Implementation of the Naïve Bayes Algorithm to Predict the Severity Status of Tuberculosis Patients in Dr. M. Djamil Central General Hospital

Nurul Abdillah[1], Alfita Dewi[2], Herman Susilo[3], Dede Fauzi[4]
[1,2,3,4] Syedza Saintika University, Padang, West Sumatra, Indonesia

Article Info

Article history:

Received April 20, 2025
Revised May 20, 2025
Accepted June 20, 2025

Keywords:

Tuberculosis (TB)
Prediction
Naïve Bayes Classifier
Data Mining

ABSTRACT

Tuberculosis (TB) remains a serious public health issue in Indonesia, including in Padang City. Accurate and timely intervention is crucial, especially in predicting the severity status of TB patients to enable faster and more appropriate medical responses. Currently, hospital systems still face limitations in classifying the severity status of TB patients, which may result in delayed treatment. This study aims to apply the Naïve Bayes Classifier (NBC) algorithm to predict the severity status of TB patients in Padang City. The medical record data used were obtained from Dr. M. Djamil Central General Hospital and Dr. Rasidin Regional General Hospital, consisting of 227 TB patient records from April to June 2024. The dataset includes 13 attributes, such as gender, age, type of cough, shortness of breath, chest pain, and severity status as the class attribute. The results show that the NBC algorithm achieved an accuracy of 71.81% in predicting patient severity status. This study is expected to support healthcare professionals in making more informed decisions for patient management planning.

Corresponding Author:

Nurul Abdillah
Bachelor of Applied Health Information Management, Syedza Saintika University
Jl. Prof. Dr. Hamka No. 228 Air Tawar Timur, Padang, West Sumatera, Indonesia
E-mail: Abdillahadik15@gmail.com

1. INTRODUCTION

Tuberculosis (TB) remains one of the deadliest infectious diseases worldwide, including in Indonesia. According to the Global Tuberculosis Report 2023 by the World Health Organization (WHO), Indonesia ranks second in the world for the highest number of TB cases, following India [1]. In Padang City, the prevalence of TB also shows an alarming trend with a high number of new cases reported each year [2]. This situation highlights the need for more effective and efficient strategies, particularly in rapidly identifying the severity status of TB patients.

The severity status of TB patients is a crucial aspect of medical management as it determines the speed and level of medical intervention provided. Delays in identifying patient severity can increase the risk of serious complications, wider transmission, and even death [3]. Therefore, hospitals and healthcare facilities require decision support systems capable of providing fast and accurate predictions of patient conditions based on available medical data.

In the digital era, the use of information technology in healthcare (e-health) offers a potential solution to improve the quality of healthcare services. One emerging approach is the application of data mining techniques, particularly classification methods, to analyze patient medical records and support clinical decision-making [4]. Classification algorithms such as Naïve Bayes have proven effective in various studies due to their simplicity and ability to generate reasonably accurate predictive results [5].

The Naïve Bayes Classifier (NBC) is a probabilistic machine learning algorithm widely used for data classification across various fields, including healthcare. NBC operates by calculating the probability of class occurrence based on the features or attributes present in the input data [6]. One of NBC's advantages is its ability to perform well on relatively small datasets and produce classification results that can be easily interpreted by healthcare professionals.

Several previous studies have implemented the Naïve Bayes algorithm in the healthcare sector. For instance, research by Sari et al. (2022) demonstrated that NBC could predict complications in diabetic patients with a good level of accuracy [7]. Similarly, a study conducted in a regional hospital in Yogyakarta found that NBC was effective in predicting inpatient risk based on initial symptoms recorded in medical records [8]. These findings suggest that NBC has potential to be applied in TB cases, particularly for determining the severity status of patients.

However, the specific application of NBC in predicting the severity status of TB patients remains limited, especially in regions such as Padang City. This presents an opportunity for further research to explore the effectiveness of this method in supporting clinical classification and decision-making processes. Digital medical record data from hospitals can serve as valuable information sources in developing predictive models.

This study aims to apply the Naïve Bayes algorithm to predict the severity status of TB patients in Padang City, using data from two major referral hospitals: Dr. M. Djamil Central General Hospital and Dr. Rasidin Regional General Hospital. The dataset includes key attributes such as age, gender, and clinical symptoms (e.g., cough, shortness of breath, and chest pain), along with the severity status label. The prediction results are expected to assist doctors in making more timely and appropriate decisions.

