Kesmas

Volume 19
Issue 5 *Special Issue 1, July*

Article 12

7-31-2024

Data Mining Analysis with Orange in the Development of Tuberculosis Among Diabetes Mellitus Patients

Malahayati Rusli Bintang
Universitas Batam, Batam, mbintang2708@gmail.com

Adang Bachtiar
Universitas Indonesia, Depok, adang@post.harvard.edu

Cicilya Candi
Universitas Indonesia, Depok, cicilyacandi@ui.ac.id

Recommended Citation

Bintang MR , Bachtiar A , Candi C , et al. Data Mining Analysis with Orange in the Development of Tuberculosis Among Diabetes Mellitus Patients. *Kesmas*. 2024; 19(5): 84-92
DOI: 10.21109/kesmas.v19isp1.1097
Available at: https://scholarhub.ui.ac.id/kesmas/vol19/iss5/12

Bintang, *et al. Kesmas: Jurnal Kesehatan Masyarakat Nasional (National Public Health Journal)*. 2024; 19 (Special Issue 1): 84-92
DOI: 10.21109/kesmas.v19isp1.1097

Data Mining Analysis with Orange in the Development of Tuberculosis Among Diabetes Mellitus Patients

Malahayati Rusli Bintang[1], Adang Bachtiar[2], Cicilya Candi[2*]

[1]Department of Medicine, Faculty of Medicine, Universitas Batam, Batam, Indonesia
[2]Department of Health Policy and Administration, Faculty of Public Health, Universitas Indonesia, Depok, Indonesia

Abstract
Prevention and treatment of diabetes will have a positive influence on tuberculosis (TB) since people may get TB because they have diabetes mellitus (DM). Recording and reporting through the TB Information System are not run optimally because of many factors. The information system must be strengthened to be used by private health facilities. This study used secondary data from the 2013 and 2018 Indonesian Basic Health Research (IBHR). The data was analyzed univariately and analyzed further using Orange Data Mining Tools to test the screening tool model used to predict TB in diabetic individuals. The total sample in this study from each data was 38,136 people. The 2013 IBHR stated that 749 people (2%) were diagnosed with pulmonary TB, while the 2018 IBHR stated that 97 people (0.3%) were diagnosed in the previous six months. The results of the Orange analysis showed that precision and recall calculations in this study were quite good, at 0.9. Therefore, the model would likely predict the occurrence of TB in diabetic individuals. According to Orange, the TB-DM electronic screening tool model tends to estimate the incidence of TB in diabetic individuals.

Keywords: diabetes mellitus, Orange analysis, screening, screening tool, tuberculosis

Introduction

Tuberculosis (TB) is the leading cause of death worldwide. In 2022, the largest number of new TB cases was in Southeast Asia (46%), followed by Africa (23%) and the Western Pacific Region (18%).[1] In 2021, 10 countries accounted for two-thirds of new TB cases worldwide: India, Indonesia, China, the Philippines, Pakistan, Nigeria, Bangladesh, the Democratic Republic of Congo, South Africa, and Myanmar.[1] In 2022, the TB case notification rate (CNR) in Indonesia reached 724,309 cases, indicating that 75% of TB patients have been confirmed, identified, detected, or reported. Also, the treatment coverage (TC) for TB reached 74.7%. This figure represents a notable improvement in TB treatment coverage compared to the preceding years but is still lower than a target of 90%. In 2022, Indonesia's TC for TB is projected to reach 48%, significantly below the government's target of 80%.[2]

In 2021, the total number of TB cases identified in Jakarta was 26,854, with East Jakarta registering the highest incidence of TB cases, followed by Central and West Jakarta.[3] The CNR in 2022 was 426 per 100,000 people. The TB TC was 100% in 2022, successfully exceeding the national target of 90%.[4] In contrast, the success rate in 2022 was 81%, slightly under the national target of 90%.[5] According to Indonesia's national TB control strategy for 2020-2024, the main efforts to improve TB case finding and treatment include mandatory TB case-finding reporting in all health facilities, active case finding in people with TB risk factors, improving the quality of recording and reporting (notification) in all health facilities, expanding and strengthening TB diagnosis and treatment services, treatment monitoring, and optimizing TB-related promotion and education in the community.[6]

Since 2016, the Regulation of the Indonesian Minister of Health No. 67 of 2016 Concerning TB Control has controlled the policy of reporting TB discoveries (notice). According to the policy, TB is an infectious disease that must be reported by each health establishment providing TB services.[7] The TB notification policy has been modified in Presidential Regulation No. 67 of 2021 Concerning TB Control, requiring case reporting within six months or at specific intervals if necessary.

