**p-ISSN 2085-7829**
**e-ISSN 2798-3455**

# Klasifikasi Penyakit Tuberkulosis Menggunakan Metode Naive Bayes (Studi Kasus: Data Pasien Di Puskesmas Petung Kabupaten Penajam Paser Utara)

## *Classification of Tuberculosis Using the Naive Bayes Method (Case Study: Patient Data at Puskesmas Petung Penajam Paser Utara)*

**Ahmad Aliful A.[1,a)], Rito Goejantoro[1,b)], dan M. Fathurahman[2,c)]**
[1]Laboratorium Statistika Komputasi FMIPA Universitas Mulawarman
[2]Laboratorium Statistika Terapan FMIPA Universitas Mulawarman

[a)]*Corresponding author:* ahmadalifulabidin@gmail.com
[b)] rto.mtk@gmail.com
[c)] fathur@fmipa.unmul.ac.id

***ABSTRACT***

*The Naive Bayes method is one of the data mining methods used in classifying data and predicting future opportunities based on experience or previous data. This method was proposed by British scientist Thomas Bayes using a branch of mathematics known as probability theory. One of the diseases that can be detected using the classification using the Naive Bayes method is Tuberculosis (TB). Tuberculosis is an infectious respiratory disease caused by the bacterium Mycobacterium tuberculosis. The purpose of this study was to determine the results and accuracy of the classification of Tuberculosis disease using the Naive Bayes method in one of the health service units, namely Puskesmas Petung, Penajam Paser Utara. The results showed that the classification of data mining using the Naive Bayes method was appropriate in classifying Tuberculosis. For training data and testing data which are divided into 90:10, the accuracy rate is 87.5% or categorized as Excellent Classification. As for the training data and testing data which is divided into 70:30, the accuracy rate is 90.9% or categorized as Excellent Classification.*

**Keywords:** *Classification, data mining, Naive Bayes, Tuberculosis*

## 1. Pendahuluan

Perkembangan dan kemajuan teknologi di era globalisasi memberikan pengaruh yang signifikan baik dalam bidang teknologi informasi, industri, kesehatan bahkan pendidikan di Indonesia. Pengaruh ini membawa suatu perubahan besar dalam berbagai bidang terutama dalam lingkup ilmu pengetahuan, salah satunya adalah statistika. Statistika merupakan ilmu dasar yang harus dikuasai seseorang dalam melakukan segala macam penelitian di dalam disiplin ilmu, sehingga hampir seluruh instansi memerlukan ilmu statistika dalam pengolahan data. Salah satu analisis statistika yang saat ini banyak digunakan oleh instansi adalah *data mining* yang digunakan untuk menggali informasi.

*Data mining* merupakan proses menelusuri pengetahuan terbaru, pola, dan tren yang dipilih dari jumlah data yang besar dan disimpan dalam repositori atau tempat penyimpanan dengan menggunakan teknik pengenalan pola statistik dan matematika. *Data mining* kemudian dikenal dengan nama *Knowledge Discovery in Databases* (KDD) merupakan kegiatan yang meliputi pengumpulan, pemakaian data historis untuk memecahkan pola atau hubungan keteraturan dalam dataset yang berukuran besar. Banyak metode yang dapat dilakukan menggunakan *data mining*, salah satunya adalah metode klasifikasi. Metode klasifikasi merupakan salah satu metode *data mining* yang merupakan suatu pengelompokan data untuk memprediksi nilai dari sekelompok atribut dalam menggambarkan dan membedakan kelas label atau target yang bertujuan untuk memprediksi kelas dari objek yang lebel kelasnya tidak diketahui. proses klasifikasi *data mining* didasarkan pada empat komponen mendasar, yaitu: kelas, atribut, training dataset, dan testing dataset (Gorunescu ,2011).

Metode Naive Bayes adalah salah satu metode *data mining* yang digunakan dalam mengklasifikasi data dan dapat memecahkan masalah dari metode-metode klasifikasi lainnya. Metode Naive Bayes dikemukakan oleh ilmuwan Inggris, Thomas Bayes. Metode ini memprediksi peluang di masa depan berdasarkan pengalaman atau data sebelumnya. Metode Naive Bayes menggunakan cabang matematika yang dikenal dengan teori probabilitas untuk mencari peluang terbesar dari kemungkinan klasifikasi dengan cara melihat frekuensi setiap klasifikasi pada data training. Klasifikasi ini didasarkan pada teorema Bayes yang memiliki kemampuan serupa dengan metode *decision tree* dan *neural network* (Kusrini & Luthfi, 2009).