Therefore, this research is expected to make a real contribution to the use of information technology for improving the quality of healthcare services, particularly in the early detection of TB patient severity. Furthermore, the findings may serve as a preliminary reference for the development of data mining–based decision support systems in hospital environments.

2. METHOD

This study employed a quantitative approach using a data mining classification method, specifically the Naïve Bayes Classifier (NBC) algorithm. This method was selected due to its ability to process relatively limited data while still producing reasonably accurate predictions. NBC operates on a probabilistic principle based on Bayes' Theorem, assuming the independence of attributes within the dataset. The model calculates the probability of each class based on the features of each data instance and assigns the class with the highest probability as the prediction result [9].

The data used in this study were obtained from the medical records of tuberculosis (TB) patients at two referral hospitals in Padang City, namely Dr. M. Djamil Central General Hospital and Dr. Rasidin Regional General Hospital. The dataset covers the period from April to June 2024 and includes a total of 227 patient records that were cleaned and preprocessed. A total of 13 attributes were used as input variables, including age, gender, type of cough, shortness of breath, and chest pain, along with one class attribute—severity status (mild, moderate, severe). The data were processed using RapidMiner software for training and testing the classification model using the Naïve Bayes algorithm [10].

Conducting research must follow a structured and systematic research framework. The research framework followed in this study is outlined below:

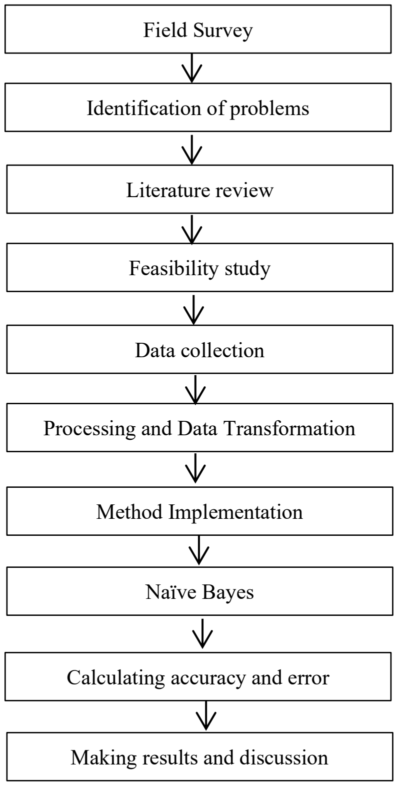

Field Survey
Identification of problems
Literature review
Feasibility study
Data collection
Processing and Data Transformation
Method Implementation
Naïve Bayes
Calculating accuracy and error
Making results and discussion

Figure 1. Research Framework

1. Description of the Research Framewor
   a. Conducting Field Survey
   Before starting the research, a field survey was conducted to obtain a qualitative overview of the severity status of patients diagnosed with TB in Padang City. The survey activities were carried out to collect data and conduct observations by directly visiting the relevant parties involved in this research.
   b. Identifying the Problem
   The problem identification stage is where the research object formulates the research problems. Problem formulation is carried out to determine the existing issues at the research site and to set boundaries for the problems to be studied. Identifying the problem is the initial step in the research, where we must determine the variables or attributes that influence the severity status of TB patients.
   c. Conducting Literature Review
   To achieve the intended objectives, it is necessary to study several related literature sources. The literature review is conducted to enrich the theoretical foundation of this research by reading journals and reference books, so that data and information analysis becomes more accurate.
   d. Conducting Feasibility Study
   The feasibility study in this research is carried out to assess whether the development of a web-based application for predicting the severity status of Tuberculosis (TB) patients in Padang City can be properly implemented. The assessment includes technical feasibility to ensure technologies such as the CodeIgniter framework and RapidMiner can be used effectively, operational feasibility to evaluate the team's capabilities and infrastructure support for application development and implementation, financial feasibility to ensure the available budget is sufficient, and legal and ethical feasibility to ensure that the use of patient data complies with applicable regulations and standards. This study is important to ensure the project can run smoothly and achieve the desired outcomes.
   e. Data Collection
   Data collection was carried out using several methods, namely:

1) Direct observation method, which involved observing directly the Medical Record unit to obtain the necessary data.
2) Interview method, which involved conducting interviews with individuals directly related to the issues being studied in this research to gain insights and explanations.
3) Literature review method, which involved referring to relevant sources such as data references and scholarly literature.
4) Browsing method, which involved collecting references from internet-based sources.

f. Data Processing and Transformation

At the Data Processing and Transformation stage, raw data will be converted and integrated into a suitable format for processing using a data mining application, namely RapidMiner.

