Scientific Research Publishing

Journal of Computer and Communications, 2019, 7, 243-251
http://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

Predicting Tuberculosis Treatment Relapse: A Decision Tree Analysis of J48 for Data Mining

Arnold P. Dela Cruz[1,2], Gilbert M. Tumibay[3]

[1]Nueva Ecija University of Science and Technology, Cabanatuan City, Philippines
[2]Graduate School, Angeles University Foundation, Angeles City, Philippines
[3]Angeles University Foundation, Angeles City, Philippines
Email: arnold@neust.edu.ph, tumibay.gibo@auf.edu.ph

How to cite this paper: Cruz, A.P.D. and Tumibay, G.M. (2019) Predicting Tuberculosis Treatment Relapse: A Decision Tree Analysis of J48 for Data Mining. *Journal of Computer and Communications*, **7**, 243-251.
https://doi.org/10.4236/jcc.2019.77020

Received: June 6, 2019
Accepted: July 26, 2019
Published: July 29, 2019

Abstract

Tuberculosis remains an important problem in public health that threatens the world, including the Philippines. Treatment relapse continues to place a severe problem on patients and TB programs worldwide. A significant reason for the development of decline is poor compliance with medical treatments. The objectives of this research are to generate a predictive data mining model to classify the treatment relapse of TB patients and to identify the features influencing the category of treatment relapse. The TB patient dataset is applied and tested in decision tree J48 algorithm using WEKA. The J48 model identified the three (3) significant independent variables (DSSM Result, Age, and Sex) as predictors of category treatment relapse.

Keywords

Data Mining, Decision Tree, J48, Tuberculosis, WEKA

1. Introduction

Tuberculosis (TB) remains the deadliest infectious disease worldwide, with 10.4 million infections and a death toll of 1.7 million people in 2016, according to the World Health Organization (WHO) statistics [1]. Tuberculosis is an infectious disease caused by a bacterium called Mycobacterium Tuberculosis that remains a global health problem. It primarily infects the lungs, bones, lymph, and digestive organs. People with diabetes, lower immunity, who use immunosuppressive drugs, old age, and people with HIV/AIDS infection, are more likely to develop TB [2]. In the Philippines, TB is still a major public health concern. Globally, the Philippines ranked ninth (9^{th}) among 22 TB high burdened countries (HBCs). One major problem in TB treatment is guaranteeing the patients to follow their

treatment, together with medical checkups and medication till completion. TB patients need retreatment if they relapse or fail in the initial treatment. "Outcomes among patients getting a standard World Health Organization Category II retreatment routine are deficient, resulting in an increased risk of disease, transmission, and drug resistance" [3].

Infection with tuberculosis can lead to life-threatening complications. Unfortunately, there is little information about how patients are considered relapse in treatment. Data mining is the exploration of large datasets to extract hidden and previously unknown patterns and relationships [4]. Data mining techniques in healthcare have been widely applied in different applications, including predicting patient outcomes and modeling health outcomes, evaluation of treatment effectiveness, infection control, and hospital ranking [5]. The term "data mining" (often called as knowledge discovery) "refers to the process of analyzing data from different perspectives and summarizing it into useful information utilizing several analytical tools and techniques, which in turn may be valuable to increase the performance of a system" [6].

Technically, the term "data mining" is the process of finding patterns or relationships among lots of attributes in large relational databases.

