Biner : Jurnal Ilmu Komputer , Teknik dan Multimedia
Volume 1, No. 3, Agustus 2023
ISSN 2988-3814 (media online)
Hal 745-759

Implementasi Data Mining Untuk Diagnosa Prediksi Penyakit Tuberculosis Dengan Menggunakan Algoritma *Naïve Bayes*

Ezza Eka Pramana[1], Aries Saifudin[1*]

[1]Fakultas Ilmu Komputer, Program Studi Teknik Informatika, Universitas Pamulang, Tangerang Selatan, Indonesia
Email: [1]ezaprambanan@gmail.com, [2*]aries.saifudin@yahoo.co.id

Abstrak − Tuberkulosis (TB) masih menjadi masalah kesehatan yang signifikan di Indonesia, dan deteksi dini kasus TB sangat penting untuk mencegah penyebaran penyakit ini. Metode klasifikasi Naive Bayes telah terbukti efektif dalam klasifikasi kasus penyakit, termasuk TB. Tujuan penelitian ini adalah untuk menerapkan metode klasifikasi Naïve Bayes pada data kasus TB di Indonesia dan mengevaluasi kinerja model klasifikasi yang dihasilkan. Data kasus TB diperoleh dari Rumah Sakit Sari Asih Ciputat dan diproses dengan teknik pengolahan data dan pembersihan data. Kemudian, dilakukan pembagian data menjadi data latih dan data uji. Metode klasifikasi Naive Bayes diimplementasikan pada data latih dan kemudian dievaluasi dengan menggunakan data uji Hasil dari penelitian ini menunjukkan bahwa metode klasifikasi Naive Bayes dapat diaplikasikan pada kasus TB di Indonesia dengan akurasi klasifikasi yang baik. Faktor-faktor yang paling mempengaruhi klasifikasi kasus TB adalah geografi, usia, dan jenis kelamin. Penelitian ini diharapkan dapat membantu meningkatkan deteksi dini kasus TB di Indonesia dan memperbaiki upaya pencegahan dan pengobatan penyakit ini.

Kata Kunci: Sistem pakar, *Naïve Bayesian*, Tuberkulosis

Abstract *− Tuberculosis (TB) is still a significant health problem in Indonesia, and early detection of TB cases is very important to prevent the spread of this disease. The Naive Bayes classification method has been proven effective in classifying disease cases, including TB. The purpose of this study was to apply the Naive Bayes classification method to TB case data in Indonesia and evaluate the performance of the resulting classification model. TB case data were obtained from Sari Asih Ciputat Hospital and processed using data processing and data cleaning techniques. Then, the data is divided into training data and test data. The Naive Bayes classification method is implemented on training data and then evaluated using test data. The results of this study indicate that the Naive Bayes classification method can be applied to TB cases in Indonesia with good classification accuracy. The factors that most influence the classification of TB cases are geography, age, and gender. This research is expected to help improve the early detection of TB cases in Indonesia and improve efforts to prevent and treat this disease.*

Keywords*: Expert systems, Naïve Bayesian, Tuberculosis*

1. PENDAHULUAN

Tuberculosis (TBC) adalah penyakit menular paru-paru yang disebabkan oleh hasil *Mycobacterium Tuberculosis*. Penyakit ini ditularkan dari penderita TBC aktif yang batuk dengan mengeluarkan titik-titik kecil air liur dan terinhalasi dari orang sehat yang tidak memiliki kekebalan dari penyakit ini. TBC termasuk salah satu dalam 10 besar penyakit yang menyebabkan kematian di dunia. Berdasarkan WHO Global TBC Report 2020, kasus TBC di Indonesia pada tahun 2019 diperkirakan sejumlah 845.000 kasus dengan insidensi 312 per 100.000 penduduk yang kemudian membawa Indonesia menjadi negara dengan jumlah kasus terbesar ke 2 di dunia setelah India (Kementrian Kesehatan Republik Indonesia, 2021).

TB dapat diklasifikasikan bedasarkan lokasi anatomi atau lokasi organ tubuh yang terserang TB. Tuberkulosis paru yaitu yang terjadi pada parenkim (jaringan paru). TB milier dianggap sebagai TB paru karena adanya lesi pada jaringan paru. Limfadenitis TB rongga dada (hilus dan atau mediastinum) atau efusi pleura tanpa terdapat gambaran radiologis yang mendukung TB pada paru, dinyatakan sebagai TB paru. Pasien yang menderita TB paru dan sekaligus juga menderita TB Ekstra paru, diklasifikasikan sebagai pasien TB paru (Isbaniah, 2021)

Data yang digunakan adalah data rekam medis yang dimiliki oleh Rumah Sakit Sari Asih Ciputat yang nantinya akan diolah dan dianalisis untuk melihat ciri-ciri penderita TB. Dari pengolahan data tersebut akan diperoleh suatu pola gejala. Untuk pengolahan data rekam medis,

Biner : Jurnal Ilmu Komputer , Teknik dan Multimedia
Volume 1, No. 3, Agustus 2023
ISSN 2988-3814 (media online)
Hal 745-759

hasil pemeriksaan formulir skrinning TB dan hasil pemeriksaan laboratorium, digunakanlah suatu teknologi data mining salah satunya Naïve Bayesian. Data mining yang dimaksud untuk memberikan hasil dalam pengambilan keputusan di dunia kesehatan untuk mengidentifikasi suatu penyakit.

Data mining adalah proses untuk mendapatkan informasi yang berguna dari basis data yang besar dan perlu diekstraksi agar menjadi informasi baru dan dapat membantu dalam pengambilan keputusan. Data mining digunakan mencari informasi bisnis berharga dari basis data yang sangat besar, yang dipakai untuk memprediksi tren dan sifat-sifat bisnis serta menemukan pola-pola yang tidak diketahui sebelumnya (Suntoro, 2019).