g. Method Implementation

After the data processing and transformation, the next stage is testing. This testing process requires computer hardware and software to carry out the implementation effectively.

h. Naïve Bayes Algorithm

In this stage, the previously proposed method will be implemented. The algorithm will be tested using RapidMiner software to evaluate the accuracy level and error rate of the Naïve Bayes algorithm implementation as a data mining classification method. The formula for Bayes' Theorem is as follows:

$$P(H|X) = \frac{\mathrm{P}(X|H)\,P(H)}{P(X)} \quad (1)$$

Naïve Bayes is a simplified form of the Bayes method. Bayes' Theorem is simplified as follows:

$$(H|X) = P(X|H)P(H) \quad (2)$$

Explanation:

X : A data sample with an unknown class (label)
H : The hypothesis that X belongs to a particular class (label)
P(H) : The probability of the hypothesis H (prior probability)
P(X) : The probability of observing the data sample X (evidence)
P(X|H): The probability of observing data sample X given that the hypothesis H is true (likelihood)

i. Calculating Accuracy and Error

At this stage, the Accuracy and Error values of the Naïve Bayes algorithm will be calculated to evaluate how close the predicted values are to the actual or true values. The higher the Accuracy, the closer the classification results are to the correct outcome. The lower the Error, the more accurate the classification results. Accuracy and Error are defined as the degree of closeness between predicted values and actual values. The formulas for calculating Accuracy and Error are as follows:

Accuracy = Jumlah yang diklasifikasi secara bena*r*/Total sampel testing yang diuji

Error = Number of incorrectly classified samples / Total number of testing samples

j. Generating Results and Discussion

The purpose of the results and discussion section is to present an overview and the findings obtained from this study. After conducting data analysis using data mining techniques, the research results can be described. The outcomes of the design phase were then implemented into a system using the CodeIgniter framework and the RapidMiner tool. The output generated from this implementation phase is a web-based application that can be used by healthcare workers and patients to predict, at an early stage, the severity status of patients diagnosed with tuberculosis (TB).

3. RESULTS AND DISCUSSION

The data used in this study consist of tuberculosis (TB) patient records from RSUP Dr. M. Djamil and RSUD Dr. Rasidin Padang, located in Padang City, covering the period from May to July 2024. A total of 227 medical records of patients who had been diagnosed with TB were analyzed.

The attributes used in the analysis include Gender, Age, Type of Cough, Shortness of Breath, Chest Pain, Fatigue, Fever, Loss of Appetite, Weight Loss, Night Sweats, Decreased Consciousness, Smoking History, and Patient Severity Status, as shown in Table 1.

Table 1. Training Data Details

No	Total	Number of Cases		Severe	Non-Severe
		227		81	146
1	Gender	L	155	60	95
		P	72	21	51
2.	Age	Adolescent	5	0	5
		Young Adult	44	18	26
		Middle Adult	91	31	60
		Elderly	87	32	55
3.	Type of Cough	Productive Cough	148	55	93
		Hemoptysis	41	13	28
		Dry Cough	38	13	25
4.	Shortness of Breath	Yes	164	74	90
		No	63	7	56
5.	Chest Pain	Yes	109	39	70
		No	118	42	76
6.	Fatigue	Yes	97	39	58
		No	130	42	88
7.	Fever	Yes	102	36	66
		No	125	45	80
8.	Appetite Loss	Yes	158	65	93
		No	69	16	53
9.	Weight Loss	Yes	158	69	89
		No	69	12	57
10.	Night Sweats	Yes	64	19	45
		No	163	62	101
11.	Decreased Consciousness	Yes	23	20	3
		No	204	61	143
12.	Smoking History	Yes	93	31	62
		No	134	50	84

1. Naïve Bayes Classification
   a. Calculating Class Prior Probabilities

      After creating a detailed data table, the first step in the Naïve Bayes calculation is to compute the Class Prior Probabilities for *Severe* and *Not Severe*.

$$P(X = Severe) = \frac{\mathrm{P}(X \cap Y)}{P(Y)}$$

$$= \frac{\mathrm{P}(Decision\ =\ Severe \cap Total\ Data)}{P(Y = Total\ Data)}$$

$$= \frac{81}{227} = \ 0{,}357$$

$$P(X = Tidak\ Gawat) = \frac{\mathrm{P}(X \cap Y)}{P(Y)}$$

$$= \frac{\mathrm{P}(Decision\ =\ Non\,Severe \cap Total\ Data)}{P(Y = Total\ Data)}$$

$$= \frac{146}{227} = \ 0{,}643$$

b. Calculating Conditional Probabilities

In the conditional calculation, the probabilities of *Severe* and *Not Severe* are computed for each variable or attribute.