Several studies in data mining have been widely used for the prognosis and diagnoses of many diseases. Ferreira *et al.* [7] used data mining to improve the diagnosis of neonatal jaundice in newborns. In their research, the dataset contained 70 attributes collected for 227 healthy newborns. Sanchez *et al.* [8] applied data mining technique to categorize TB related to identifying patients' sickness. The study was analyzing the TB diagnostic categories based on given variables. The dataset containing 1655 instances with 56 attributes was used as a raw dataset. The 56 qualities are trimmed into five attributes which are backgrounds, age category, bacteriology results, extrapulmonary, and pulmonary tuberculosis. Several techniques in data mining were useful, including J48, Naive Bayes classifier, CART, multilayer perceptron, simple logistic, and SMO. Vikas and Pal [9] conducted research to recognize the most common data mining algorithms applied in modern Medical Diagnosis and assessed their performance on several medical datasets. Five (5) algorithms were selected: Simple Logistic, RBF Network, Naıve Bayes, J48, and Decision Tree. Some studies have discovered the decision tree method to analyze medical data. Sharma *et al.* [10], and Zolbanin *et al.* [11] have used the decision tree algorithm in their research work, to examine the data, and make the tree and its rules to make a prediction. The two authors have used the decision tree to the dataset to increase the correctness of predictive performance.

However, none of the authors mentioned above ventured on the utilization of J48 decision tree algorithm which engaged to predict the TB patient category treatment relapse which this paper addressed. J48 is very instrumental in this kind of study since it handles both categorical and continuous attributes to build a decision tree. To handle continuous attributes, J48 splits the attribute values

into two partitions based on the selected threshold such that all the values above the threshold as one child and the remaining as another child. It also handles missing attribute values. J48 uses the gain ratio as an attribute selection measure to build a decision tree. It removes the biases of information gain when there are many outcome values of an attribute. J48 uses pessimistic pruning to remove unnecessary branches in the decision tree to improve the accuracy of classification [12].

This article aims to make a predictive data mining model for treatment relapse to classify which factors influence the category treatment relapse of TB. The model was built by applying data mining technique to the data provided by the Integrated Tuberculosis Information System (ITIS) website of the Department of Health (DOH).

2. Methodology

This research is a quantitative research design that uses the computational, mathematical, and statistical tools to analyze and examine the data to simplify outcomes from the datasets. The dataset of TB patients came from Cabanatuan City, which was mined last 2017 from the Integrated Tuberculosis Information System database of City Health Office. This approves the comparability and accuracy of data, which are essential features in the J48 model. The collected data were scrutinized and coded using the WEKA (Waikato Environment for Knowledge Analysis). WEKA is a collection of machine learning algorithms for data mining tasks. The study procedure of interview and data collection was permitted by the Office of the City Mayor and the City Health Director of Cabanatuan City.

The researchers used data mining as a tool with the J48 decision tree as a method to design the prediction model treatment relapse of TB patients. Data mining sort enormous datasets to classify patterns and established relationship to unravel complications with the use of data analysis. Data mining tools were utilized by enterprises to forecast coming trends [13]. Classification trees are mostly used in several fields, such as medicine, botany, psychology, and computer science [14]. "These classification trees promptly give themselves to being presented graphically, assisting in making them easier to analyze than they would be if only a strict numerical interpretation were possible". J48 Decision Tree Algorithm is an implementation by the WEKA project team of the famous tree training algorithm C4.5 [15]. The decision tree classification model advantages are easy to understand and identified to have comparable accuracy to other classification models.

3. Results and Discussions

The WEKA platform [16] was used in this research. WEKA is a state-of-the-art tool for developing Machine Learning (ML) techniques and their application to real-world data mining problems. It is a popular data mining software with a

user-friendly graphical user interface that supports a wide range of data mining algorithms. The J48 model was built using 10-fold cross-validation. Decision Tree J48 is the implementation of algorithm ID3 (Iterative Dichotomiser 3) developed by the WEKA project team.

3.1. J48 Analysis Application

Table 1 shows the description of the user attributes to form a J48 decision tree algorithm was completed as follows: The Patient Category attribute is distinct as a dependent variable. The attribute type of Patient Category is nominal with two values (relapse, non-relapse). As identified in Table 1, as to age, 166 patients were less than 10 years old; 32 patients were in the age group of 10 to 19 years; 56 patients were age in the age group of 20 to 28 years; 56 patients were age in the age group of 29 to 38 years; 59 patients were age in the age group of 39 to 47 years; 92 patients were age in the group of 48 to 57 years; 37 patients were in the

Table 1. Structure of Attributes used in J48 Analysis.