Metode *Naïve Bayes* adalah metode klasifikasi statistik yang dapat memprediksi kelas suatu anggota probabilitas, algoritma ini memanfaatkan teori probabilitas yang dikemukakan oleh ilmuan Inggris yaitu memprediksi probabilitas di masa depan berdasarkan pengalaman di masa sekarang (Amalia, 2020). Penerapan metode *Naïve Bayes* ini diharapkan dapat membantu menghitung probabilitas pada sampel data untuk mengidentifikasi adanya penyakit Tuberculosis (TB).

2. METODE

2.1 Data Mining

Sebagai bidang ilmu yang baru, saat ini Data Mining menjadi salah satu pusat perhatian para akademis maupun praktisi. Data mining adalah proses untuk mendapatkan informasi yang berguna dari basis data yang besar dan perlu diekstraksi agar menjadi informasi baru dan dapat membantu dalam pengambilan keputusan (Suntoro, 2019).

Data Mining adalah kegiatan menemukan pola yang menarik dari data dalam jumlah besar, data dapat disimpan dalam database, data warehouse, atau penyimpanan informasi (Saputro, 2019). Data Mining memiliki suatu rangkaian proses yang harus dilakukan sebelum dapat memperoleh informasi baru. Tahap dalam data mining adalah sebagai berikut:

a. *Data cleaning*

   Pembersihan dari merupakan proses menghilangkan noise dan data yang tidak konsisten.

b. *Data integration*

   Proses dimana menggabungkan data dari berbagai macam sumber data. Proses ini dilakukan ketika menggunakan sumber data yang lebih dari satu.

c. *Data selection*

   Proses menyeleksi data dimana data yang akan digunakan dalam proses data mining diambil dan membiarkan data yang tidak digunakan.

d. *Data transformation*

   Proses mengubah data ke dalam bentuk yang dapat digunakan dalam perhitungan suatu algoritma.

e. *Data mining*

   Proses menemukan pola dari dataset yang digunakan sebagai basis pengetahuan.

f. Pattern evaluation

   Merupakan proses menganalisis hasil dari proses mining menggunakan suatu satuan ukur.

g. *Knowledge presentation*

   Merupakan proses untuk menampilkan hasil dari proses mining.

Biner : Jurnal Ilmu Komputer , Teknik dan Multimedia
Volume 1, No. 3, Agustus 2023
ISSN 2988-3814 (media online)
Hal 745-759

2.2 Klasifikasi

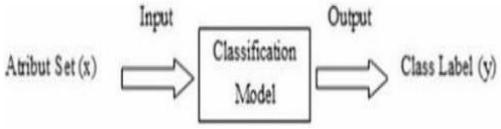
Klasifikasi menggambarkan sebuah topik pembelajaran yang ada di dalam data mining serta machine learning. Pengertian dari klasifikasi adalah sebuah pengelompokkan beberapa data kedalam kelas atau label tertentu yang sudah ditentukan sebelumnya. Dalam menyelesaikan kasus klasifikasi terdapat algoritmaalgoritma yang berjalan, algoritma ini termasuk dalam supervised learning (Nurdiansyah, Cholissodin, & Adikara, 2020).

Dalam klasifikasi, terdapat target variabel kategori. Sebagai contoh, penggolongan pendapatan dapat dipisahkan dalam tiga kategori, yaitu pendapatan tinggi, pendapatan sedang, dan pendapatan rendah. Contoh lain klasifikasi dalam bisnis dan penelitian adalah:

a. Menentukan apakah suatu transaksi kartu kredit merupakan transaksi yang curang atau bukan.

b. Mendiagnosis penyakit seorang pasien untuk mendapatkan termasuk penyakit apa.

Klasifikasi merupakan proses pembelajaran suatu fungsi tujuan (target) f yang memetakan tiap himpunan atribut x ke satu dari label kelas y yang didefinisikan sebelumya. Fungsi target disebut juga model klasifikasi.



Gambar 1. Diagram Model Klasifikasi

Ada dua jenis model Klasifikasi, yaitu :

a. Pemodelan Deskriptif (*descriptive modelling*) : Model klasifikasi yang dapat berfungsi sebagai suatu alat penjelasan untuk membedakan objekobjek dalam kelas-kelas yang berbeda.

b. Pemodelan Prediktif (*predictive modelling*) : Model klasifikasi yang dapat digunakan untuk memprediksi label kelas record yang tidak diketahui.

2.3 Algortima *Naïve Bayes*

Naive Bayes Classifier merupakan sebuah metode klasifikasi yang berakar pada teorema Bayes. Metode pengklasifikasian dengan menggunakan metode probabilitas dan statistik. *Naïve Bayes Classifier* memprediksi peluang di masa depan berdasarkan pengalaman di masa sebelumnya (Asfi & Fitrianingsih, 2020).

Naive Bayes Classifier menggunakan asumsi yang sangat kuat (naif) akan independensi dari masing-masing kondisi atau kejadian, dimana masing-masing petunjuk saling bebas (independen) satu sama lain. Dengan asumsi tersebut, maka berlaku suatu persamaan sebagai berikut:

$$P(Y|X) = \frac{P(X|Y)\,P(Y)}{P(X)}$$

Keterangan :

X = Data testing yang kelasnya belum diketahui.

Y = Hipotesis.

P(Y|X) = Probabilitas prior dari hipotesis Y, yaitu probabilitas bersyarat dari hipotesis Y berdasarkan kondisi X.

P(Y) = Probabilitas prior dari hipotesis Y, yaitu probabilitas bahwa hipotesis Y bernilai benar sebelum data X muncul.

P(X) = Probabilitas dari data X.