Klasifikasi *data mining* dapat digunakan dalam berbagai bidang, salah satunya adalah bidang kesehatan dengan mengklasifikasi data penyakit yang dapat membantu petugas medis dalam mengambil keputusan diagnosis penyakit tersebut. Diagnosis dini perlu dilakukan agar dapat mengurangi penularan dan tingkat kematian masyarakat serta sebagai tindakan pencegahan sehingga dapat memungkinkan penderita penyakit dapat disembuhkan lebih cepat. Salah satu penyakit yang dapat dideteksi menggunakan klasifikasi

**p-ISSN 2085-7829**
**e-ISSN 2798-3455**

dengan menggunakan metode Naive Bayes adalah Tuberkulosis (TB). TB adalah penyakit saluran pernafasan menular yang disebabkan oleh bakteri yang disebut *mycobacterium tuberculosis* dan merupakan penyakit yang paling menular serta merupakan penyebab kematian tertinggi setelah stroke.

Indonesia merupakan negara dengan penderita penyakit TB terbesar ketiga setelah India dan Cina. Setiap tahunnya terjadi sekitar 245.000 kasus TB baru dengan jumlah kematian mencapai 46.000 setiap tahunnya. Kalimantan Timur merupakan salah satu provinsi di Indonesia dengan peningkatan jumlah kasus baru yang signifikan. Provinsi Kalimantan Timur berada pada urutan ke 20 pada angka notifikasi kasus TB per 100.000 penduduk di Indonesia pada tahun 2020. Sedangkan, Kabupaten Penajam Paser Utara total kasus TB hingga tahun 2020 mencapai 188 kasus. Meskipun penyakit ini dapat disembuhkan dan dicegah, hal itu tetap menjadikan penyakit TB sebagai ancaman global yang sangat serius termasuk salah satunya di Kabupaten Penajam Paser Utara (Profil Kesehatan Indonesia, 2020).

Penelitian yang dilakukan Saputra dan Widodo (2014) tentang diagnosis awal tuberkulosis paru menggunakan metode ensembel C4.5, Naive Bayes, *Neural Network*, dan *Logistic Regression*, diperoleh bahwa metode Naive Bayes memiliki ketepatan klasifikasi yang tertinggi, yaitu 91,61%. Oleh karena itu, penelitian ini akan menerapkan metode Naive Bayes dalam memprediksi diagnosis penyakit TB yang merupakan metode yang akurat dalam klasifikasi *data mining*.

## 2. Tinjauan Pustaka

### 2.1 *Data Mining*

*Data mining* merupakan proses menelusuri pengetahuan terbaru, pola, dan tren yang dipilih dari jumlah data yang besar dan disimpan dalam repositori atau tempat penyimpanan dengan menggunakan teknik pengenalan pola statistik dan matematika. *Data mining* berisi pencarian tren atau pola yang diinginkan dalam database besar untuk membantu pengambilan keputusan di waktu yang akan datang (Hermawati, 2013).

### 2.2 Klasifikasi *Data Mining*

Klasifikasi adalah metode *data mining* yang dapat digunakan untuk proses pencarian sekumpulan fungsi yang dapat menjelaskan dan membedakan kelas-kelas data atau konsep. Tujuannya supaya metode tersebut dapat digunakan memprediksi objek kelas yang labelnya tidak diketahui atau dapat memprediksi kecenderungan data-data yang muncul di masa depan. Metode klasifikasi juga bertujuan untuk melakukan pemetaan data ke dalam kelas yang sudah didefinisikan sebelumnya berdasarkan pada nilai atribut data (Han & Kamber, 2006).

### 2.3 Naive Bayes untuk Klasifikasi

Menurut Saleh (2015), metode Naive Bayes merupakan pengklasifikasian dengan metode probabilitas dan statistik yang dikemukan oleh ilmuwan Inggris Thomas Bayes, yaitu memprediksi peluang di masa depan berdasarkan pengalaman di masa sebelumnya. Metode ini mengasumsikan semua atribut independen atau tidak saling ketergantungan yang diberikan oleh nilai pada kelas atau label. Metode ini menggunakan cabang matematika yang dikenal dengan teori probabilitas untuk mencari peluang terbesar dari beberapa kemungkinan klasifikasi, dengan melihat frekuensi tiap klasifikasi pada *data training*. Metode Naive Bayes merupakan pengklasifikasian statistik yang digunakan untuk memprediksi probabilitas keanggotaan suatu kelas. Klasifikasi bayesian didasarkan pada teorema bayes yang memiliki kemampuan klasifikasi serupa dengan *decision tree* dan *neural network* (Dewi, 2016).