Attribute	Value (modalities)	Structure		Type of Attribute	
		fi	%		
Age	<10	166	33	Numeric	Independent
	10 - 19	32	6		
	20 -28	56	11		
	29 - 38	56	11		
	39 - 47	59	12		
	48 - 57	92	18		
	58 - 66	37	7		
	67 - 76	8	2		
	>76	4	1		
Sex	Male	343	67	Nominal	Independent
	Female	167	33		
BacStatus (Bacteriologically Status)	Bacteriologically-Confirmed TB	258	51	Nominal	Independent
	Clinically-Diagnosed TB	252	49		
DSSMResult (Direct Sputum Smear Microscopy)	ODT (Observed Direct Treatment)	174	34	Nominal	Independent
	0	130	26		
	1+	87	17		
	2+	57	11		
	3+	62	12		
Classification	Pulmonary	507	99	Nominal	Independent
	Extra-Pulmonary	3	1		
Patient Category	Non-Relapse	455	89	Nominal	Dependent
	Relapse	55	11		

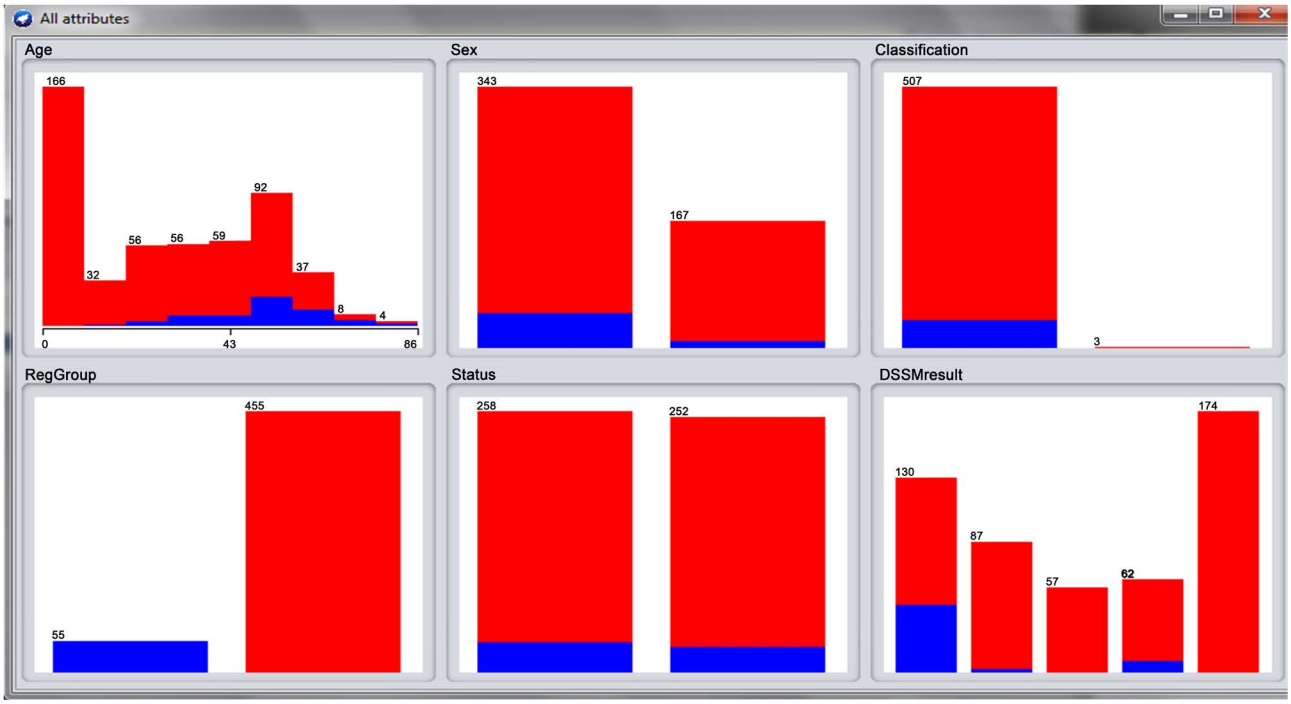

Figure 1. Attributes of TB patient's dataset using WEKA.