P(X|Y) = Probabilitas bersyarat dari X bedasarkan kondisi pada hipotesis Y, dan biasa disebut likehood.

Biner : Jurnal Ilmu Komputer , Teknik dan Multimedia
Volume 1, No. 3, Agustus 2023
ISSN 2988-3814 (media online)
Hal 745-759

Naïve Bayes Classifier dalam proses klasifikasi data memerlukan banyak petunjuk berupa atribut untuk mendapatkan kesimpulan berupa label kelas yang sesuai untuk sampel data (Irmayani, 2021). Oleh karena itu, Teorema Bayes disesuaikan sebagai berikut:

$$P(Y|X) = \frac{P(X1|Y)P(X2|Y) \dots P(Xn|Y)P(Y)}{P(X)}$$

Keterangan :

X = Himpunan data training

Y = Hipotesis

(Y|X) = Probabilitas prior dari hipotesis Y, yaitu probabilitas bersyarat dari hipotesis Y berdasarkan kondisi X

(Y) = Probabilitas prior dari hipotesis Y, yaitu probabilitas bahwa hipotesis Y bernilai benar sebelum data X muncul.

(X) = probabilitas dari data X.

P(X1 | Y)P(X2 | Y)..P(Xn | Y)P(Y) = Probabilitas dari X1, X2, Xn untuk hipotesis Y, biasa disebut dengan likehood. Karena P(X) irrelevant, maka untuk mencari peluang hanya menggunakan rumus berikut ini:

$$P(Y|X) = P(X1|Y)P(X2|Y) \dots P(Xn|Y)P(Y)$$

Jika nilai P(Xn|Y) adalah 0, maka nilai P(Y|X) = 0. Maka klasifikasi *Naïve Bayesian* tidak bisa dilakukan, karena klasifikasi *Naïve Bayesian* tidak bisa memprediksi record yang salah satu atributnya memiliki probabilitas bersyarat (likehood) = 0. Untuk mengatasi hal tersebut, dilakukan penambahan nilai 1 ke setiap evidence / P(X) dalam perhitungan sehingga probabilitas tidak akan bernilai 0.

2.4 *Laplace Correction*

Laplace correction atau *laplace smoothing* merupakan salah satu metode atau algoritma pemulusan (*smoothing*) tertua yang digunakan. *Laplacian smoothing* merupakan metode yang berpengaruh untuk mencegah masalah probabilitas nol (Aprianto, Prasetyo, & Sahartian, 2021). Metode *laplacian smoothing* juga dikenal *sebagai add-one smoothing yaitu* menambahkan angka 1 pada setiap frekuensi token yang didapat (Bentuk perhitungan dengan *laplacian smoothing* dapat dilihat pada rumus sebegai berikut:

$$P(Wi|class) = \frac{freq(Wi, class) + 1}{N\ class + V\ class}$$

Laplacian Smoothing menambahkan nilai 1 pada setiap frekuensi kelas dan menambahkan V class yaitu jumlah atribut pada kelas tertentu untuk menghasilkan probabilitas tanpa hasil nol. Pada penelitian ini, laplacian smoothing digunakan untuk menghindari nilai nol pada probabilitas yang dihasilkan oleh metode naïve bayes. Nilai probabilitas nol akan berpengaruh terhadap hasil klasifikasi karena mengkalkulasi beberapa probabilitas untuk mendapatkan nilai akhir yang menentukan nilai probabilitas.

3. ANALISA DAN PEMBAHASAN

3.1 Akuisisi Pengetahuan

Setelah penyusunan basis pengetahuan dengan tabel keputusan diagnosa penyakit sesuai pengamatan pada penyakit Tuberkulosis. Hanya beberapa gejala yang paling nampak digunakan dalam tabel keputusan diagnosa, kemudian ditentukan hasil diagnosanya. Represetansi pengetahuan dibuat dalam bentuk tabel yang akan digunakan dalam pembuatan aturan-aturan untuk melakukan pengambilan keputusan diagnosa pada penyakit tuberkulosis.

Biner : Jurnal Ilmu Komputer , Teknik dan Multimedia
Volume 1, No. 3, Agustus 2023
ISSN 2988-3814 (media online)
Hal 745-759

Tabel 1. Tabel Keputusan Diagnosa

Penyakit	Kode Gejala	Gejala Penyakit
Tuberculosis	G1	Batuk berdahak selama > 2-3 minggu
	G2	Batuk berdarah
	G3	Demam hilang timbul > 1 bulan
	G4	keringat malam tanpa aktifitas
	G5	Penurunan berat badan tanpa penyebab yang jelas
	G6	Sesak nafas dan nyeri dada

3.2 Akuisisi Pengetahuan

Setelah penyusunan basis pengetahuan dengan tabel keputusan diagnosa penyakit sesuai pengamatan pada penyakit Tuberkulosis. Hanya beberapa gejala yang paling nampak digunakan dalam tabel keputusan diagnosa, kemudian ditentukan hasil diagnosanya. Represetansi pengetahuan dibuat dalam bentuk tabel yang akan digunakan dalam pembuatan aturan-aturan untuk melakukan pengambilan keputusan diagnosa pada penyakit tuberkulosis.

Tabel 2. Gejala Penyakit Tuberkulosis

Penyakit	Kode Gejala	Gejala Penyakit
Tuberculosis	G1	Batuk berdahak selama > 2-3 minggu
	G2	Batuk berdarah
	G3	Demam hilang timbul > 1 bulan
	G4	keringat malam tanpa aktifitas
	G5	Penurunan berat badan tanpa penyebab yang jelas
	G6	Sesak nafas dan nyeri dada

3.3 Metode Algoritma *Naïve Bayes*

Naïve Bayes merupakan sebuah pengklasifikasian probalistik sederhana yang menghitung sekumpulan probabilitas dengan menjumlahkan frekuensi dan kombinasi nilai dari data set yang diberikan. Algoritma menggunakan Teorema *Bayes* dan mengasumsikan semua atribut independen atau tidak saling ketergantungan yang diberikan oleh nilai pada variable kelas. *Naïve Bayes* juga didefinisikan sebagai pengklasifikasian dengan metode probabilitas dan statistic yang dikemukakan oleh Thomas Bayes yaitu memprediksi peluang di masa depan berdasarkan pengalaman di masa sebelumnya.