***Correspondence*:** Cicilya Candi, Department of Health Policy and Administration, Faculty of Public Health, Universitas Indonesia, F Building 1st Floor Kampus Baru UI Depok 16424, Indonesia. Email: cicilyacandi@ui.ac.id. Phone: +62-889 06199017*

Received : February 10, 2024
Accepted : July 24, 2024
Published: July 31, 2024

Each region carries out the outcomes and reports at the district/city level to the province until they reach the applicable ministry or agency.[8] However, private health facilities continue to play a small role in TB case detection and treatment in Indonesia. The intricacy of the form for reporting TB patients is the cause of low notification among private practitioners. Another determinant is staff reporting TB cases; 50% of the personnel submitting TB notifications have not attended training or workshops for WiFi TB applications, an Android-based application to report the suspect or new TB cases.[9]

According to Ronacher *et al.,* more people across Africa suffer from TB because they have diabetes mellitus (DM) than because they contract HIV. Diabetes prevention and treatment will have a positive impact on TB prevention. The TB Prevention Therapy Program is critical not only for children and people living with HIV/AIDS but also for family members living in the same house as TB patients, especially if they have DM.[10] TB screening is an endeavor to discover TB patients in a specific community, such as the diabetic population. Based on the Indonesian consensus in the management of TB-DM, TB screening procedures for diabetic individuals can be carried out at First Level Health Facilities/*Fasilitas Kesehatan Tingkat Pertama* (FKTP) and advanced referral health facilities by conducting interviews to observe TB symptoms or risk factors in the diabetics and chest x-ray examinations to examine abnormalities in the lungs. If the FKTP examination is not accessible, the patient is directed to an advanced referral health facility or network radiology laboratory.[11]

In the age of digital transformation, the Indonesian Ministry of Health (MoH) developed an Android-based application called WiFi TB in 2017 to streamline the notification reporting procedure.[12] This application is designed to allow private health facilities to record case findings and link them to primary health care (PHC). It includes tools for uploading data on TB suspects, TB cases, patient treatment monitoring, and patient treatment monitoring results reporting. Another application, called SOBAT TB, is used to share information regarding TB control among the general public, patients, communities, patient organizations, and medical personnel. This application has a self-filtering feature and is intended to provide accurate information about TB to the general public and facilitate access to TB health service facilities. Last, the EMPATI TB application is designed to help with TB therapy, especially for those who are drug-resistant. However, these applications are not used optimally. There are still problems with the WiFi TB application launched since its implementation because of the lack of human resource capability in entering data into the application, problems with network disruption, quota availability, and a lack of outreach of WiFi TB to private health facilities.[13] Because the SOBAT TB and EMPATI TB applications are relatively new, they are currently still in the evaluation stage and require community outreach in their implementation and use.[14]

TB Information System, the main software of the TB case recording and reporting system in Indonesia used by all stakeholders from primary to secondary health facilities (primary health care, hospitals, private practitioners, clinics, laboratories, and pharmacies), health offices, and the MoH, has not been optimally utilized. This condition is caused by insufficient human resources, less funding for recording and reporting programs in health institutions, inadequate integrated information systems, and poor internal networking.[15] As a result, the TB Information System must be strengthened so that health facilities, particularly FKTP, can use this existing application effectively.

This study analyzed the 2013 and 2018 Indonesian Basic Health Research (IBHR) data with a target population of patients with DM records to identify determinants of TB disease, which was used as a screening guideline and as reinforcement in creating information systems and digital transformation in the form of e-screening. This e-screening tool increases tuberculosis notification, particularly in private health facilities. The Orange Data Mining Tool application was used to test the screening tool model to predict TB in people with DM. Data mining is the process of collecting valuable or relevant information from large amounts of data to uncover certain patterns in the data and use it to predict or make judgments.

This program can also be utilized for machine learning for candidates in the fields of biology, biomedicine, and informatics. The data mining in the Orange software uses a widget system, where each widget has its function, which can receive input or produce output. Orange has data entry and filtering capabilities, categorization, evaluation with cross-validation and sampling-based processes, regression, and data visualization.[16] The final result of a series of model testing processes is accuracy, describing how accurately the e-model of the TB-DM screening tool classifies correctly, precision describing the accuracy between the requested data and the prediction results provided by the model, and recall or sensitivity describing the success of the model in retrieving information. This study provided benefits to TB-DM control programs by increasing the implementation of the TB Information System in private health facillities and improving information on TB suspects or new cases.

Method

A quantitative method was used in this study. The predictors of TB in diabetic individuals were studied using secondary data from the 2013 and 2018 IBHR. This study used Lemeshow's (1991) two-proportion difference hypothesis test formula to select the sample.[17] To find out the description of each variable, univariate data analysis was used in which the findings were reported in a distribution table. Data from the 2013 and 2018 IBHR were examined to determine the availability of data required, specifically data on TB risk factors in diabetic individuals (host and environmental factors) and the totals of TB and DM cases. The data were entered into the SPSS Version 26.0 (Licensed under IBM, New York, USA) data analysis tool, and data cleaning was performed to identify missing data or outliers prior to data analysis.