age group of 58 to 66 group; 8 patients were in the age group of 67 to 76 years; and 4 patients were more than 76 years old. In attribute sex, the data showed that there were more males than females. In their bacteriologically status, 51% of the TB patients were under bacteriologically-confirmed TB, while 49% belonged to clinically-diagnosed TB. In DSSM result, "ODT" and "O" had the highest percentage (60%) of TB patients. Almost 100% of the tuberculosis cases on the classification were in pulmonary value; while in the patient category, the majority was under non-relapse.

In Figure 1, the attribute simulation describes different factors for tuberculosis patient category using WEKA. The dataset contains 510 instances and 6 attributes.

3.2. Modeling Results Analysis

The graphical representation in Figure 2 is the J48 Decision Tree Relapse Model.

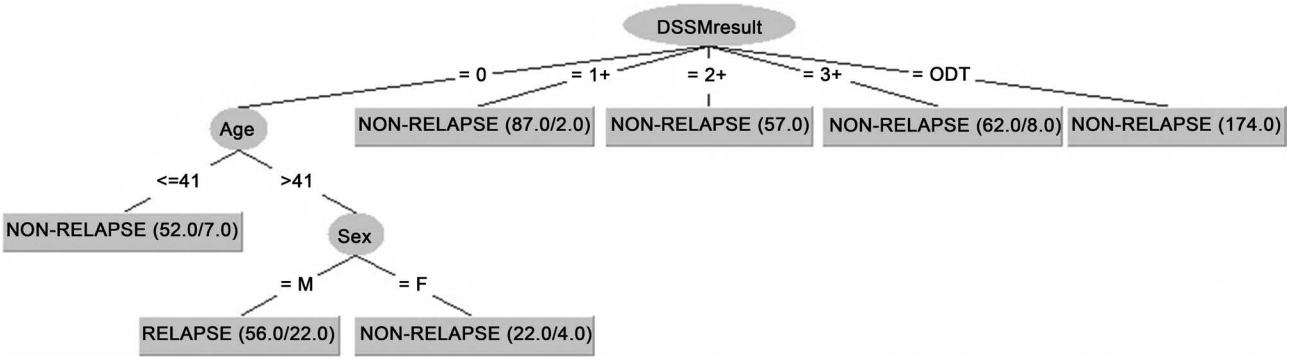

Figure 2. J48 decision tree relapse model.

The main objective of data visualization is to link information efficiently and clearly. It makes composite data more usable, accessible, and understandable.

J48 Model Rule Sets

Rule 1:
If (DSSMResult = "1+") or (DSSMResult = "2+") or (DSSMResult = "3+") or (DSSMResult = "ODT") then Prediction = "Non-Relapse"
Rule 2:
If (DSSMResult = "0") and (Age <= 41) then
Prediction = "Non-Relapse"
Rule 3:
If (DSSMResult = "0") and (Age > 41) and (Sex = "F") then
Prediction = "Non-Relapse"
Rule 4:
If (DSSMResult = "0") and (Age > 41) and (Sex = "M") then
Prediction = "Relapse"

This J48 model shows that the three important factors for predicting relapse are DSSMresult, age, and sex. The attribute DSSMResult appears as the first splitting attribute. This specifies the significance of this information. The model can be interpreted as follows: if the TB patient's DSSMResult is equal to "1+", or "2+", or "3+", or "ODT", the model predicts non-relapse. However, if the patient's DSSMResult is equal to "O", then the model examines the age of the patient. If the age of the patient is less than or equal to 41, the model predicts non-relapse.