Konsep dasar teorema Bayes yaitu melakukan klasifikasi dengan melakukan perhitungan nilai probabilitas $P(C|X)$, yaitu probabilitas kelas $C$ jika diketahui data $X$. Rumus dari teorema Bayes menurut Prasetyo (2012) adalah seperti di bawah ini:

$$P(C \mid X) = \frac{P(X \mid C)P(C)}{P(X)} \tag{1}$$

di mana:

$X$ : Atribut dengan kelas yang belum diketahui.
$C$ : Kelas spesifik.
$P(C|X)$ : Probabilitas kelas $C$ bersyarat kondisi $X$ (*posterior probability*).
$P(C)$ : Probabilitas kelas C (*prior probability*).
$P(X|C)$ : Probabilitas $X$ bersyarat kondisi pada $C$.
$P(X)$ : Probabilitas dari $X$

Untuk menjelaskan teorema Bayes, perlu diketahui bahwa proses klasifikasi memerlukan sejumlah atribut untuk menentukan kelas yang tepat bagi objek yang dianalisis tersebut. Oleh karena itu, teorema Bayes di atas disesuaikan sebagai berikut:

$$P(C \mid X_1,...,X_q) = \frac{P(C)P(X_1,...,X_q \mid C)}{P(X_1,...,X_q)} \tag{2}$$

Penjabaran lebih lanjut dari teorema Bayes dapat dilakukan dengan menjabarkan $P(C|X_1,...,X_q)$ mengunakan

**p-ISSN 2085-7829**
**e-ISSN 2798-3455**

aturan perkalian berikut:

$$\begin{aligned} P\left(C\middle|X_1,...,X_q\right) &= P(C)\,P\left(X_1,...,X_q\middle|C\right) \\ &= P(C)\,P\left(X_1\middle|C\right)P\left(X_2\middle|C,X_1\right) \\ &\quad P\left(X_3\middle|C,X_1,X_2\right) \\ &\quad ...P\left(X_q\middle|C,X_1,X_2,...,X_{q-1}\right) \end{aligned} \tag{3}$$

Dapat dilihat bahwa hasil penjabaran tersebut menyebabkan semakin banyak dan semakin kompleksnya faktor-faktor syarat yang mempengaruhi nilai probabilitas mengakibatkan perhitungan tersebut sulit untuk dilakukan. Oleh karena itu digunakan asumsi independensi yang sangat tinggi (*naive*) dengan mengasumsikan semua atribut saling bebas satu sama lain. Dengan asumsi tersebut, berlaku suatu persamaan sebagai berikut:

$$\begin{aligned} P\left(X_i\middle|X_j\right) &= \frac{P\left(X_i \cap X_j\right)}{X_j} \\ &= \frac{P\left(X_i\right) \cap P\left(X_j\right)}{P\left(X_j\right)} = P\left(X_j\right) \end{aligned} \tag{4}$$

Untuk $i \neq j$, sehingga

$$P\left(X_i\middle|C,X_j\right) = P\left(X_i\middle|C\right) \tag{5}$$

Dari persamaan (5), dapat disimpulkan bahwa asumsi independensi tersebut membuat syarat peluang menjadi sederhana, sehingga perhitungan menjadi mudah untuk dilakukan. Selanjutnya penjabaran $P(C|X_1,...,X_q)$ dapat disederhanakan menjadi

$$\begin{aligned} P\left(C\middle|X_1,...,X_q\right) &= P(C)\,P\left(X_1\middle|C\right)P\left(X_2\middle|C\right)...\left(X_q\middle|C\right) \\ &= P(C)\prod_{i=1}^{q} P(X_i \mid C) \end{aligned} \tag{6}$$

### 2.4 *Laplace Correction*

Menurut Imanidanantoyo, Ananta dan Kirana (2020), pada proses klasifikasi bisa saja terdapat probababilitas yang bernilai 0 (nol) dan dapat menyebabkan metode Naive Bayes tidak dapat mengklasifikasi sebuah data dengan baik. Oleh karena itu, digunakan teknik *Laplace Correction* yaitu suatu teknik yang menambahkan nilai 1 pada setiap kombinasi atribut. Untuk jumlah *record* yang banyak hingga ribuan, teknik ini akan sangat akurat karena tidak akan membuat perbedaan yang berarti pada estimasi probabilitas. Rumus dari *Laplace Correction* dinyatakan sebagai berikut:

$$P(X_i \mid C) = \frac{P(X_i \mid C)+1}{P(C)+|V|} \tag{7}$$

dimana:
$P(X_i|C)$ : Probabilitas tiap atribut $X_i$.
$P(C)$ : Total probabilitas dalam $X_i$.
$|V|$ : Banyak kategori nilai $X_i$.

