com

1. INTRODUCTION

Data mining and machine learning algorithms are regarded as powerful tools in discovering embedded hidden patterns that are already within the dataset. More often than not, algorithms could not be employed as data classifier especially when the data set is not suitable. For algorithms to work well data set should be large enough to include all possible patterns and equally important that the data set must be suitable with the techniques that will be used. Suitability refers to the type data to be used and the numbers of data attributes which most of the time the data needs some statistical techniques to prepare for its use. The machine learning and data mining techniques enable driving understandable and interpretable models; these models can be used for prediction or explanatory purposes [1].

The Naïve Bayes (NB) classifier algorithm is a machine learning algorithm that builds on the assumption of independence between events or attributes. According to Keogh [2] the independent assumption rarely happen but can positively contribute to be relaxing in order to improve the performance. Not on all cases that the relaxing of independence can lead to a favorable improvement on the performance. Once in a while, this process could possibly lead to a state of reduced accuracy. The accurateness of the naïve Bayes classifier is primarily dependent upon the features of the dataset [3].

On the one hand, a Bayesian network (BN) is a directed acrylic graph that is commonly used in many domains which includes software reliability assessment and prediction [4, 5], medical diagnosis [6], Artificial Intelligence [7], and Bioinformatics [8]. Bayesian Network is considered to be a probabilistic representation model which uses a direct cycle graph to represent the random variables by nodes, and connecting these nodes by edges. The edge that connect nodes A to B, where A is a parent node, represents a conditional probability P(B, A) [7]. There are numerous algorithms that are used for medical diagnosis, which among them are logistic regression, artificial neural networks, and decision trees. However, the selection of algorithms depends on various factors; one of the most important factors is the availability of data which has a great affection on the algorithm's performance. One character of Naïve Bayes is its tolerance for noisy and incomplete data [9], and it also provides to achieve good performance when the attributes number is large.

In Sudan, like any developing country, a number of its population suffers from dangerous, infectious diseases such as Tuberculosis (TB) which is additionally worsened by lack or insufficient medical treatment, thereby increasing potential patients' mortality. Usually, medical treatment for patients afflicted with TB is the use of antibiotics to kill the bacteria that causes the spread of the disease in a patient's body. However, this method of treatment is more problematic than the short courses of antibiotics used to treat most bacterial infections since a longer period of treatment, which takes around six to 12 months, is needed to entirely eliminate mycobacteria afflicting the body.

Aiding medical practitioners in decision making by inferring information from a database at early stages will help them to predict potential problematic cases that could possibly match some criteria such as defaulting, while using their limited resources wisely and effectively.

2. Related Works

Several studies have been done on data mining and medical applications. In article [10] the researchers proposed for a prediction model to ensure a best estimate concerning the risk of smear negative pulmonary tuberculosis in Brazil. The researcher has engaged some 551 patients who provided their medical status to be used as dataset for study purposes. The data contains attribute related to Symptoms, X-rays, and clinical impact. The algorithms used in this study are Logistic Regression and Regression Tree which yielded and achieved accuracy range from 64 percent to 76 percent.

There were numerous articles that employed machine learning techniques to provide solutions to medical problem. In [11] the researcher introduced the Naïve Bayes classifier as an efficient technique to apply for medical datasets, and using it to analyze genes dataset. However, the researcher had focused simply on the technical aspect of the dataset and ignored the medical details since the objective of their study was only limited in describing how the algorithm can be used. In [6] three predictive models were proposed to predict breast cancer survivability. The dataset were gathered from cancer incidence database in USA containing 72 attributes. Three algorithms were applied to construct this model which includes Artificial Neural Networks, Decision Trees, and Logistic Regression. Of the three algorithms, the Decision Trees provided the best performance as it achieved 93.6% and was described as the best result reported in the published literature; the model which was constructed in our study achieved higher than this ratio. However, the medical problems showed quite differently. Another work conducted in [12] where the main objective is to exploit the effects of missing data in Bayesian Network model for predicting medical problems. The researcher used pneumonia-case dataset on their study with and without missing values. Furthermore, there were more research studies that tends to address TB problems but all these articles regard medical studies for example the articles [13] and [14].

3. Data Preparation

Machine Learning requires gathering of data as cases, examples, and instances of all possible object classes. This is due to the created model where these data could be used later to know the groups or similarities in the data under the unsupervised learning, or predict the class of new objects under the supervised learning.