On the other hand, if the age is higher than 41 years, the model examines the sex of the patient. If the sex of the patient is female (F), the model predicts non-relapse. Otherwise, relapse is predicted. According to this model, when the DSSMResult is "O" and age is greater than 41 years, and sex is equal to "M", the patient is at high risk to repeat the TB treatment, or the TB patient becomes a relapse.

3.3. Classification Model Accuracy Assessment

In this research, the researchers used the Waikato Environment for Knowledge Analysis tool in **Figure 3** for calculating accuracy based on incorrect and correct classes produced by the confusion matrix.

Table 2 shows the tabular display using WEKA that assesses the forecasting accuracy of a predictive model from the dataset of TB patient in the year 2017 with 510 instances and 6 attributes.

$$\text{Accuracy}(\text{ACC}) = \frac{\text{TN} + \text{TP}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}$$

$$\text{Accuracy}(\text{ACC}) = \frac{434 + 27}{27 + 28 + 434 + 21}$$

$$\text{Accuracy}(\text{ACC}) = 90.39\%$$

```
=== Summary ===

Correctly Classified Instances         461               90.3922 %
Incorrectly Classified Instances        49                9.6078 %
Kappa statistic                          0.4711
Mean absolute error                      0.1314
Root mean squared error                  0.268
Relative absolute error                 67.8223 %
Root relative squared error             86.404  %
Total Number of Instances              510

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.491    0.046    0.563      0.491    0.524      0.472    0.889     0.426     RELAPSE
               0.954    0.509    0.939      0.954    0.947      0.472    0.889     0.985     NON-RELAPSE
Weighted Avg.  0.904    0.459    0.899      0.904    0.901      0.472    0.889     0.925

=== Confusion Matrix ===

   a   b   <-- classified as
  27  28 |   a = RELAPSE
  21 434 |   b = NON-RELAPSE
```

Figure 3. Accuracy rate of J48 decision tree.

Table 2. Confusion matrix.

	TB Treatment Category		
Predicted	**n = 510**	**Relapse**	**Non-Relapse**
	Relapse	True Positive TP = 27	False Positive FP = 28
	Non-Relapse	False Negative FN = 21	True Negative TN = 434

Accuracy (ACC) signifies the amount of the total number of TB patient predictions that are correct. True Positive (TP) means the amount of actual outcome of TB patient relapse that is accurately classified as predicted relapse category, and True Negative (TN) refers to the number of TB patient non-relapse that are rightly classified as predicted TB patient non-relapse category. The accurately classified instance is equal to 90.39%.

4. Conclusions and Recommendations

This paper deals with efficient data mining procedure for predicting the TB relapse from medical records of patients. The J48 classifier was developed by the researchers using WEKA and trained it on a preprocessed TB dataset. The J48 classifier is used to increase the accuracy rate of the data mining procedure. From the results, algorithm J48 predicted the patient category of TB data with the accuracy of 90.39%, which is reasonable enough for the system to be depended on for prediction of category relapse. In order to measure the unbiased prediction accuracy of the method, the 10-fold cross-validation procedure was used. The J48 prediction model was handy and advantageous to assess the consistency among attributes that are used to predict the TB treatment category re-

lapse. A J48 model was established based on the input variables gathered from the ITIS database of the City Health Office of Cabanatuan. The attributes DSSM Result, Age, and Sex are the most critical factors for the prediction of patient category relapse.

It is recommended other IT and computer science experts to venture on studies that are medically related using J48 predication model and continue to investigate and evaluate the technique [17] they are using so as to increase the overall performance of their healthcare delivery system [18].