1) Calculation of Conditional Probabilities for the Gender Attribute

At this stage, the conditional probabilities for the gender attribute (Male and Female) will be calculated for both the *Severe* and *Not Severe* classes.

$$P(JK = M|Severe) = \frac{\mathrm{P}(X \cap Y)}{P(Y)} = \frac{P(JK\ =\ M \cap Severe)}{P(Y\ =\ Severe)} = \frac{60}{81} = \ 0{,}74$$

$$P(JK = M|Non\ Severe) = \frac{\mathrm{P}(X \cap Y)}{P(Y)} = \frac{P(JK\ =\ M\ \cap Non\ Severe)}{P(Y\ =\ Non\ Severe)} = \frac{95}{146} = \ 0{,}65$$

$$P(JK = M|Severe) = \frac{\mathrm{P}(X \cap Y)}{P(Y)} = \frac{P(JK\ =\ F \cap Severe)}{P(Y\ =\ Severe)} = \frac{21}{81} = \ 0{,}26$$

$$P(JK = M|Non\ Severe) = \frac{\mathrm{P}(X \cap Y)}{P(Y)} = \frac{P(JK\ =\ F\ \cap Non\ Severe)}{P(Y\ =\ Non\ Severe)} = \frac{51}{146} = \ 0{,}35$$

1. Calculating Conditional Probabilities for Other Attributes
2. Determining the Decision Class of the Case

Setelah After calculating the prior and conditional probabilities, the next step is to perform Naive Bayes calculations aimed at predicting the decision for a given case.

Case Example:

A patient is male, 35 years old, has a productive cough, shortness of breath, loss of appetite, weight loss, night sweats, decreased consciousness, and a history of smoking.

c. Probability of Severe (Severe)

P(H|X) x P(X) = ((P(Gender=Male | Severe) × P(Age=Early Adulthood | Severe) × P(Type of Cough=Productive Cough | Severe) × P(Shortness of Breath=Yes | Severe) × P(Chest Pain=No | Severe) × P(Weakness=No | Severe) × P(Fever=No | Severe) × P(Loss of Appetite=Yes | Severe) × P(Weight Loss=Yes | Severe) × P(Night Sweats=Yes | Severe) × P(Decreased Consciousness=Yes | Severe) × P(Smoking History=Yes | Severe)) × P(X = Severe)
= (0,74 x 0,22 x 0,68 x 0,91 x 0,52 x 0,56 x 0,8 x 0,85 x 0,23 x 0,25 x 0,38) x 0,36
= 0,000157