### 2.5 Evaluasi *Confusion Matrix* dan *Receiver Operating Charateristic* (ROC) *Curve*

Menurut Rosandy (2016), *confusion matrix* adalah suatu metode untuk melakukan evaluasi dengan menggunakan tabel matriks yang digunakan pada konsep *data mining* untuk melakukan perhitungan akurasi. Evaluasi dengan menggunakan fungsi *confusion matrix* akan menghasilkan nilai *accuracy* yang merupakan persentase dari jumlah *record* data yang diklasifikasikan secara baik dan benar dengan menggunakan sebuah metode dan dapat membuat klasifikasi setelah dilakukan pengujian hasil klasifikasi tersebut, *precision* yang merupakan proporsi dari jumlah kasus yang diprediksi mendapatkan hasil positif di mana nilainya juga akan positif pada data sebenarnya, dan *recall* atau *sensivity value* merupakan proporsi dari jumlah kasus yang bernilai positif yang sebenarnya dan diprediksi positif secara benar.

**Tabel 1.** *Confusion Matrix*

| ***Actual*** | ***Classified as*** | |
|---|---|---|
| | + | - |
| + | *True Positives* | *False Negative* |
| - | *False Positives* | *True Negatives* |

Setelah data uji dimasukkan ke dalam *confusion matrix*, maka dihitung nilai-nilai yang telah dimasukkan tersebut untuk dihitung jumlah *accuracy*, *sensitivity*, *spesificity*, *Positive Predictive Value* (PPV)

**p-ISSN 2085-7829**
**e-ISSN 2798-3455**

dan *Negative Predictive Value* (NPV) berdasarkan rumus berikut

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{8}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{9}$$

$$Spesificity = \frac{TN}{TN + FP} \tag{10}$$

$$PPV = \frac{TP}{TP + FP} \tag{11}$$

$$NPV = \frac{TN}{TN + FN} \tag{12}$$

Sedangkan fungsi ROC *Curve* adalah untuk memperlihatkan akurasi dan membandingkan klasifikasi secara visual. ROC mengekspresikan *Confusion Matrix*. ROC merupakan grafik dua dimensi dengan garis horizontal sebagai *false positive* dan garis vertikal sebagai *true positive*. (Buani, 2016).

Menurut Gronescu (2011), visualisasi dengan ROC *curve* memiliki beberapa kategori yang dapat dilihat melalui luas area dibawah kurva atau *Area Under Curve* (AUC). Berikut tingkat nilai diagnosis dari ROC *curve*, yaitu:

**Tabel 2.** Kategori nilai ROC *Curve*

| Nilai | Kategori |
|---|---|
| 0,90 – 1,00 | *Excellent Classification* |
| 0,80 – 0,90 | *Good Classification* |
| 0,70 – 0,80 | *Fair Classification* |
| 0,60 – 0,70 | *Poor Classification* |
| 0,50 – 0,60 | *Failure* |

### 2.6 Tuberkulosis (TB)

Kata "Tuberkulosis" diciptakan oleh Johann Lukas Schonle pada tahun 1839 dari Bahasa Latin "*tuberculum*" yang berarti benjolan kecil, pembengkakan atau jerawat. TB disebabkan oleh suatu penyakit infeksi kuman menular yang mempunyai nama yang sama yaitu bakteri *Mycobacterium Tuberculosis*. Kuman atau bekteri ini dapat menyerang berbagai organ terutama paru-paru, penyakit TB ini bila tidak diobati atau pengobatannya tidak tuntas atau tidak selesai maka dapat menimbulkan komplikasi berbahaya hingga kematian (Ekata, 2016).

## 3. Metode Penelitian

### 3.1. Sumber Data

Penelitian ini menggunakan rancangan yang bersifat *ex post facto* dengan sampel yang digunakan adalah pasien yang positif TB dan negatif TB di Puskesmas Petung, Kabupaten Penajam Paser Utara tahun 2020-2021. Data yang digunakan merupakan data sekunder yang diperoleh dari data internal puskesmas berupa rekapitulasi medik pasien yang terdiri dari sepuluh variabel dengan salah satunya adalah hasil atau kelas yang diambil dari hasil laboratorium.

### 3.2. Variabel Penelitian

Variabel yang digunakan dalam penelitian ini terdiri dari sepuluh variabel yaitu sembilan variabel atribut yang terdiri dari lama batuk ($X_1$), jenis batuk ($X_2$), sesak nafas ($X_3$), sakit dada ($X_4$), lemas ($X_5$), demam ($X_6$), hilang nafsu makan ($X_7$), berat badan menurun ($X_8$), dan keringat malam ($X_9$).dan satu variabel hasil atau kelas yaitu hasil diagnosis ($C$).