Naive Bayes didasarkan pada asumsi penyederhanaan bahwa nilai atribut secara kondisional saling bebas jika diberikan nilai output. Dengan kata lain, diberikan nilai output, probabilitas mengamati secara bersama adalah produk dari probabilitas individu. Keuntungan penggunaan *Naive Bayes* adalah bahwa metode ini hanya membutuhkan jumlah data pelatihan (*Training Data*) yang kecil untuk menentukan estimasi paremeter yang diperlukan dalam proses pengklasifikasian. *Naive Bayes* sering bekerja jauh lebih baik dalam kebanyakan situasi dunia nyata yang kompleks dari pada yang diharapkan (Rifai, 2019).

Biner : Jurnal Ilmu Komputer , Teknik dan Multimedia
Volume 1, No. 3, Agustus 2023
ISSN 2988-3814 (media online)
Hal 745-759

3.4 Data

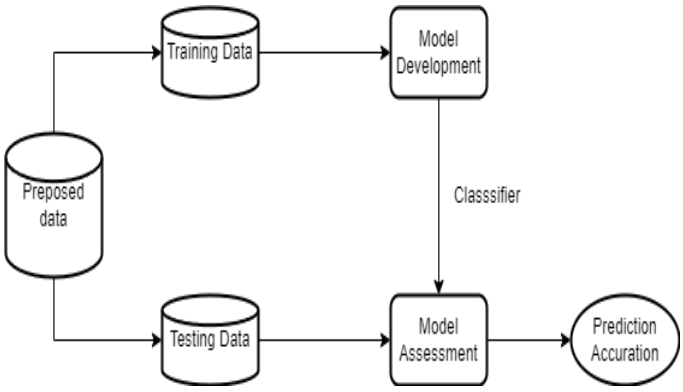
Peneliti mendapatkan data dari pihak RS Sari Asih Ciputat berupa *hard copy* file rekam medis pasien. Data rekam medis tersebut kemudian dianalisis guna mendapatkan data yang spesifik dan menuangkan data yang didapatkan dalam bentuk excel, guna mempermudah pengolahan data. Total data yang diambil sebanyak 100 kasus, pasien terdiagnosa positif tuberculosis sebanyak 70 dan pasien terdiagnosa negatife tuberkulosis sebanyak 30 orang.

Tabel 3. Contoh Data Pasien Tuberkulosis (TB)

		Batuk berdahak selama > 2-3	Batuk berdarah	Demam hilang timbul > 1 bulan	keringat malam tanpa aktifitas	Penurunan berat badan tanpa penyebab	Sesak nafas dan nyeri dada	
NO	ID Kode MR	G1	G2	G3	G4	G5	G6	HASIL
1	01.68.90	ya	tidak	tidak	tidak	tidak	tidak	NEGATIF
2	30.45.98	ya	tidak	tidak	tidak	tidak	tidak	NEGATIF
3	33.89.48	ya	tidak	tidak	tidak	tidak	tidak	NEGATIF
4	33.77.22	ya	tidak	tidak	tidak	tidak	tidak	NEGATIF
5	12.05.79	ya	tidak	tidak	tidak	tidak	tidak	NEGATIF
6	05.43.55	tidak	tidak	ya	tidak	tidak	ya	NEGATIF
7	33.88.68	ya	tidak	tidak	tidak	tidak	tidak	NEGATIF
8	29.21.63	ya	tidak	tidak	tidak	ya	ya	POSITIF
9	33.87.69	ya	tidak	tidak	tidak	tidak	ya	POSITIF
10	32.33.49	ya	tidak	tidak	tidak	ya	tidak	POSITIF
11	33.49.07	tidak	ya	tidak	tidak	tidak	tidak	POSITIF
12	23.89.58	ya	tidak	tidak	tidak	ya	tidak	POSITIF
13	10.46.90	ya	tidak	ya	tidak	ya	ya	POSITIF
14	09.54.41	ya	tidak	tidak	tidak	tidak	ya	POSITIF
15	33.93.81	ya	tidak	ya	tidak	ya	ya	POSITIF
16	30.98.34	ya	tidak	tidak	tidak	ya	tidak	POSITIF
17	33.32.66	ya	tidak	ya	tidak	ya	ya	POSITIF
18	32.28.52	ya	tidak	tidak	tidak	ya	tidak	POSITIF
19	32.47.49	ya	tidak	tidak	tidak	ya	ya	POSITIF

3.5 Model Yang Diusulkan

Model yang diusulkan untuk klasifikasi menggunakan algoritma *Naïve bayes* adalah menggunakan model *split validation. Split validation* membagi data menjadi dua subset data yaitu data trainning dan data testing. Data trainning merupakan data yang digunakan untuk pelatihan, sedangkan data testing akan digunakan untuk pengujian. Adapaun untuk melihat secara lebih jelas dari model split validation dapat dilihat pada gambar berikut:



Gambar 2. Model Yang Diusulkan

Biner : Jurnal Ilmu Komputer , Teknik dan Multimedia
Volume 1, No. 3, Agustus 2023
ISSN 2988-3814 (media online)
Hal 745-759

Pada gambar 3.2 akan digunakan untuk melakukan pengujian dengan masing-masing proporsi pembagian datanya dapat dilihat pada table berikut:

Tabel 4. Pembagian Data

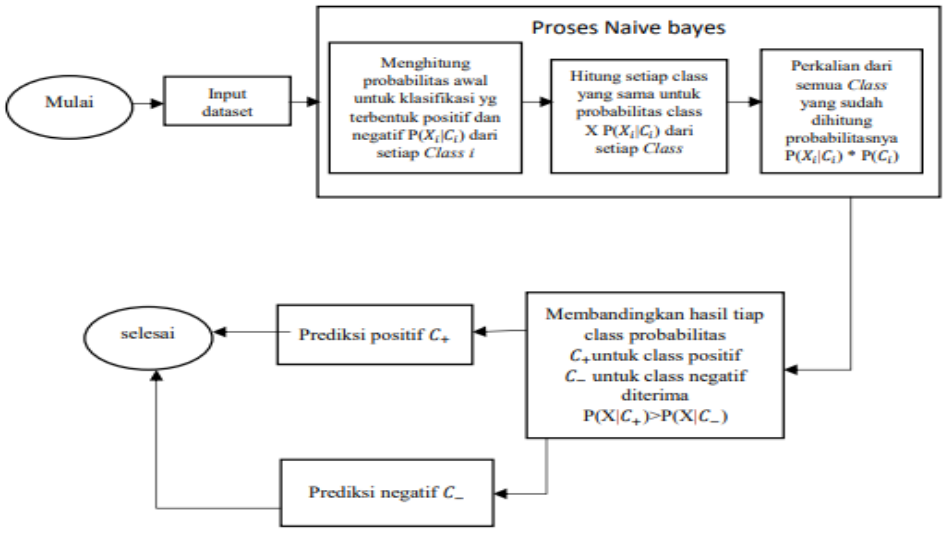
Training	Testing
60%	40%
70%	30%
80%	20%
90%	10%

Dari empat kali percobaan yang dilakukan bedasarkan dari tabel 3.6.3 setiap hasil yang diperoleh akan ditentukan jumlahnya.

4. HASIL DAN PEMBAHASAN

4.1 Lagkah Perhitungan

Pada tahap ini metode yang digunakan dalam perhitungan tingkat akurasi adalah algoritma naive bayes dengan melakukan pengujian akurasi data set dan perhitungan manual. Berikut langkah metode algoritma naive bayes:



Gambar 3. Model Algoritma *Naïve Bayes*

Pada gambar 3 mulai identifikasi sampel dari data set baca data. selanjutnya P(Xi|Ci) menghitung jumlah class dari klasifikasi yang sudah terbentuk yaitu class positif dan negatif untuk setiap class. Kemudian P(X|Ci) menghitung jumlah kasus yang sama dari kelas yang sama X, dalam kasus data set pada penelitian ini terdiri dari 2 class yaitu positif yang dinyatakan dengan simbul (+) dan negatif yang dinyatakan dengan simbul (-). Kemudian hitung P(X|C+),i=+,- untuk setiap kelas atau atribut. Setelah itu dibandingkan, jika P(X|C+)>P(X|$C-$) maka kesimpulannya adalah C+ atau pada penelitian ini berarti diagnosa penyakit positif. Jika P(X|C+).

Biner : Jurnal Ilmu Komputer , Teknik dan Multimedia
Volume 1, No. 3, Agustus 2023
ISSN 2988-3814 (media online)
Hal 745-759

4.2 Hasil Pengujian Prediksi Diagnosa

Tabel 5. Contoh Potongan Data Training

NO	ID Kode MR	G1	G2	G3	G4	G5	G6	HASIL
1	01.68.90	ya	tidak	tidak	tidak	tidak	tidak	NEGATIF
2	30.45.98	ya	tidak	tidak	tidak	tidak	tidak	NEGATIF
3	33.89.48	ya	tidak	tidak	tidak	tidak	tidak	NEGATIF
4	33.77.22	ya	tidak	tidak	tidak	tidak	tidak	NEGATIF
5	12.05.79	ya	tidak	tidak	tidak	tidak	tidak	NEGATIF
6	05.43.55	tidak	tidak	ya	tidak	tidak	ya	NEGATIF
7	33.88.68	ya	tidak	tidak	tidak	tidak	tidak	NEGATIF
8	29.21.63	ya	tidak	tidak	tidak	ya	ya	POSITIF
9	33.87.69	ya	tidak	tidak	tidak	tidak	ya	POSITIF
10	32.33.49	ya	tidak	tidak	tidak	ya	tidak	POSITIF
11	33.49.07	tidak	ya	tidak	tidak	tidak	tidak	POSITIF
12	23.89.58	ya	tidak	tidak	tidak	ya	tidak	POSITIF
13	10.46.90	ya	tidak	ya	tidak	ya	ya	POSITIF
14	09.54.41	ya	tidak	tidak	tidak	tidak	ya	POSITIF
15	33.93.81	ya	tidak	ya	tidak	ya	ya	POSITIF
16	30.98.34	ya	tidak	tidak	tidak	ya	tidak	POSITIF
17	33.32.66	ya	tidak	ya	tidak	ya	ya	POSITIF
18	32.28.52	ya	tidak	tidak	tidak	ya	tidak	POSITIF
19	32.47.49	ya	tidak	tidak	tidak	ya	ya	POSITIF
20	32.06.97	ya	tidak	tidak	tidak	tidak	ya	POSITIF
21	31.50.84	ya	tidak	tidak	tidak	tidak	ya	POSITIF
22	31.26.93	ya	tidak	ya	tidak	ya	ya	POSITIF
23	18.61.10	ya	tidak	tidak	tidak	tidak	ya	POSITIF
24	31.11.80	ya	tidak	ya	tidak	ya	ya	POSITIF

Data *training* adalah data yang akan di latih untuk menentukan hasil dari data *testing*.