Acknowledgements

The researchers are grateful for the support offered to them by the City Health Office of Cabanatuan and the Nueva Ecija Provincial Health Office.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] World Health Organization (2017) Global Tuberculosis Report 2017.

[2] World Health Organization (2014) Global Tuberculosis Control: WHO Report 2014. World Health Organization, Geneva.

[3] Dooley, K.E., *et al.* (2011) Risk Factors for Tuberculosis Treatment Failure, Default, or Relapse and Outcomes of Retreatment in Morocco. *BMC Public Health*, **11**, 140. https://doi.org/10.1186/1471-2458-11-140

[4] Han, J. and Kamber, M. (2011) Data Mining: Concepts and Techniques. 3rd Edition, The Morgan Kaufmann Series in Data Management Systems, Morgan Kaufmann, Burlington.

[5] Suh, S.C. (2012) Practical Applications of Data Mining. Jones & Bartlett Publishers, Burlington.

[6] Larose, D.T. (2005) Discovering Knowledge in Data: An Introduction to Data Mining. John Wiley, New York, 203-231.

[7] Ferreira, D., Oliveira, A. and Freitas, A. (2012) Applying Data Mining Techniques to Improve Diagnosis in Neonatal Jaundice. *BMC Medical Informatics and Decision Making*, **12**, 143. https://doi.org/10.1186/1472-6947-12-143

[8] Sánchez, M.A., Uremovich, S. and Acrogliano, P. (2009) Mining Tuberculosis Data. In: Berka, P., Rauch, J. and Zighed, D.A., Eds., *Data Mining and Medical Knowledge Management*: *Cases and Applications*, Medical Information Science Reference, New York.

[9] Vikas, C. and Pal, S. (2014) Performance Analysis of Data Mining Algorithms for Diagnosis and Prediction of Heart and Breast Cancer Disease. *International Journal of Innovative Computing, Information & Control*: *IJICIC*, **3**, 1-13.

[10] Sharma, N. and Om, H. (2013) Data Mining Models for Predicting Oral Cancer Survivability. *Network Modeling Analysis in Health Informatics and Bioinformatics*, **2**, 285-295. https://doi.org/10.1007/s13721-013-0045-7

[11] Zolbanin, H.M., Delen, D. and Hassan Zadeh, A. (2015) Predicting Overall Surviva-

bility in Comorbidity of Cancers: A Data Mining Approach. *Decision Support Systems*, **74**, 150-161. https://doi.org/10.1016/j.dss.2015.04.003

[12] Quinlan, J.R. (1986) Induction of Decision Trees. *Machine Learning*, **1**, 81-106. https://doi.org/10.1007/BF00116251

[13] Rouse, M. (2019) AWS Analytics Tools Help Make Sense of Big Data. https://searchsqlserver.techtarget.com/definition/data-mining

[14] Camdeviren, H.A., Yazici, A.C., Akkus, Z., Bugday, R. and Sungur, M.A. (2007) Comparison of Logistic Regression Model and Classification Tree: An Application to Postpartum Depression Data. *Expert Systems with Applications*, **32**, 987-994. https://doi.org/10.1016/j.eswa.2006.02.022

[15] Quinlan, R. (1993) C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, San Mateo.

[16] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. and Witten, I.H. (2009) The WEKA Data Mining Software: An Update. *ACM SIGKDD Explorations Newsletter*, **11**, 10-18. https://doi.org/10.1145/1656274.1656278

[17] Subia, G.S. (2018) Comprehensible Technique in Solving Consecutive Number Problems in Algebra. *Journal of Applied Mathematics and Physics*, **6**, 447-457. https://doi.org/10.4236/jamp.2018.63041

[18] Cruz, A. and Tumibay, G. (2019) An Efficiency Assessment of Tuberculosis Treatment on Health Centers: A Data Envelopment Analysis Approach. *Journal of Computer and Communications*, **7**, 11-20. https://doi.org/10.4236/jcc.2019.74002