Furthermore, an analysis was performed using the Orange software Version 3.0 (general public license/free version) in March-April 2023 to test the screening tool model used to predict TB in diabetic individuals. In this study, regression analysis of predictor factors was performed using Orange. The first step was converting data entry into a dataset the file widget could read. The previous input dataset would be taken to the data table widget, where the input dataset was displayed in spreadsheet form to retrieve the data or filter the required data with the select row widget feature, producing suitable data. This data would be entered into a logistic regression model to determine the logistic regression coefficient and model regularization.

The model was evaluated using the test widget and scored through supervised machine learning, testing the learning algorithm on the data and displaying evaluation results that could be used to analyze the performance of the classifier using the confusion matrix widget, displaying the number/proportion of occurrences between predicted classes (predicted values) and actual class (actual value). The final results of this model testing process were "accuracy," describing how accurately the model was in classifying accurately; "precision," describing the accuracy between the requested data and the anticipated results provided by the model; and "recall," or "sensitivity," describing the success of the model in retrieving information.

The data consisted of several variables, each with a certain category. For pulmonary TB, categories included individuals diagnosed within the past year, more than a year ago, and those not diagnosed at all, with some missing data. The DM was categorized based on individuals diagnosed with diabetes, those undiagnosed, and missing data. Age was divided into seven groups: 12-17, 17-25, 25-35, 35-45, 45-55, 55-65, and over 65 years. Sex was categorized as male or female, with some missing data. Education level included individuals who were uneducated, not completing elementary school, attained elementary school, junior high school, senior high school, diploma (DI/DII/DIII), and higher education, with some missing data. Employment status was categorized into unemployed, students, civil servants/police/armed forces, private employees, self-employed, farmers, fishermen, laborers, other occupations, fresh graduates, and missing data. The weight and height measures variables indicated whether these measures were taken, with some missing data. The variables kidney failure diagnosis, cancer diagnosis, and rheumatic/joint disease diagnosis indicate whether a person was diagnosed with the condition, with some missing data.

Smoking records included regular and occasional smokers, former regular and occasional smokers, non-smokers, and missing data. The alcohol intake record was categorized by those consuming alcohol, not consuming alcohol, and missing data. Fasting blood sugar and temporary blood sugar variables were divided into controlled and uncontrolled categories. Ventilation variables for the bedroom, kitchen, and living room indicated whether the ventilation area was adequate (≥10% of the floor area), inadequate (<10% of the floor area), or nonexistent, with some missing data. The availability of health facilities, such as public and private hospitals, primary health care, clinics/physician practices, midwife/maternity homes, integrated health care, village health posts, and village maternity clinics, was indicated, with some respondents unsure or missing data.

In addition to univariate analysis, this study also used the Orange analysis to obtain the results of accuracy, precision, and recall calculations by using tests and scores which could be obtained through logistic regression, random forest, Support Vector Machine (SVM), and Naïve Bayes. Logistic regression is employed when the predicted variables are categorical in nature. Therefore, it is necessary to perform numeric encoding of the variables.[18] Random Forest is a multiple decision tree algorithm created in a randomized manner to ensure greater stability in the developed method.[19] The SVM can generate the input-output mapping function from labeled training data in the form of a classification function.

This function exhibits high simplification performance, thus deemed an effective classifier due to its independence from prior knowledge.[20]

Naïve Bayes comprises a collection of simple probabilistic algorithms based on Bayes' theorem, rendering it one of the most frequently utilized classification algorithms.[21] This model has good precision and recall at 0.9 due to imbalanced data, which shows that most do not have TB diagnoses, so the model will tend to predict it. Furthermore, the confusion matrix, which is a table containing predicted and actual values, is used for evaluation. The confusion matrix is useful for assessing the performance of classification models with two or more classes as output.[22] A Receiver Operating Characteristic Curve is constructed from this conflation matrix, which can be used to measure the overall diagnostic performance of a test and compare two or more diagnostic tests.[23]