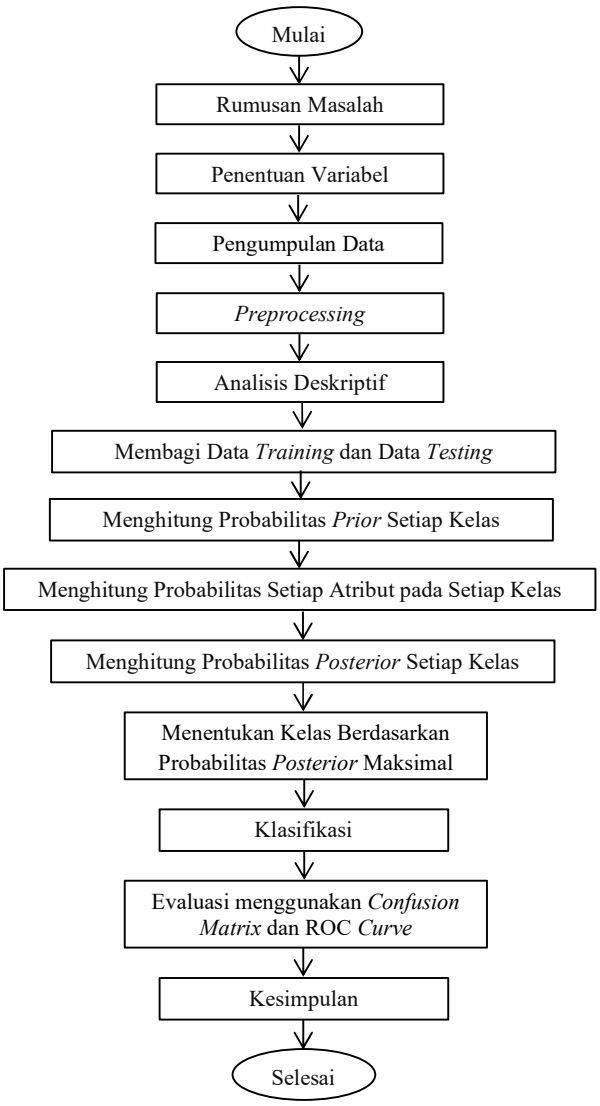
### 3.3. Teknik Sampling

Teknik sampling yang digunakan dalam penelitian ini adalah teknik *purposive sampling.* Dalam hal ini, pertimbangannya adalah ketersediaan dan kelengkapan data terbaru rekam medik pasien TB pada puskesmas. Selain itu, meningkatnya jumlah kasus TB di Provinsi Kalimantan Timur selama sepuluh tahun terakhir termasuk di Kabupaten Penajam Paser Utara juga menjadi pertimbangan dalam menentukan sampel.

### 3.4. Teknik Analisis Data

Teknik analisis data yang digunakan dalam penelitian ini adalah analisis statistika deskriptif kemudian data diolah dan diuji menggunakan metode Klasifikasi Naive Bayes. Setelah itu hasil analisis yang diperoleh dari metode Naive Bayes dihitung tingkat akurasinya menggunakan *Confusion Matrix* dan ROC *Curve*. Secara umum, tahapan analisis pada penelitian ini adalah sebagai berikut:

**p-ISSN 2085-7829**
**e-ISSN 2798-3455**




**Gambar 1.** Tahapan Analisis Data

## 4. Hasil Dan Pembahasan

Data yang digunakan dalam penelitian ini adalah pasien Tuberkulosis di Puskesmas Petung, Kabupaten Penajam Paser Utara tahun 2020-2021. Jumlah data yang diperoleh pada penelitian ini sebanyak 86 *record*, baik pasien positif TB dan negatif TB, tetapi data tersebut masih terdapat *missing values* sehingga perlu dilakukan *preprocessing* dengan teknik *data size reduction* untuk mendapatkan data yang berkualitas dan informatif. Setelah dilakukan *preprocessing*, data yang awalnya sebanyak 86 record direduksi dengan menghapus data yang terdapat *missing value* menjadi 72 *record*.

### 4.1 Analisis Statistika Deskriptif

Analisis statistika deskriptif dilakukan untuk mengetahui gambaran umum pasien Tuberkulosis. Gambaran umum pada analisis statistika deskriptif meliputi persentase dari variabel-variabel yang digunakan antara lain diagnosis pasien TB, lama batuk, jenis batuk, sesak nafas, sakit dada, lemas, demam, hilang nafsu makan, berat badan menurun, dan keringat malam yang dapat dilihat pada Tabel 4.