In this study, we used the data collected by Mahojub [15] Epidemiology laboratory (Epi- lab) in Sudan. The dataset contains about 54 attributes and 714 instances, including personal information of the patient, symptoms the patient suffers from, HIV tests, history of the disease, diagnostic tools used, treatment that includes regimens for the type of the disease and doses given, with its drug reaction, the follow-up results for the whole treatment period, also costs and hospitalization paid. However, attributes that are likely to affect the patient behavior towards the treatment (treatment outcome is one of the following: cured, treatment completed, death, failure, defaulted, and transferred out). The data were categorized as the following:

Attributes that are related to patient in particular (sex, age, income, etc).
Attributes that are related to regimen
Attributes that are related to proximity to the health center
Attributes that are related to side effects of treatment (social or clinical impact)
Attributes that are related to duration of treatment

4. Applying Naïve Bayes and Bayesian Networks algorithms:

In this section, both the Naïve Bayes and Bayesian Networks algorithms were applied to the smear-positive pulmonary TB dataset. In the first section, this study showed some interesting statistics about the attributes concerning the class distribution over the values of each attribute. In the second and third sections, the results using both the Naïve Bayes and Bayesian Networks were presented respectively. Discussion about the results is provided in the final section.

There are five steps in this work and Figure 1 has illustrated them. Gathering data is the first step. In the second step, the data were processed before the analysis process, the operations include attribute's selections, data type transformation from one type to another type such as a number to be nominal, process the missed data and so on. The third and fourth steps include the actual data mining tasks. In the third step learning of the relationship between overall attributes was done and in the fourth step targeted algorithms were implemented in order to construct the prediction model. In the last step, we must evaluate our model by domain expert, by comparing the results with other researchers, or by sensitivity analysis. If the model achieves the acceptable accuracy then the process will terminate otherwise the third step will be repeated.

4.1 Interesting Statistics

After loading the data into the software (Figure2 shows the loading process), this study has found some interesting statistics about the attributes. For example, the class targeted (treatment outcome) distribution over the cases of each attribute was shown at the bottom of the selected attribute area (see Figure 3.a and 3.b data visualization) that could view the class distribution in all attributes. The class distribution in Figure 3.a indicates the general trends. However, the final result must be concluded from the training algorithms which specify the class probability distribution over all attributes.

In Figures 3.a and 3.b, the diagrams showed an inverse affection of each attribute over the classes. For example the (piw) attribute has a clear effect to the cured class (there are two color blue and red which refer to cured and complete-treatment classes). The statistics shown here was regarded as general trends in the data; the significant results are shown in Section 4.2 and Section 4.3.

4.2 Results

By applying the Naïve Bayes techniques and Bayesian Networks algorithm for TB treatment, outcomes classification and further looking into the underlying mechanisms at the first step, this study discusses the classification model built by Naïve Bayes and consequently investigates how the Bayesian network extends it and how it can improve the classification accuracy.

4.2.1 Naïve Bayes Results

The application of Naïve Bayes technique with 10-fold cross validation as[6] in evaluating the prediction model based on the correctly classified instances, the model has produced 90.7563 percent accuracy rate (see Table 2) showing the confusion matrix in the left side there are five classes (a, b, c, d, e, f) representing treatment outcome groups. A confusion matrix is a matrix which represents per class classification (true and false classification). In Table 2, it is shown that the cured class of 616 cases were classified as true, but 7 were actually a (from a - e) classified as false (3 cases from a to b and 4 cases from a to e). Class a and b has no significant difference between them, hence the error in the classification do not have significant effect. However, the general percentage obtained from the software (correctly classified instances in Naïve Bayes is 90.7563 percent and 93.2773 percent in Bayesian Network, respectively). Table 1 below shows the number of each case (instances) in the data set which were obtained from Figure 2. Section 4.2.2 discusses how the percentage showed in Table 2 can be enhanced.

Table 1. Number of Each case (instances) in the Dataset

Target variable classifications	count
cured	625
complete_treatment	12
died	13
failure	6
defaulter	48
transferred-out	10