Results

Table 1. Frequency Distribution Table

2013 Indonesian Basic Health Research			**2018 Indonesian Basic Health Research**		
Variable	**n**	**%**	**Variable**	**n**	**%**
Pulmonary Tuberculosis			**Pulmonary Tuberculosis**		
Yes, within the last ≤1 year	185	0.5	Yes, within the last ≤1 year	97	0.3
Yes, >1 year	749	2	Yes, >1 year	150	0.4
No	37,202	97.6	No	37,213	97.6
			Missing data	676	1.8
Diabetes Mellitus			**Diabetes Mellitus**		
Yes	795	2.1	Yes	1,061	2.8
No	37,341	97.4	No	35,399	95.4
			Missing data	1,676	1.8
Age			**Age**		
12–17 years	1,364	3.6	12–17 years	1,287	3.4
17–25 years	4,331	11.4	17–25 years	4,417	11.6
25–35 years	6,781	17.8	25–35 years	6,354	16.7
35–45 years	9,042	23.7	35–45 years	8,527	22.4
45–55 years	7,875	20.6	45–55 years	8,067	21.2
55–65 years	5,280	13.8	55–65 years	5,778	15.2
>65 years	3,481	9.1	>65 years	3,706	9.7
Sex			**Sex**		
Male	16,100	42.2	Male	15,386	40.3
Female	22,036	57.8	Female	22,074	57.9
			Missing data	676	1.8
Education Level			**Education Level**		
Uneducated	2,940	7.7	Uneducated	2,769	7.3
Not completing elementary school	5,532	14.5	Not completing elementary school	5,505	14.4
Elementary school	12,825	33.6	Elementary school	11,085	29.1
Junior high school	7,055	18.5	Junior high school	7,369	19.3
Senior high school	7,873	20.6	Senior high school	8,450	22.2
Diploma I/II/III	779	2	Diploma I/II/III	845	2.2
Higher education	1,132	3	Higher education	1,439	3.8
			Missing data	674	1.8
Employment Status			**Employment Status**		
Unemployed	13,326	34.9	Unemployed	12,955	34
Students	2,050	5.4	Students	2,098	5.5
Civil Servants/Police/Armed Forces	1,090	2.9	Civil servants/Police/Armed Forces	618	1.6
Private Employees	2,487	6.5	Private Employees	2,734	7.2
Self-employed	4,761	12.5	Self-employed	5,350	14
Farmers	8,134	21.3	Farmers	7,865	20.6
Fishermen	267	0.7	Fishermen	163	0.4
Laborers	4,127	10.8	Laborers	3,640	9.5
Other occupations	1,429	3.7	Other occupations	2,037	5.3
Fresh Graduates	465	1.2	Fresh Graduates	-	-
			Missing data	676	1.8
Weight Measurement			**Weight Measurement**		
Yes	38,004	99.7	Yes	37,318	97.9
No	132	0.3	No	142	0.4
			Missing data	676	1.8
Height Measurement			**Height Measurement**		
Yes	37,928	99.5	Yes	-	-
No	208	0.5	No	-	-
Cancer Diagnosis			**Cancer Diagnosis**		
Yes	90	0.2	Yes	107	0.3
No	38,046	99.8	No	37,353	97.9
			Missing data	676	1.8

2013 Indonesian Basic Health Research			2018 Indonesian Basic Health Research		
Variable	**n**	**%**	**Variable**	**n**	**%**
Kidney Failure Diagnosis			**Kidney Failure Diagnosis**		
Yes	114	0.3	Yes	163	0.4
No	38,022	99.7	No	37,297	97.8
			Missing data	676	1.8
Rheumatic/Joint Disease Diagnosis			**Rheumatic/Joint Disease Diagnosis**		
Yes	5,607	14.7	Yes	3,634	9.5
No	32,529	85.3	No	33,826	88.7
			Missing data	676	1.8
Smoking Record			**Smoking Record**		
Yes, every day (regular)	9,371	24.6	Yes, every day (regular)	9,524	25
Yes, sometimes (occasional)	1,962	5.1	Yes, sometimes (occasional)	2,998	7.9
No, but previously smoking every day	1,052	2.8	No, but previously smoking every day	-	-
No, but previously smoking sometimes	917	2.4	No, but previously smoking sometimes	-	-
Never at all	24,834	65.1	Never at all	24,938	65.4
			Missing data	676	1.8
Alcohol Intake Record			**Alcohol Intake Record**		
Yes	-	-	Yes	738	1.9
No	-	-	No	36,722	96.3
			Missing data	676	1.8
Fasting Blood Sugar			**Fasting Blood Sugar**		
Controlled	25,441	66.7	Controlled	29,060	76.2
Uncontrolled	12,695	33.3	Uncontrolled	9,076	23.8
Temporary Blood Sugar			**Temporary Blood Sugar**		
Controlled	37,864	99.3	Controlled	37,864	99.3
Uncontrolled	272	0.7	Uncontrolled	272	0.7
Bedroom Ventilation			**Bedroom Ventilation**		
Adequate, the area is ≥ 10% of the floor area	16,462	43.2	Adequate, the area is ≥10% of the floor area	16,515	43.3
Inadequate, the area is < 10% of the floor area	16,700	43.8	Inadequate, the area is <10% of the floor area	15,071	39.5
Nonexistent	4,974	13	Nonexistent	5,473	14.4
			Missing data	1,077	2.8
Kitchen Ventilation			**Kitchen Ventilation**		
Adequate, the area is ≥10% of the floor area	15,715	41.2	Adequate, the area is ≥10% of the floor area	14,611	38.3
Inadequate, the area is <10% of the floor area	16,088	42.2	Inadequate, the area is <10% of the floor area	14,131	37.1
Nonexistent	6,333	16.6	Nonexistent	7,838	20.6
			Missing data	1,556	4.1
Living Room Ventilation			**Living Room Ventilation**		
Adequate, the area is ≥ 10% of the floor area	18,505	48.5	Adequate, the area is ≥10% of the floor area	19,819	52
Inadequate, the area is < 10% of the floor area	15,398	40.4	Inadequate, the area is <10% of the floor area	12,489	32.7
Nonexistent	4,233	11.1	Nonexistent	3,482	9.1
			Missing data	2,346	6.2
Public Hospital			**Public Hospital**		
Available	26,664	69.9	Available within the district/city	30,754	80.6
Unavailable	11,472	30.1	Available in the nearest district/city	3,636	9.5
			Unavailable	531	1.4
			Do not know	2,539	6.7
			Missing data	676	1.8
Private Hospital			**Private Hospital**		
Available	20,568	53.9	Available within the district/city	30,754	80.6
Unavailable	17,568	46.1	Available in the nearest district/city	3,636	9.5
			Unavailable	531	1.4
			Do not know	2,539	6.7
			Missing data	676	1.8
Public Primary Health Care			**Primary Health Care**		
Available	34,628	90.8	Available within the district/city	34,316	90
Unavailable	3,508	9.2	Available in the nearest district/city	2,363	6.2
			Unavailable	64	0.2
			Do not know	717	1.9
			Missing data	676	1.8
Clinic/Physician Practice			**Clinic/Physician Practice**		
Available	20,987	55	Available within the district/city	30,437	79.8
Unavailable	17,149	45	Available in the nearest district/city	1,852	4.9
			Unavailable	1,405	3.7
			Do not know	3,766	9.9
			Missing data	676	1.8
Midwife/Maternity Home			**Midwife/Maternity Home**		
Available	25,655	67.3	Available	-	-
Unavailable	12,481	37.3	Unavailable	-	-
Integrated Health Care			**Integrated Health Care**		
Available	25,407	66.6	Available	-	-
Unavailable	12,729	33.4	Unavailable	-	-