Tabel 6. Contoh Potongan Data Testing

no.mr	g1	g2	g3	g4	g5	g6	hasil
38.71.87	ya	tidak	ya	tidak	ya	ya	positif
30.94.11	ya	tidak	tidak	tidak	tidak	ya	positif
08.23.05	ya	ya	ya	tidak	tidak	tidak	positif
35.11.39	ya	tidak	tidak	tidak	ya	tidak	positif
36.36.53	ya	tidak	tidak	tidak	tidak	ya	positif
17.94.26	ya	tidak	ya	tidak	ya	ya	positif
36.26.01	ya	tidak	ya	tidak	tidak	ya	positif
36.26.40	ya	tidak	ya	ya	tidak	ya	positif
00.40.47	ya	tidak	tidak	tidak	tidak	ya	positif
35.76.54	ya	tidak	ya	tidak	ya	ya	positif
02.11.31	tidak	ya	ya	tidak	tidak	ya	positif
02.32.12	ya	ya	ya	tidak	tidak	tidak	positif
01.03.22	ya	tidak	tidak	tidak	ya	tidak	negatif
01.40.22	ya	tidak	tidak	tidak	tidak	tidak	negatif
01.50.23	ya	tidak	tidak	tidak	tidak	tidak	negatif
32.13.45	ya	ya	ya	tidak	ya	tidak	positif

Tabel berisi tentang data diagnosa dari rumah sakit yang kemudian akan di testing menggunakan hitung manual dan rapid miner.

4.3 Prediksi Menggunakan Perhitungan Manual

Berikut ini perhitungan dalam penelitian ini menggunakan 100 data training terdiri dari 6 atribut untuk menentukan sebuah class, yang mana dari 100 data training tersebut akan digunakan untuk melakukan perhitungan algoritma *Naive bayes* dan *Laplacian smooting* untuk mengcegah

Biner : Jurnal Ilmu Komputer , Teknik dan Multimedia
Volume 1, No. 3, Agustus 2023
ISSN 2988-3814 (media online)
Hal 745-759

masalah probabilitas nol. Adapun contoh potongan data training tersebut dapat dilihat pada tabel 4.2 :

Data testing 1 : X = (No.mr=”32.13.45”, g1=”ya”, g2= “ya”, g3=”ya”, g4= “tidak”, g5= “ya”, g6=”tidak”)

Tahap 1 menghitung jumlah kelas atau prediksi data testing

$P(Ci)$

$P(Positif) = \frac{73}{100} = 0{,}73$

$P(Negatif) = \frac{27}{100} = 0{,}27$

Tahap 2 menghitung jumlah kasus yang sama dengan kelas yang sama.

$P(X \mid Ci)$

$(g1)\ Batuk\ berdahak\ 23\ minggu$

$(Ya|Positif) = \frac{70}{79} = 0{,}88607$

$(g1)\ Batuk\ berdahak > 2-3\ minggu\ (Ya|Negatif) = \frac{20}{33} = 0{,}60606$

$(g2)\ Batuk\ berdarah\ (Ya|Positif) = \frac{12}{79} = 0{,}15189$

$(g2)\ Batuk\ berdarah\ (Ya|Negatif) = \frac{1}{33} = 0{,}03030$

$(g3)\ Demam > 1\ bulan\ (Ya|Positif) = \frac{30}{79} = 0{,}37974$

$(g3)\ Demam > 1\ bulan\ (Ya|Negatif) = \frac{9}{33} = 0{,}27272$

$(g4)\ Keringat\ malam\ tanpa\ aktifitas$

$(Tidak|Positif) = \frac{61}{79} = 0{,}77215$

$(g4)\ Keringat\ malam\ tanpa\ aktifitas$

$(Tidak|Negatif) = \frac{28}{33} = 0{,}84848$

$(g5)\ Penurunan\ berat\ badan\ (Ya|Positif) = \frac{50}{79} = 0{,}63291$

$(g5)\ Penurunan\ berat\ badan\ (Ya|Negatif) = \frac{2}{33} = 0{,}06060$

$(g6)\ Sesak\ nafas\ (Tidak|Positif) = \frac{21}{79} = 0{,}26582$

$(g6)\ Sesak\ nafas\ (Tidak|Negatif) = \frac{20}{33} = 0{,}60606$

Biner : Jurnal Ilmu Komputer , Teknik dan Multimedia
Volume 1, No. 3, Agustus 2023
ISSN 2988-3814 (media online)
Hal 745-759

Tahap 3 mengkalikan semua hasil dari atribut Positif dan Negatif.

$$P(X \mid Positif) = 0{,}88607 * 0{,}15189 * 0{,}37974 * 0{,}77215 * 0{,}63291 * 0{,}26582 = 0{,}0066391861$$

$$P(X \mid Negatif) = 0{,}60606 * 0{,}03030 * 0{,}27272 * 0{,}84848 * 0{,}06060 * 0{,}60606 = 0{,}0001560648$$

Tahap 4 membandingkan nilai kelas Positif dan Negatif.