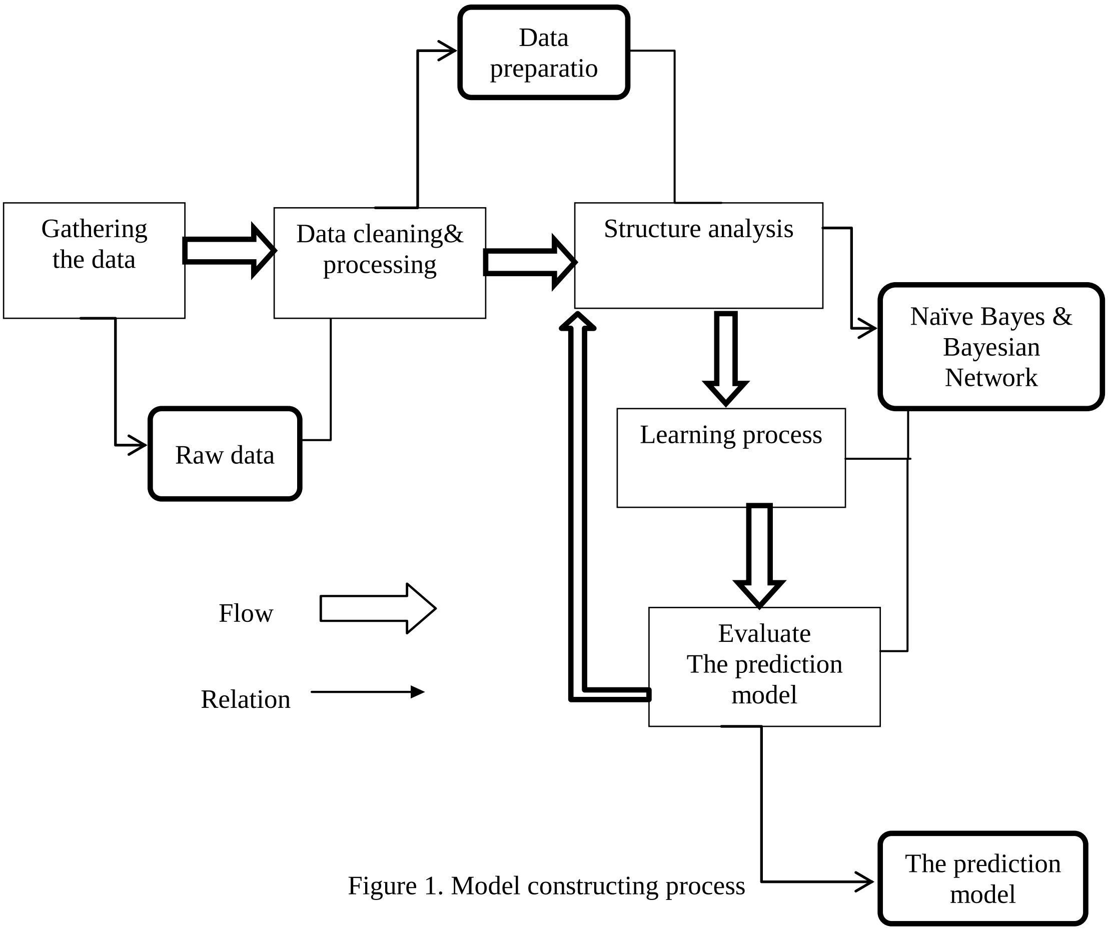
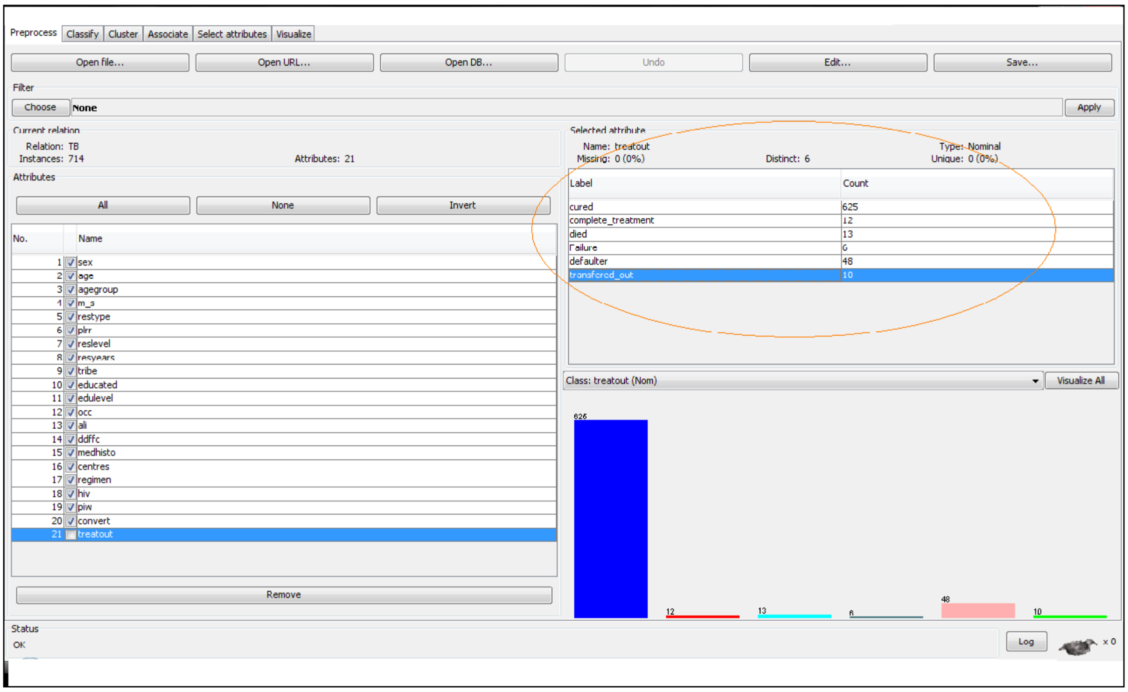