2013 Indonesian Basic Health Research			**2018 Indonesian Basic Health Research**		
Variable	**n**	**%**	**Variable**	**n**	**%**
Village Health Post			**Village Health Post**		
Available	4,811	12.6	Available	-	-
Unavailable	33,325	87.4	Unavailable	-	-
Village Maternity Clinic			**Village Maternity Clinic**		
Available	5,677	14.9	Available	-	-
Unavailable	32,459	85.1	Unavailable	-	-

The selected respondents in this study were 38,136 people from each of the 2013 and 2018 IBHR data. Table 1 depicts the frequency distribution of TB risk factors according to the 2013 and 2018 IBHR. The 2013 IBHR revealed that 749 out of 38,136 participants (2%) had been diagnosed with pulmonary TB in the previous year, while, in the 2018 data, 97 out of 38,136 participants (0.3%) had been diagnosed with pulmonary TB in the previous year. DM affected 795 people (2.1%) in 2013 and 1,061 (2.8%) in 2018. More than half of respondents in the 2013 and 2018 IBHR were in the age group of 35-45 years at 23.7% and 22.4%, respectively. Furthermore, most respondents in this study, according to the 2013 and 2018 IBHR, were mostly females (at 57.8% and 57.9%, respectively), graduated from elementary school (33.6% in 2013 and 29.1% in 2018), and worked as farmers (at 21.3% and 20.6%, respectively).

Of 38,136 respondents, 99.7% were weighed in 2013, and 97.9% were weighed in 2018. In addition to weight, height was recorded in 2013, with 37,004 respondents (99.5%) having their height measured. In 2018, there was no data on height measurement. For cancer diagnosis, 90 respondents (0.2%) had the diagnosis in 2013 and 107 respondents (0.3%) in 2018. A total of 114 respondents (0.3%) in 2013 and 163 respondents (0.4%) in 2018 had been diagnosed with kidney failure for three months or longer. Then, 5,607 respondents (14.7%) in 2013 and 3,634 people (9.5%) in 2018 had rheumatic/joint disease.

Most respondents in 2013 and 2018 (65.1% and 65.4%, respectively) were non-smokers. There was no data on alcohol intake records in 2013; however, most respondents (96.3%) did not consume alcohol in the previous month in 2018. A total of 25,441 respondents (66.7%) in 2013 and 29,060 respondents (76.2%) in 2018 had control of their fasting blood sugar. In terms of temporary blood sugar, 99.3% of respondents in 2013 and 0.7% of respondents in 2018 had control of it.

In 2013, respondents had ventilation in their bedroom (43.2%), kitchen (41.2%), and living room (48.5%), with an area of 10% of floor area for each. While in 2018, respondents had an area of 10% of the floor area for bedroom ventilation at 43.3%, 38.3% for kitchen ventilation, and 52% for the living room. In terms of the availability of health facilities, 69.9% of respondents confirmed a public hospital available in their area in 2013, while 80.6% said there was a hospital in their district/city in 2018. In addition to public hospitals, the presence of private hospitals was explored, and 53.9% of respondents in 2013 and 80.6% in 2018 confirmed the availability of private hospitals in their area.