**p-ISSN 2085-7829**
**e-ISSN 2798-3455**

**Tabel 3.** Statistika Deskriptif

| Variabel | Kategori | Persentase |
|---|---|---|
| Hasil Diagnosis | Negatif | 51,39% |
| | Positif | 48,61% |
| Lama Batuk | Tidak Batuk | 6,94% |
| | < 1 Minggu | 27,78% |
| | 1 Minggu - 1 Bulan | 43,06% |
| | > 1 Bulan | 22,22% |
| Jenis Batuk | Tidak Batuk | 6,94% |
| | Batuk Kering | 44,45% |
| | Batuk Berdahak | 20,83% |
| | Batuk Berdarah | 27,78% |
| Sesak Nafas | Ya | 72,22% |
| | Tidak | 27,78% |
| Sakit Dada | Ya | 51,39% |
| | Tidak | 48,61% |
| Lemas | Tidak Lemas | 15,28% |
| | ≤ 1 minggu | 61,11% |
| | > 1 minggu | 23,61% |
| Demam | Tidak Demam | 23,61% |
| | ≤ 1 minggu | 58,33% |
| | > 1 minggu | 18,06% |
| Hilang Nafsu Makan | Ya | 56,94% |
| | Tidak | 43,06% |
| Berat Badan Menurun | Ya | 34,72% |
| | Tidak | 65,28% |
| Keringat Malam | Ya | 33,33% |
| | Tidak | 66,67% |

Sebelum dilakukan klasifikasi menggunakan metode Naive Bayes, data yang diperoleh dibagi menjadi *data training* dan *data testing* sebesar 90:10 dan 70:30. Dari 72 data dengan proporsi sebesar 90:10, 64 data yang diacak menggunakan *software R* berfungsi sebagai *data training* dan 8 data lainnya berfungsi sebagai *data testing*. Sedangkan untuk proporsi sebesar 70:30, 50 data berfungsi sebagai *data training* dan 22 data lainnya berfungsi sebagai *data testing.*

### 4.2 Klasifikasi Naive Bayes

Klasifikasi *data mining* menggunakan metode Naive Bayes terdiri dari tiga tahapan, yaitu menghitung probabilitas *prior* setiap kelas, menghitung probabilitas setiap atribut pada setiap kelas, dan menghitung probabilitas *posterior* setiap kelas. Tahapan yang pertama yaitu probabilitas *prior* setiap kelas menggunakan data training berdasarkan persamaan (4). Adapun nilai probabilitas *prior* atau probabilitas awal setiap kelas atau diagnosis pasien Tuberkulosis dengan kategori negatif TB dan positif TB dapat dilihat pada Tabel 4.

**Tabel 4.** Probabilitas *Prior* Hasil Diagnosis

| Diagnosis | Probabilitas | |
|---|---|---|
| | 90:10 | 70:30 |
| Negatif | 0,531 | 0,5 |
| Positif | 0,469 | 0,5 |

Tahapan selanjutnya yaitu menghitung probabilitas setiap atribut pada setiap kelas yaitu hasil diagnosis berdasarkan persamaan (5). Perhitungan nilai probabilitas setiap atribut yaitu lama batuk ($X_1$), jenis batuk ($X_2$), sesak nafas ($X_3$), sakit dada ($X_4$), lemas ($X_5$), demam ($X_6$), hilang nafsu makan ($X_7$), berat badan menurun ($X_8$), dan keringat malam ($X_9$).

**p-ISSN 2085-7829**
**e-ISSN 2798-3455**

**Tabel 5.** Probabilitas Atribut pada Hasil Diagnosis

| Atribut | Kategori | Probabilitas | | | |
|---|---|---|---|---|---|
| | | 90:10 | | 70:30 | |
| | | Negatif | Positif | Negatif | Positif |
| Lama Batuk | Tidak Batuk | 0,088 | 0,067 | 0,08 | 0,08 |
| | < 1 Minggu | 0,412 | 0,167 | 0,44 | 0,04 |
| | 1 Minggu - 1 Bulan | 0,441 | 0,400 | 0,4 | 0,44 |
| | > 1 Bulan | 0,059 | 0,367 | 0,08 | 0,44 |
| Jenis Batuk | Tidak Batuk | 0,088 | 0,033 | 0,08 | 0,04 |
| | Batuk Kering | 0,647 | 0,300 | 0,6 | 0,36 |
| | Batuk Berdahak | 0,265 | 0,100 | 0,32 | 0,08 |
| | Batuk Berdarah | 0,000 | 0,567 | 0,00 | 0,52 |
| Sesak Nafas | Ya | 0,529 | 0,900 | 0,52 | 0,84 |
| | Tidak | 0,471 | 0,100 | 0,48 | 0,16 |
| Sakit Dada | Ya | 0,382 | 0,700 | 0,36 | 0,60 |
| | Tidak | 0,618 | 0,300 | 0,64 | 0,40 |
| Lemas | Tidak Lemas | 0,265 | 0,033 | 0,28 | 0,00 |
| | ≤ 1 minggu | 0,676 | 0,567 | 0,64 | 0,52 |
| | > 1 minggu | 0,059 | 0,400 | 0,08 | 0,48 |
| Demam | Tidak Demam | 0,412 | 0,033 | 0,36 | 0,04 |
| | ≤ 1 minggu | 0,588 | 0,633 | 0,64 | 0,64 |
| | > 1 minggu | 0,000 | 0,333 | 0,00 | 0,32 |
| Hilang Nafsu Makan | Ya | 0,353 | 0,767 | 0,32 | 0,76 |
| | Tidak | 0,647 | 0,233 | 0,68 | 0,24 |
| Berat Badan Menurun | Ya | 0,235 | 0,433 | 0,28 | 0,44 |
| | Tidak | 0,765 | 0,567 | 0,72 | 0,56 |
| Keringat Malam | Ya | 0,118 | 0,567 | 0,16 | 0,52 |
| | Tidak | 0,882 | 0,433 | 0,84 | 0,48 |