Figure 1. Model constructing process



Filter

Choose **None**

Current relation

Relation: TB
Instances: 714 Attributes: 21

Attributes

No.		Name
1	☑	sex
2	☑	age
3	☑	agegroup
4	☑	m_s
5	☑	restype
6	☑	plrr
7	☑	reslevel
8	☑	resyears
9	☑	tribe
10	☑	educated
11	☑	edulevel
12	☑	occ
13	☑	ali
14	☑	ddffc
15	☑	medhisto
16	☑	centres
17	☑	regimen
18	☑	hiv
19	☑	piw
20	☑	convert
21	☐	treatout

Selected attribute

Name: treatout Type: Nominal
Missing: 0 (0%) Distinct: 6 Unique: 0 (0%)

Label	Count
cured	625
complete_treatment	12
died	13
Failure	6
defaulter	48
transfered_out	10

Class: treatout (Nom)

Status
OK

Figure 2. Data loading

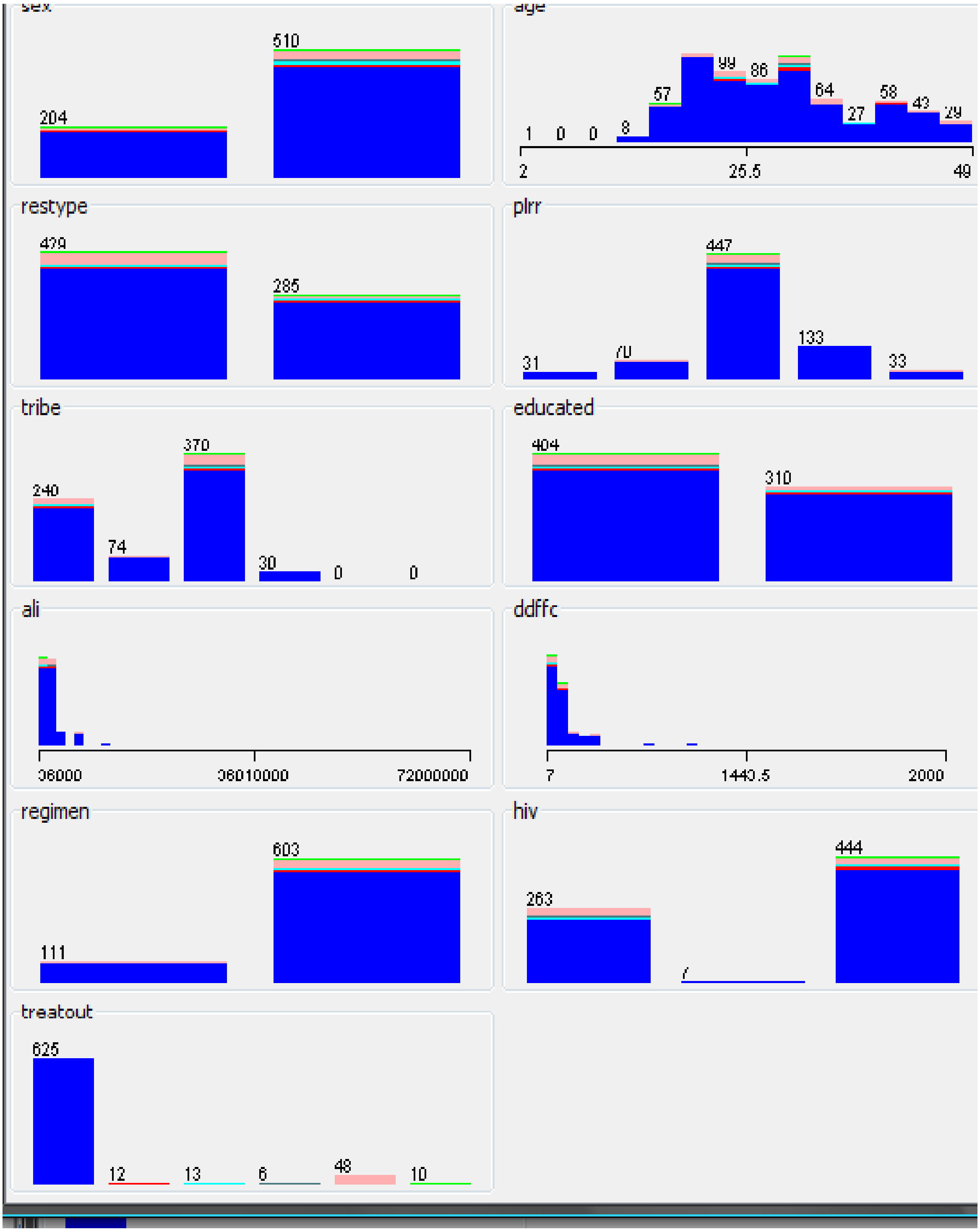


Figure 3a. Data visualization Part of the distribution of treatment outcome classes over all attributes.

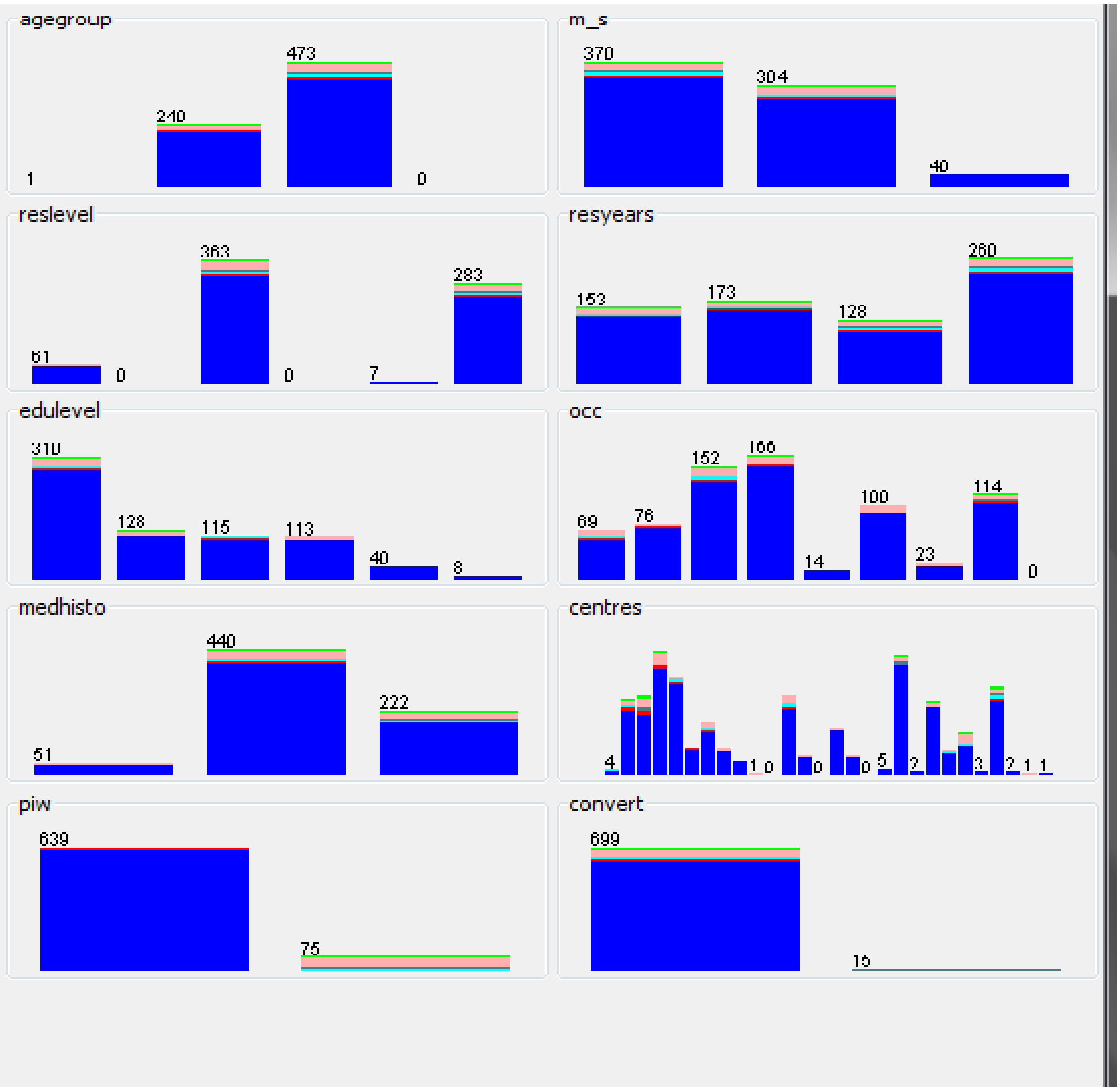


Figure 3b. Data visualization Another part of the distribution of treatment outcome classes over allattributes. By examining the diagrams (the colors indicate the treatment outcome groups) in Figures 3.a and 3.b, one can find the most specific terms for each group.