Most respondents (90.8%) claimed primary health care was available in their neighborhood in 2013; also, 90% (34,316) respondents confirmed the same in their district/city in 2018. In 2013, 20,987 individuals (55%) said there was a doctor's clinic in their neighborhood; in 2018, 30,437 people (79.8%) said there was a doctor's clinic in their district/city. In 2013, 25,655 people (67.3%) said there were midwife practices/maternity homes in their area, but there was no data on midwife practices/maternity homes in 2018. There were 25,407 respondents (66.6%) reporting that they found integrated health care in their neighborhood unit in 2013, but there was no data on the existence of integrated health care in 2018. In 2013, 33,325 people (87.4%) said there were no village health posts in their area, but there was no data on the existence of village health posts in 2018. In 2013, 85.1% of respondents said there were no village maternity clinics in their area, but there was no data on the existence of them in 2018.

Discussion

DM has been highlighted as a significant risk factor for pulmonary TB by the World Health Organization (WHO). Previous systematic review and meta-analysis studies have found that the incidence and prevalence of TB cases in patients with diabetes are quite high.[24] A previous study indicated that the risk of TB increased by 2-4 times in diabetic patients compared to non-diabetic individuals.[25] Diabetic individuals are at a higher risk of acquiring pulmonary TB due to weakened immunity, which increases the chance of infection in the lungs.[26] Another study found that TB is more likely to occur in diabetic individuals due to their reduced immune systems.[27] Early detection is critical in preventing TB infection in diabetic individuals; however, early detection of TB-DM is still uncommon, particularly in developing countries.[27] In Indonesia, there is no strategy for screening for TB-DM, even though diabetic individuals are three times more likely to get TB than people with HIV/AIDS.[28]

Patients with diabetes, cancer, and/or three of the five primary symptoms of TB are the objectives of e-screening. The same symptoms of pulmonary TB and lung cancer make identification and early detection difficult.[29] Diabetic individuals should be checked for TB, according to the WHO. Before developing a TB-DM screening tool, the model must be tested, which may be done using the Orange software to predict TB in diabetic individuals. This model aims to predict whether or not someone has TB. The TB variable in this study was divided into three categories: those diagnosed within the past year, more than a year ago, and those not diagnosed at all. The diagnosed and never-diagnosed categories have been combined into the "TB" category. In addition to Orange, other software such as WEKA, KNIME, and SPSS Modeler can be utilized, each with advantages and downsides. Orange has the advantage of displaying more interactive data analysis and data visualization.[30]

Orange software can be utilized to diagnose diseases. In a previous study, Orange was employed to determine the diagnosis of liver disease in individuals using methods such as Decision Tree, Random Forest, SVM, Neural Network, Naïve Bayes, K-Nearest Neighbors, and Logistic Regression. Through the confusion matrix, accuracy rate, precision rate, training time, and testing time were evaluated, yielding results indicating relatively brief testing and training times. Additionally, it was identified that the four best-performing methods based on accuracy rates are Logistic Regression, Neural Network, Random Forest, and Naïve Bayes.[31] A previous study has used Orange to analyze models to predict the quality-of-life domain of DM patients in a population using sensitivity, precision, and accuracy. The neural network model based on the Area Under Control (AUC), precision, accuracy, and ROC analysis values is the best model for predicting the utility of the quality-of-life domain of DM patients.[32]

In addition to evaluating models, Orange can be used to compare the classification algorithms of several models. A study by Mohi revealed that Orange was used to compare the classification algorithms of the decision tree, Naïve Bayes, and K-Nearest Neighbors models to classify two types of medical data to be tested based on previously conducted health tests.[33] Another study showed that Orange can be used to predict and classify benign and malignant breast cancer by using three classification models: Random Forest, Naïve Bayes, and AdaBoost. The prediction accuracy was 100%, 80%, and 80%, respectively.[34] Random Forest produced the best categorization results of the three models. In contrast, Orange can diagnose disease, predict outcomes, and evaluate models with high accuracy and efficient testing times. However, Orange also has some limitations. It offers fewer variants of data mining methods compared to other tools, and it does not support workflow modularity. Orange is effective for certain practical applications, but it may not be as efficient as other tools in terms of the variety of methods offered and the ability to manage complex workflows.[35]

Conclusion

This study's contribution will benefit the TB-DM control program. The developed e-screening tool has the potential to significantly enhance TB notification and control efforts, especially within private health facilities. Integrating this tool into existing health information systems makes it possible to improve the early detection and management of TB in diabetic individuals, thereby contributing to better health outcomes. The TB-DM e-screening tool will improve the implementation of the TB Information System, increasing TB notification at private health facilities. Future studies should focus on refining the model and expanding its application to other regions and populations to validate its effectiveness and scalability.