Tahapan selanjutnya adalah menghitung probabilitas *posterior* setiap kelas atau hasil perkalian probabilitas *prior* dan probabilitas setiap atribut pada setiap kelas berdasarkan persamaan (6). Pada perhitungan ini akan digunakan salah satu *data testing* dengan atribut atau gejala batuk kurang dari satu minggu, batuk berdahak, mengalami sesak nafas, tidak mengalami sakit dada, lemas kurang dari satu minggu, tidak demam, tidak kehilangan nafsu makan, berat badan tidak turun, dan tidak berkeringat pada malam hari.

**Tabel 6.** Probabilitas *Posterior*

| Diagnosis | Probabilitas | |
|---|---|---|
| | 90:10 | 70:30 |
| Negatif | 0,00231 | 0,00292 |
| Positif | 0,00000226 | 0,000345 |

Berdasarkan perhitungan probabilitas atribut atau gejala yang dialami oleh pasien Tuberkulosis, gejala yang dominan adalah batuk kurang dari satu minggu, batuk kering, mengalami sesak nafas namun tidak mengalami sakit dada, lemas dan demam kurang dari satu minggu, tidak kehilangan nafsu makan, berat badan tidak berkurang dan tidak berkeringat pada malam hari sedangkan untuk pasien yang didiagnosis positif TB gejalanya didominasi oleh batuk dalam rentang satu minggu sampai satu bulan, batuk berdarah, mengalami sesak nafas dan sakit dada, lemas dan demam kurang dari satu minggu, kehilangan nafsu makan, berat badan tidak berkurang, dan berkeringat pada malam hari. Berdasarkan perhitungan probabilitas *posterior*, dapat diketahui bahwa probabilitas pasien yang didagnosis negatif TB lebih besar dibandingkan probabilitas positif TB sehingga pasien tersebut dapat diklasifikasikan ke dalam pasien yang didiagnosis negatif TB.

### 4.3 Evaluasi Klasifikasi

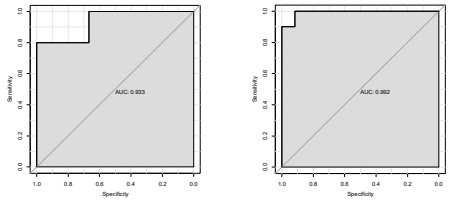
Setelah dilakukan klasifikasi pasien Tuberkulosis menggunakan metode Naive Bayes akan dilakukan evaluasi menggunakan *data testing*. Evaluasi yang digunakan dalam klasifikasi *data mining* pasien Tuberkulosis menggunakan metode Naive Bayes adalah *Confusion Matrix* dan *Receiver Operating Charateristic* (ROC) *Curve* terhadap masing-masing data training dan data testing dengan proporsi 90:10 dan 70:30. Untuk *Confusion Matrix*, dapat dihitung nilai *Accuracy*, *Sensitivity*, *Spesificity*, *Positive Predictive Value*, dan *Negative Predictive Value* menggunakan persamaan (8) sampai dengan (12).

**p-ISSN 2085-7829**
**e-ISSN 2798-3455**

**Tabel 7**. Nilai *Accuracy*, *Sensitivity*, *Spesificity*, PPV, dan NPV

| Nilai | 90:10 | 70:30 |
|---|---|---|
| *Accuracy* | 0,875 | 0,909 |
| *Sensitivity* | 1 | 0,917 |
| *Spesificity* | 0,8 | 0,9 |
| PPV | 0,75 | 0,917 |
| NPV | 1 | 0,9 |

Berdasarkan Tabel 7, dapat diketahui bahwa untuk proporsi 90:10, diperoleh nilai *Accuracy, Spesificity*, PPV lebih besar dibandingkan proporsi 70:30 sehingga dapat dikatakan bahwa proporsi 70:30 lebih baik dibandingkan proporso 90:30. Untuk ROC *Curve* dari *data training* dan *data testing* dengan proporsi 90:10 dan 70:30 dapat dilihat pada Gambar 1.