Table 2 Confusion Matrix and the Percentage Achieved by Each Algorithm

	Bayes Naïve 90.7563%						Bayesian Network 93.2773%					
class	a	b	c	d	e	f	a	b	c	d	e	f
a=cured	616	3	0	0	4	2	624	1	0	0	0	0
b=complete_treatment	12	0	0	0	0	0	12	0	0	0	0	0
c=died	1	0	3	0	6	3	1	0	3	0	8	1
d=Failure	0	0	0	1	2	3	0	0	0	0	4	2
e=defaulter	1	3	6	0	28	10	0	0	3	2	39	4
f=transferred_out	1	0	4	1	4	0	1	0	1	1	7	0

4.2.2 Bayesian Network Result

In this stage, this study applied the Bayesian Network classifier to the dataset by using the default parameter setting and 10-fold cross validation. As a result, the model produced a 93.2773 percent accuracy rate. See Figure 5. which shows the graphical model.

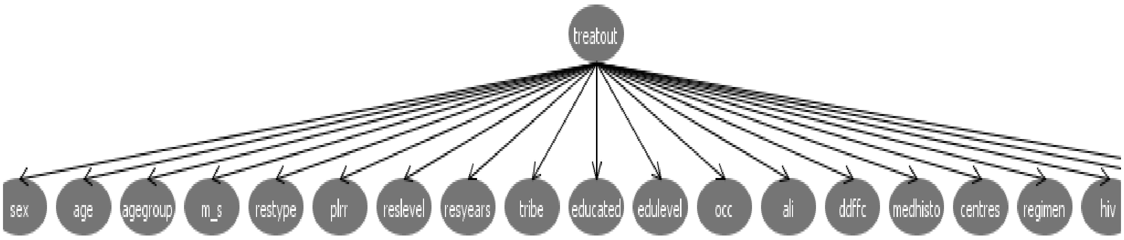
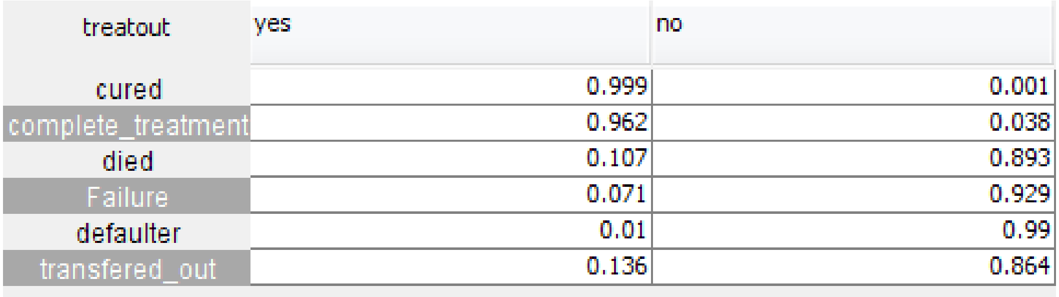


Figure 4. The graphical model for Bayesian Network

In Figure 3, there are root nodes representing the treatment outcome attribute and independent leaves representing the other attributes on the data set. From Figure 4, it should be noted that this structure of the network makes this model equivalent to the Naïve Bayes classifier. One of the Bayesian Network algorithm assumptions, states that the variables are statistically independent, and this assumption meets Naïve Bayes algorithm assumption (Naïve Bayes has one assumption, the independence assumption). On the other hand, the reason that makes the model achieve little higher accuracy (93.2773%) is due to the set of the probabilities estimation [16]. In the weka software, the naïve Bayes technique employed Laplace estimator which was set to the value 1 for the initial value for the frequency count, where Bayesian Network used 0.5 [17]. This estimator was also applied to the prior probability calculation of the target variable.

As mentioned earlier, the Bayesian Network algorithm has more than one assumption which affects the accuracy and the network structure. It is assumed that not all the variables are completely independentaccording to [2] and [17] where the independent assumption could rarely arise. Furthermore, there is a relation between variables shown in Table 3, which represent the probabilities of the variable. It should benoticed that the cured class is dependent upon the variable (piw) by 0.999 probabilities. Thus, we could violate the independent assumption “no more than one parent” which is expressed by set (maxNrOfParents=1), while in weka software it is set as (maxNrOfParents=2), with which this process leads to achieve 94.2577 percent accuracy. And the graphical model change as shown in Figure 5. In Figure 5 the graphical models describe how the probability distribution effect on the model representation. Clearly we notice the relation between attributes and their affection to the prediction result.