Abbreviations

TB: tuberculosis; CNR: case notification rate; TC: treatment coverage; DM: diabetes mellitus; FKTP: *Fasilitas Kesehatan Tingkat Pertama/*First Level Health Facilities; MoH: Ministry of Health; IBHR: Indonesian Basic Health Research; WHO: World Health Organization.

Ethics Approval and Consent to Participate

This study was approved by the Community Health Research and Ethics Commission, Faculty of Public Health, Universitas Indonesia (Reference 46/UN2.F10.D11/PPM.00.02/2023 dated March 03, 2023).

Competing Interest

The author declared no significant competing financial, professional or personal interests that might have affected the performance or presentation of the work described in this manuscript.

Availability of Data and Materials

The data used in this study were not publicly available. A reasonable request for the dataset can be sent to the National Institute of Health Research and Development/*Badan Penelitian dan Pengembangan Kesehatan* of the Indonesian Ministry of Health.

Authors' Contribution

MRB was responsible for the entire process, including the conceptualization, design, analysis, writing, and revision of the manuscript. AB was responsible for supervising the findings of this work. CC was involved in the analysis of the findings. All authors discussed and approved the final manuscript.

Acknowledgment

The authors would like to thank the Universitas Indonesia Postgraduate International Indexed Publication (PUTI) for supporting and assisting in the writing of this article.

References

1. World Health Organization. Global Tuberculosis Report 2023. Geneva: World Health Organization; 2023.
2. Kementerian Kesehatan Republik Indonesia. Laporan Program Penanggulangan Tuberkulosis Tahun 2022. Jakarta: Kementerian Kesehatan Republik Indonesia; 2023.
3. Badan Pusat Statistik Provinsi DKI Jakarta. Jumlah Kasus Penyakit Menurut Provinsi/Kabupaten/Kota dan Jenis Penyakit 2018-2021. Jakarta: Badan Pusat Statistik Provinsi DKI Jakarta; 2022.
4. Dinas Kesehatan Provinsi DKI Jakarta. Profil Kesehatan Provinsi DKI Jakarta Tahun 2022. Jakarta: Dinas Kesehatan Provinsi DKI Jakarta; 2023.
5. Kementerian Kesehatan Republik Indonesia. Dashboard TB Indonesia. Jakarta: Kementerian Kesehatan Republik Indonesia; 2024.
6. Kementerian Kesehatan Republik Indonesia. Strategi Nasional Penanggulangan Tuberkulosis di Indonesia 2020-2024. Jakarta: Kementerian Kesehatan Republik Indonesia; 2022.
7. Menteri Kesehatan Republik Indonesia. Peraturan Menteri Kesehatan Nomor 67 Tahun 2016 Tentang Penanggulangan Tuberkulosis. Jakarta: Kementerian Kesehatan Republik Indonesia; 2016.
8. Presiden Republik Indonesia. Peraturan Presiden Nomor 67 Tahun 2021 tentang Penanggulangan Tuberkulosis. Jakarta: Pemerintah Pusat; 2021.
9. Wadudah F, Prasetyowati I, Bumi C. Pelaksanaan wajib notifikasi (WiFi) TB di Dinas Kesehatan Kabupaten Jember. Pustaka Kesehat. 2023; 11 (1): 54-61. DOI: 10.19184/pk.v11i1.20358.
10. Ronacher K, Joosten SA, van Crevel R, et al. Acquired immunodeficiencies and tuberculosis: Focus on HIV/AIDS and diabetes mellitus. Immunol Rev. 2015; 264 (1): 121-137. DOI: 10.1111/imr.12257.
11. Kementerian Kesehatan Republik Indonesia. Konsensus pengelolaan tuberkulosis dan diabetes melitus (TB-DM) di Indonesia. Jakarta: Kementerian Kesehatan Republik Indonesia; 2015.
12. Kementerian Kesehatan Republik Indonesia. Strategi Nasional Penanggulangan Tuberkulosis di Indonesia 2020-2024. Jakarta: Kementerian Kesehatan Republik Indonesia; 2020.
13. Kurniadi A, Widianawati E, Kusuma EJ, et al. Implementasi aplikasi Wifi TB berdasarkan persepsi kemudahan dan kemanfaatan di Kota Semarang. J Kesehatan Vokasional. 2020; 5 (2): 102-109. DOI: 10.22146/jkesvo.50483.
14. Septiani D, Haniifah FN, Riswaluyo MA, et al. Optimalisasi aplikasi SOBAT TB dalam pelayanan TB selama masa pandemi COVID-19. J Biostat Kependuduk Inform Kesehat. 2022; 2 (2): 117-125. DOI: 10.51181/bikfokes.v2i2.5794.
15. Ratnasari Y, Sjaaf AC, Djunawan A. Evaluasi sistem pencatatan dan pelaporan kasus tuberculosis di Rumah Sakit Syarif Hidayatullah. J Manaj Kesehat Yayasan RS Dr. Soetomo. 2021; 7 (1): 115-124. DOI: 10.29241/jmk.v7i1.608.
16. Ishak A, Siregar K, Aspriyati A, et al. Orange Software Usage in Data Mining Classification Method on The Dataset Lenses. In: IOP Conf Ser Mater Sci Eng. 2020; 1003: 012113. DOI: 10.1088/1757-899X/1003/1/012113.
17. Egbuchulem KI. The basics of sample size estimation: An editor's view. Ann Ib Postgrad Med. 2023; 21 (1): 5-10.
18. Olson DL, Wu D. Predictive data mining models. 1st ed. Singapore: Springer Singapore; 2017.
19. Ratra R, Gulia P. Experimental evaluation of open source data mining tools (WEKA and Orange). Int J Eng Trends Technol. 2020; 68 (8): 30-35. DOI: 10.14445/22315381/ijett-v68i8p206s.
20. Dash S, Pani SK, Balamurugan S, et al. Biomedical data mining for information retrieval: Methodologies, techniques and applications. USA: Scrivener Publishing; 2021.
21. Zhou H. Learn data mining through Excel: A step-by-step approach for understanding machine learning methods. 2nd ed. Berkeley, CA: Apress; 2023.
22. Ikhromr FN, Sugiyarto I, Faddillah U, et al. Implementasi data mining untuk memprediksi penyakit diabetes menggunakan algoritma Naïve Bayes dan K-Nearest Neighbor. INTECOMS J Inform Technol Comput Sci. 2023; 6 (1): 416-428. DOI: 10.31539/intecoms.v6i1.5916.
23. Nahm FS. Receiver operating characteristic curve: Overview and practical use for clinicians. Korean J Anesthesiol. 2022; 75 (1): 25-36. DOI: 10.4097/kja.21209.
24. Wu Q, Liu Y, Ma YB, et al. Incidence and prevalence of pulmonary tuberculosis among patients with type 2 diabetes mellitus: A systematic review and meta-analysis. Ann Med. 2022; 54 (1): 1657-1666. DOI: 10.1080/07853890.2022.2085318.
25. Al-Rifai RH, Pearson F, Critchley JA, et al. Association between diabetes mellitus and active tuberculosis: A systematic review and meta-analysis. PLoS One. 2017; 12 (11): 1-26. DOI: 10.1371/journal.pone.0187967.
26. Carabalí-Isajar ML, Rodríguez-Bejarano OH, Amado T, et al. Clinical manifestations and immune response to tuberculosis. World J Microbiol Biotechnol. 2023; **39**: 206. DOI: 10.1007/s11274-023-03636-x.
27. Tenaye L, Mengiste B, Baraki N, et al. Diabetes mellitus among adult tuberculosis patients attending tuberculosis clinics in eastern Ethiopia. Biomed Res Int. 2019; 7640836. DOI: 10.1155/2019/7640836.
28. Alisjahbana B, McAllister SM, Ugarte-Gil C, et al. Screening diabetes mellitus patients for pulmonary tuberculosis: A multisite study in Indonesia, Peru, Romania and South Africa. Trans R Soc Trop Med Hyg. 2021; 115 (6): 634-643. DOI: 10.1093/trstmh/traa100.