$$P(X \mid Ci) * P(Ci)$$

$$P(X \mid Positif) * P(Positif) = 0{,}006639 * 0{,}73 = 0{,}00484647$$

$$P(X \mid Negatif) * P(Negatif) = 0{,}000156 * 0{,}27 = 0{,}00004212$$

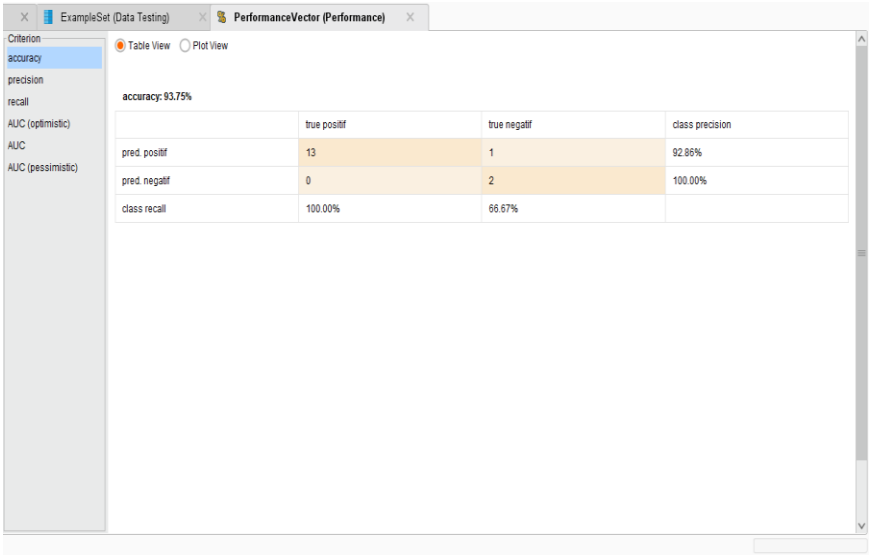
Jadi untuk (No. MR=”32.13.45”, g1=”ya”, g2= “ya”, g3=”ya”, g4= “tidak”, g5= “ya”, g6=”tidak”), hasilnya “**Positif**” Tuberculosis.

4.4 Hasil Performance

Proses klasifikasi dengan rapidminer dengan metode *naive bayes* yang digunakan mengklasifikasi data pasien sebanyak 100 data pada penelitian ini sehingga diperoleh nilai *Accuracy*, *Precision* dan *Recall* dengan menggunakan data testing sebanyak 16 data.

1. *Accuracy* / akurasi

   Dengan mengetahui jumlah data yang di klasifikasikan secara benar maka dapat diketahui akurasi hasil prediksi yaitu 93,75% dari hasil data pasien. Di bawah ini merupakan hasil dari testing menggunakan rapidminer 9.1.



Gambar 4. *Accuracy*

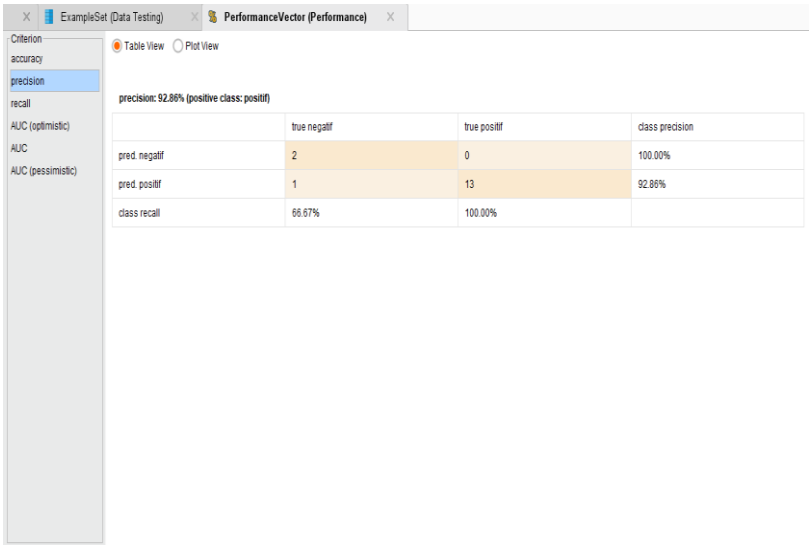
Biner : Jurnal Ilmu Komputer , Teknik dan Multimedia
Volume 1, No. 3, Agustus 2023
ISSN 2988-3814 (media online)
Hal 745-759

Berikut ini perhitungan manual untuk menentukan *accuracy*.

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)} x\ 100\%$$
$$= \frac{(13+2)}{(13+2+1+0)} x\ 100\% = \frac{15}{16} x100\%$$
$$= 0{,}9375\ x\ 100\% = 93.75\%$$

2. *Precision*

Precision adalah jumlah data yang true positive (jumlah data positif yang dikenali secara benar sebagai positif) dibagi dengan jumlah data dikenali sebagai positif. Dari hasil pengujian nilai precision yaitu 92,86% untuk class Positif dan nilai precision 100,00% untuk class Negatif.



Gambar 5. *Precision*

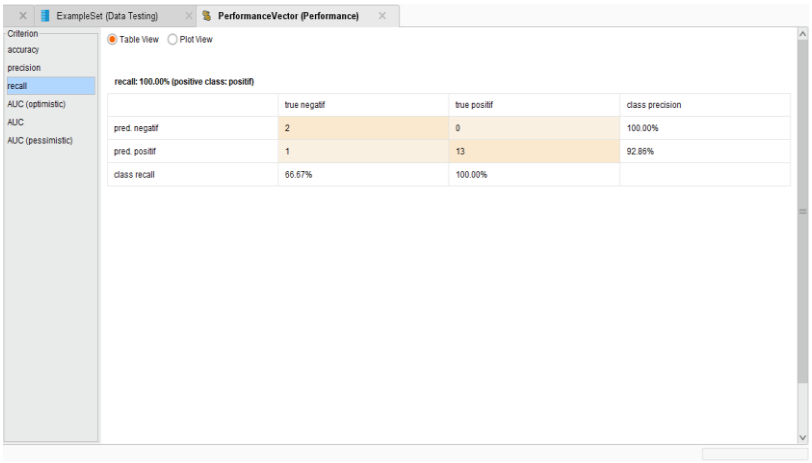
Berikut ini perhitungan manual untuk menentukan Precision.

$$Precision\ positive = \frac{TP}{(TP+FP)} = x\ 100\% = \frac{13}{(13+1)} x\ 100\% = 0{,}9285\ x\ 100\% = 92{,}85\%$$
$$Precision\ positive = \frac{TN}{(TN+FN)} = x\ 100\% = \frac{2}{(2+0)} x\ 100\% = 1\ x\ 100\% = 100{,}00\%$$

3. *Recall*

Recall merupakan jumlah data yang *true positive* dibagi dengan jumlah data yang sebenarnya positive (*true positive* + *true negative*). Untuk nilai recall yaitu 100,00% pada class positif dan nilai recall 66,67% pada class negatif.