Table 3Probability Distribution of (piw) Variable.



treatout	yes	no
cured	0.999	0.001
complete_treatment	0.962	0.038
died	0.107	0.893
Failure	0.071	0.929
defaulter	0.01	0.99
transfered_out	0.136	0.864

Table 3 shows that the cured class depend on the variable (piw) by probability 0.999. This result allows violating the independent assumption.

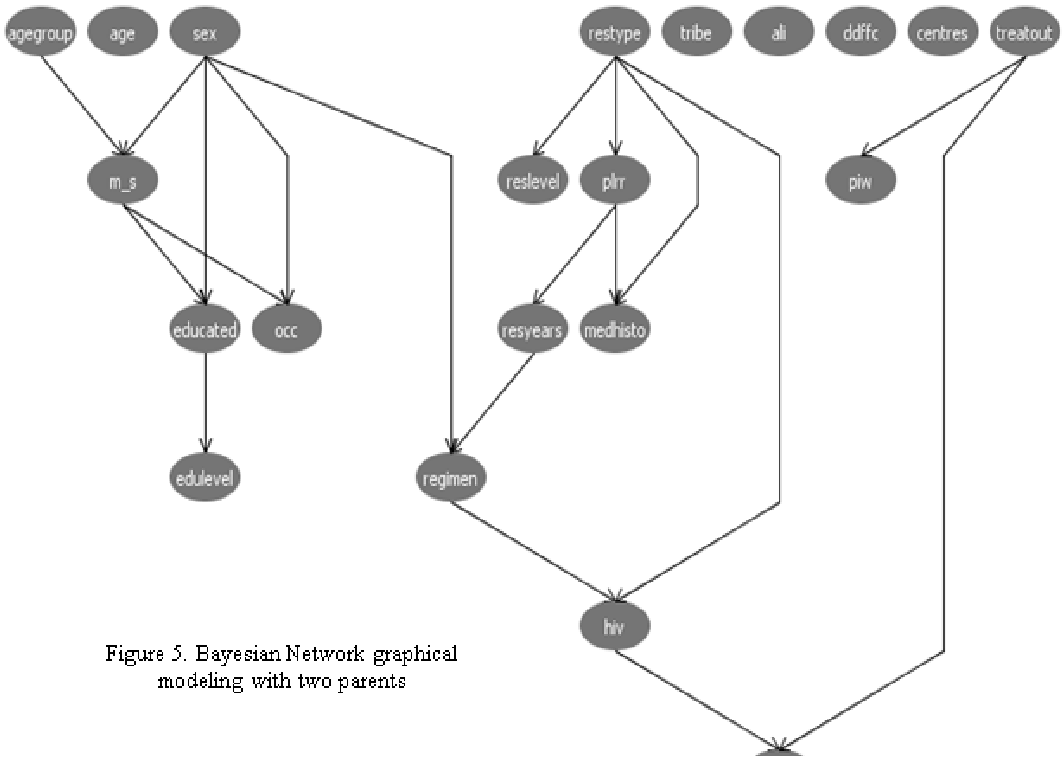


Figure 5. Bayesian Network graphical modeling with two parents

5. Conclusion and future work

In this study, two machine learning algorithms, the Naïve Bayes and Bayesian Network, were proposed as predictive techniques that can be adopted towards constructing a model for predicting potential problematic cases in Tuberculosis disease. Moreover, this study used medical dataset gathered from Epidemiology laboratory. Naïve Bayes produced acceptable accuracy while Bayesian Network has led to enhance its accuracy and showed better performance through the violation of Naïve Bayes independent assumptions. The violation was based on the probability distribution of attributes. The current proposed models, if prospectively confirmed, would be useful in guiding medical and health practitioners in estimating the risk of smear-positive pulmonary TB with or without HIV test, optimizing the use of more expensive tests, and the unnecessary anti-PT treatment. The models could be useful as a cost-effective tool in a health care system with limited resources. Furthermore, the Naïve Bayes and Bayesian Network can be used as best predictive tools for constructing prediction models to solve problems on others domains such as reliability prediction of component based systems. There are various machine learning techniques that were not examined here (due to time constraints) as they may be also effective in dealing with uncertainty. These other machine learning techniques include, but not limited to Rule Induction, and Support Vector Machines, among others. Finally, the authors of this study believed that the machine learning techniques used in this paper warrant further investigation, particularly to explore under which conditions and attributes, in this underlined problem where they are most likely to be effective.