29. Nugroho NP, Wati FF. Koeksistensi kanker paru dan tuberkulosis. Syifa' Medika J Kedokt Kesehat. 2020; 11 (1): 49-64. DOI: 10.32502/sm.v11i1.2583.
30. Hosseini S, Sardo SR. Data mining tools: A case study for network intrusion detection. Multimed Tools Appl. 2021; 80 (4): 4999-5019. DOI: 10.1007/s11042-020-09916-0.
31. Hikmah IR, Yasa RN. Perbandingan hasil prediksi diagnosis pada Indian Liver Patient Dataset (ILPD) dengan teknik supervised learning menggunakan software Orange. J Telematik. 2021; 16 (2): 69-76. DOI: 10.61769/telematika.v16i2.402.
32. Arifin B, Perwitasari DA, Zulkarnain Z, et al. Models for predicting the quality of life domains on the general population through the orange data mining approach. Pharmaciana. 2022; 12 (1): 72-82. DOI: 10.12928/pharmaciana.v12i1.20827.
33. Mohi ZR. Orange data mining as a tool to compare classification algorithms. Dijlah J Sci Eng. 2020; 3 (3): 13-23.
34. Imran B, Hambali H, Subki A, et al. Data mining using random forest, naïve bayes, and adaboost models for prediction and classification of benign and malignant breast cancer. J Pilar Nusa Mandiri. 2022; 18 (1): 37-46. DOI: 10.33480/pilar.v18i1.2912.
35. Dobesova Z. Evaluation of Orange data mining software and examples for lecturing machine learning tasks in geoinformatics. Comput Appl Eng Educ. 2024; 32 (4): 1-18. DOI: 10.1002/cae.22735.