**(a)** **(b)**

**Gambar 2.** Evaluasi Menggunakan ROC *Curve* (a) Proporsi 90:10 dan (b) Proporsi 70:30

Berdasarkan Gambar 2, dapat diketahui bahwa ROC *Curve* untuk proporsi 90:10 memiliki kurva yang lebih dekat dengan garis diagonal dibandingkan proporsi 70:30 sehingga dapat diartikan bahwa luas dibawah kurva lebih besar untuk proporsi 70:30. Luas tersebut juga disebut nilai *Area Under Curve* (AUC) yang dapat dilihat pada Tabel 8.

**Tabel 8.** Nilai AUC

| Proporsi | AUC | Kategori |
|---|---|---|
| 90:10 | 0,933 | *Excellent Classification* |
| 70:30 | 0,992 | *Excellent Classification* |

## 5. Kesimpulan

Berdasarkan hasil dan pembahasan, maka kesimpulan yang dapat diambil adalah sebagai berikut:

1. Metode Naive Bayes dapat digunakan untuk mengklasifikasi pasien Tuberkulosis di Puskesmas Petung, Kabupaten Penajam Paser Utara.
2. Tingkat akurasi klasifikasi menggunakan metode Naive Bayes untuk proporsi 90:10 adalah sebesar 87,5% dan dari evaluasi menggunakan ROC *Curve* dapat dikategorikan sebagai *Excellent Classification*. Sedangkan untuk proporsi 70:30 diperoleh tingkat akurasi sebesar 90,9% dan dari evaluasi menggunakan ROC *Curve* dapat dikategorikan sebagai *Excellent Classification*.

## Daftar Pustaka

Buani, D. P. (2016). Optimasi Algoritma Naïve Bayes dengan Menggunakan Algoritma Genetika untuk Prediksi Kesuburan (Fertility). *Jurnal Evolusi*, 4(1), 54-63.

Dewi, A. (2016). Komparasi 5 Metode Algoritma Klasifikasi Data Mining Pada Prediksi Keberhasilan Pemasaran Produk Layanan Perbankan. *Jurnal Techno Nusa Mandiri*, 8(1), 7-9.

Ekata, Tyagi, P. K., Gupta, N. K., & Gupta, S. (2016). Diagnosis of Pulmonary Tuberculosis using Fuzzy Inference System. IEEE Second International Innovative Applications of Computational Intelligence on Power, Energy and Controls with their Impact on Humanity, 3-7. doi: 10.1109/CIPECH.2016.7918726.

Gorunescu, F. (2011). *Data Mining: Concepts, Model and Techniques*. Romania.

Han, J. dan M. Kamber. (2006). *Data Mining Concepts and Techniques Second Edition*. San Francisco: Morgan Kaufmann.

Hermawati. (2013). *Data Mining*. Yogyakarta: Penerbit ANDI.

Imanidanantoyo, A. I., Ananta, A. Y., & Kirana, A. P. (2020). Implementasi Naive Bayes Dan Pos Tagging Menggunakan Metode Hidden Markov Model Viterbi Pada Analisa Sentimen Terhadap Akun Twitter Presiden Joko Widodo Di Saat Pandemi COVID-19. *Seminar Informatika Aplikatif Polinema*, 235–241.

Kementerian Kesehatan Republik Indonesia. (2021). *Profil Kesehatan IndonesiaTahun 2020*. Jakarta :

**p-ISSN 2085-7829**
**e-ISSN 2798-3455**

Kementerian Kesehatan RI.

Kusrini, dan Luthfi, E.M. (2009). *Algoritma Data Mining*. Yogyakarta : CV. Andi Offset.

Prasetyo, E., (2012). *Data Mining Konsep dan Aplikasi Menggunakan Matlab*. Yogyakarta: Andi Offset.

Rosandy, T. (2016). Perbandingan Metode Naive Bayes Classifier dengan Metode Decison Tree (C4.5) Untuk Menganalisa Kelancaran Pembiayaan (Study Kasus: KSPPS/BMT Al-Fadhila). *Jurnal TIM Darmajaya*, 2(1), 52-62.

Saleh, Alfa. (2015). *Implementasi Metode Klasifikasi Naïve Bayes Dalam Memprediksi Besarnya Penggunaan Listrik Rumah Tangga*. Jakarta: Sumber Utama.

**p-ISSN 2085-7829**
**e-ISSN 2798-3455**