ACKNOWLEDGEMENTS

The author would like to thank UniversitiTeknologi Malaysia and software engineering research group for their support.

REFERENCES

[1] C. Helma, et al., "Data mining and machine learning techniques for the identification of mutagenicity inducing substructures and structure activity relationships of noncongeneric compounds," Journal of chemical information and computer sciences, vol. 44, pp. 1402-1411, 2004.

[2] E. Keogh and M. Pazzani, "Learning augmented Bayesian classifiers: A comparison of distribution-based and classification-based approaches," 1999, pp. 225-230.

[3] J. Hucaljuk and A. Rakipovic, "Predicting football scores using machine learning techniques," in MIPRO, 2011 Proceedings of the 34th International Convention, 2011, pp. 1623-1627.

[4] H. Singh, et al., "A bayesian approach to reliability prediction and assessment of component based systems," 2001, pp. 12-21.

[5] R. Roshandel, et al., "A Bayesian model for predicting reliability of software systems at the architectural level," Software Architectures, Components, and Applications, pp. 108-126, 2007.

[6] D. Delen, et al., "Predicting breast cancer survivability: a comparison of three data mining methods," Artificial intelligence in medicine, vol. 34, pp. 113-127, 2005.

[7] S. Russell, et al., "A modern approach," Artificial Intelligence. Prentice-Hall, Egnlewood Cliffs, 1995.

[8] N. Friedman, et al., "Using Bayesian networks to analyze expression data," Journal of computational biology, vol. 7, pp. 601-620, 2000.

[9] J. Kazmierska and J. Malicki, "Application of the Naive Bayesian Classifier to optimize treatment decisions," Radiother Oncol, vol. 86, pp. 211-6, 2008.

[10] F. C. Q. Mello, et al., "Predicting smear negative pulmonary tuberculosis with classification trees and logistic regression: a cross-sectional study," BMC Public Health, vol. 6, p. 43, 2006.

[11] R. Bhuvaneswari and K. Kalaiselvi, "Naive Bayesian Classification Approach in Healthcare Applications."

[12] J. H. Lin and P. J. Haug, "Exploiting missing clinical data in Bayesian network modeling for predicting medical problems," Journal of biomedical informatics, vol. 41, pp. 1-14, 2008.

[13] A. Pablos-Méndez, et al., "Nonadherence in tuberculosis treatment: predictors and consequences in New York City," The American journal of medicine, vol. 102, pp. 164-170, 1997.

[14] A. Van Rie, et al., "Analysis for a limited number of gene codons can predict drug resistance of Mycobacterium tuberculosis in a high-incidence community," Journal of clinical microbiology, vol. 39, pp. 636-641, 2001.

[15] L. Mahgoub, "On Building predication system for public Health-A Comparative Study Using SNTB programme " master, computer science, khartoum, khartoum, 2007.

[16] R. R. Bouckaert. (2008). Bayesian Network Classifiers in Wekafor Version 3-5-7. Available: http://www.cs.waikato.ac.nz/~remco/weka_bn/index.html

[17] Z. Markov and I. Russell, "Probabilistic Reasoning with Naïve Bayes and Bayesian Networks," 2007.

BIOGRAPHY OF AUTHORS

Mr. Awad Ali Recived the B.Sc. from University of Alneelean and M.Sc. from University of Khartoum, Sudan.He worked as a lecturer at University of Kassala. Currently he is a Ph.D candideat at Facltyof Computer Science & information system, Universiti Teknologi Malaysia UTM. He is a member of SERG research group. His research interests include data mining, software engineering, software reliability, and component based software engineering.

Assoc. Prof. Moawia Elfaki Recived the B.Sc. and M.Sc. from University of Khartoum, Sudan, and his Ph.D from Putra University, Malaysia in 1999. He is now an Associate Profesor at the College of Computer Science and IT, King Faisal University, Sudi Arabia. He worked as head of computer science department at University of Khartoum. His research interests are intelligent hybrid systems, expert systems, data mining, text mining, neural networks, rough set theory, and intelligent information retrieval.

Dr. Dayang Norhayati Abang Jawawi is a lecturer at the Department of Software Engineering, Faculty of Computer Science and Information Systems, UniversitiTeknologi Malaysia (UTM). She received her B.Sc. Degree in Software Engineering from Sheffield Hallam University, UK and M.Sc. and Ph.D research in software engineering from Universiti Teknologi Malaysia. Currently she is Head of Software Engineering Department, Faculty of Computer Science and Information System, UTM, and a member of Software Engineering Research Group (SERG), K-Economy, UTM. Her research areas are software reuse, component-based software engineering and embedded real-time software.