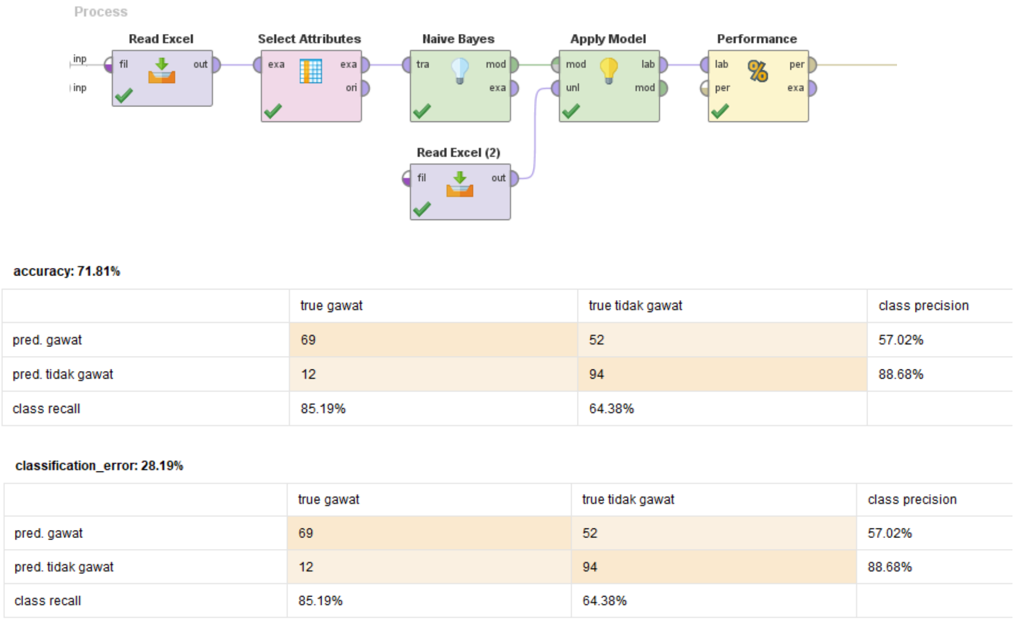
2. Probability Non Severe

P(H|X) × P(X) = ((P(Gender=Male | Severe) × P(Age=Early Adulthood | Severe) × P(Type of Cough=Productive Cough | Severe) × P(Shortness of Breath=Yes | Severe) × P(Chest Pain=No | Severe) × P(Weakness=No | Severe) × P(Fever=No | Severe) × P(Loss of Appetite=Yes | Severe) × P(Weight Loss=Yes | Severe) × P(Night Sweats=Yes | Severe) × P(Decreased Consciousness=Yes | Severe) × P(Smoking History=Yes | Severe)) × P(X = Severe)
= (0,65 x 0,18 x 0,64 x 0,62 x 0,52 x 0,6 x 0,55 x 0,64 x 0,61 x 0,31 x 0,02 x 0,42) x 0,64
= 0,00000518

Based on the calculation of the probabilities for severe and non-severe cases, the predicted decision is **severe**, because the probability value for the severe class is higher than that of the non-severe class.

3. NBC Accuracy and Error

Calculating accuracy and error aims to evaluate the accuracy and error values of predictions compared to the actual or assumed correct values. The higher the accuracy value, the closer the classification is to being correct; conversely, the lower the error value, the more accurate the classification becomes. The number of correctly classified data points is 164 out of the total dataset (71.81%), while the number of incorrectly classified data points is 63 (28.19%).



accuracy: 71.81%

	true gawat	true tidak gawat	class precision
pred. gawat	69	52	57.02%
pred. tidak gawat	12	94	88.68%
class recall	85.19%	64.38%	

classification_error: 28.19%

	true gawat	true tidak gawat	class precision
pred. gawat	69	52	57.02%
pred. tidak gawat	12	94	88.68%
class recall	85.19%	64.38%	

Figure 2. Display of Accuracy and Error Values in RapidMiner

4. CONCLUSION

The development of a web-based application using the CodeIgniter framework and the Naïve Bayes Classifier (NBC) algorithm has successfully produced a prediction model with an accuracy rate of 71.81% based on data from 227 tuberculosis-diagnosed patients to predict the severity status of tuberculosis patients in Padang City. Although the accuracy level is not yet optimal, this application is expected to assist medical personnel in making faster and more accurate decisions regarding the treatment of TB patients based on the predicted severity status in Padang City.

REFERENCES

[1] World Health Organization. (2023). Global Tuberculosis Report 2023. Geneva: WHO Press.

[2] Dinas Kesehatan Provinsi Sumatera Barat. (2023). Laporan Tahunan Kasus TBC Kota Padang. Padang: Dinkes Sumbar.

[3] Suryani, T., & Wibowo, A. (2021). Analisis Faktor Keterlambatan Penanganan Pasien TBC di Rumah Sakit. Jurnal Kesehatan Masyarakat, 12(2), 110–118.

[4] Nugroho, A., & Prasetyo, E. (2020). Penerapan Data Mining untuk Prediksi Penyakit Menggunakan Metode Klasifikasi. Jurnal Teknologi dan Sistem Komputer, 8(3), 325–332.

[5] Kurniawan, D., & Yuliani, L. (2019). Perbandingan Algoritma Naïve Bayes dan Decision Tree dalam Memprediksi Penyakit. Jurnal Ilmu Komputer dan Informatika, 5(1), 45–53.

[6] Han, J., Kamber, M., & Pei, J. (2012). Data Mining: Concepts and Techniques (3rd ed.). Elsevier.

[7] Sari, R., & Rahmawati, D. (2022). Prediksi Komplikasi Diabetes Menggunakan Algoritma Naïve Bayes. Jurnal Teknologi Kesehatan, 10(1), 88–96.

[8] Widyastuti, S., & Fauzan, M. (2021). Penerapan Naïve Bayes untuk Prediksi Risiko Rawat Inap di RSUD. Jurnal Sistem Informasi Kesehatan, 9(2), 74–81.

[9] Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques* (3rd ed.). Elsevier.

[10] Witten, I. H., Frank, E., & Hall, M. A. (2016). *Data Mining: Practical Machine Learning Tools and Techniques* (4th ed.). Morgan Kaufmann.