Biner : Jurnal Ilmu Komputer , Teknik dan Multimedia
Volume 1, No. 3, Agustus 2023
ISSN 2988-3814 (media online)
Hal 745-759



Gambar 6. *Recall*

Berikut ini perhitungan manual untuk menentukan Recall positive dan Recall negative.

$$Recall\ (Positive) = \frac{TP}{(TP + FN)} x\ 100\% = \frac{13}{(13 + 0)} x\ 100\% = 1\ x\ 100\%\ =\ 100{,}00\%$$

$$Recall\ (Negatif) = \frac{TN}{(TN+FP)} x\ 100\% = \frac{2}{(2+1)} x\ 100\% = 0{,}6667 x\ 100\% = 66{,}67\%$$

4. AUC (*Area Under Curve*)

Kurva *Receiver Operating Characteristic* (ROC) juga dihasilkan oleh *Rapidminer*. Kurva tersebut dapat dilihat pada gambar berikut:

Gambar 7. AUC (*Area Under Curve*)

Kurva ROC digunakan untuk mengekspresikan data confusion matrix. Dari gambar dapat diketahui bahwa nilai Area Under Curve (AUC) model algoritma naive bayes adalah 0.885. Hal ini menunjukkan bahwa model algoritma naive bayes mencapai klasifikasi hampir sempurna.

5. Pada gambar dibawah ini adalah hasil *confusion matrix* dari rapidminer.

accuracy: 93.75%

	true positif	true negatif	class precision
pred. positif	13	1	92.86%
pred. negatif	0	2	100.00%
class recall	100.00%	66.67%	

Gambar 8. *Confusion Matrix*

Berdasarkan hasil Confusion Matrix dari rapidminer didapat accuracy 93.75%.

Biner : Jurnal Ilmu Komputer , Teknik dan Multimedia
Volume 1, No. 3, Agustus 2023
ISSN 2988-3814 (media online)
Hal 745-759

4.5 Evaluasi dan Validasi

Tahapan evaluasi yang dilakukan dalam penelitian ini adalah untuk memberikan penilaian dari hasil penggunaan algoritma *naive bayes* saja dan *naive bayes* yang disertai dengan *confusion matrix* untuk mengklasifikasi diagnosa prediksi penyakit tuberculosis menggunakan *Split validation*. Bagian yang akan dievaluasi adalah presentase data, jumlah data training, jumlah data testing dan nilai akurasi yang di hasilkan. Adapaun secara keseluruhan dapat dilihat pada tabel berikut:

Tabel 6. Hasil Akurasi Dari Perbandingan Data Training Data Testing

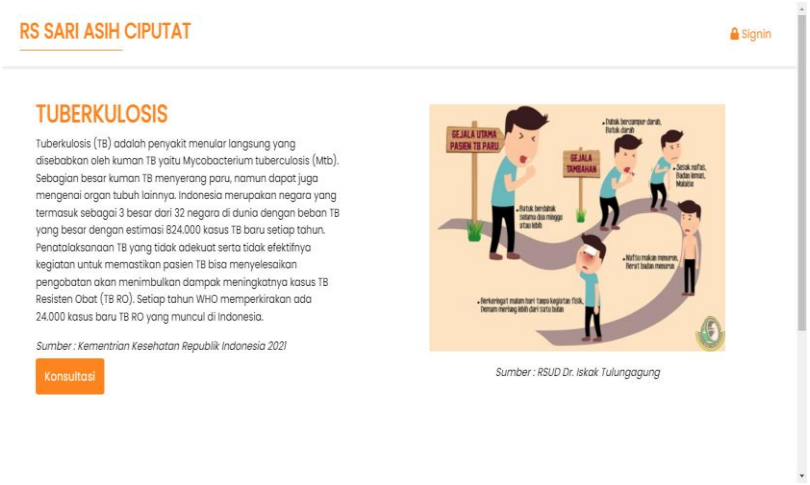
No.	Presentase Data	Data Training	Data Testing	Akurasi
1	60:40	60	40	85.00%
2	70:30	70	30	96.67%
3	80:20	80	20	95.00%
4	90:10	90	10	100.00%

Pada Tabel berisi tentang hasil akurasi dari perbandingan data training data testing sebagai berikut:

1. Presentase data dengan perbandingan data 60 : 40 yaitu data training sebanyak 60 data dan data testing sebesar 40 data menghasilkan akurasi 85.00%.
2. Presentase data dengan perbandingan data 70 : 30 yaitu data training sebanyak 70 data dan data testing sebesar 30 data menghasilkan akurasi 96.67%.
3. Presentase data dengan perbandingan data 80 : 20 yaitu data training sebanyak 80 data dan data testing sebesar 20 data menghasilkan akurasi 95,00%.
4. Presentase data dengan perbandingan data 90 : 10 yaitu data training sebanyak 90 data dan data testing sebesar 10 data menghasilkan akurasi 100,00%

4.6 Form Halaman Utama

Form Halaman utama adalah form utama yang terdiri dari form konsultasi dan form login. Gambar form home sebagai berikut:



Gambar 9. Halaman Utama

Keterangan gambar 9:

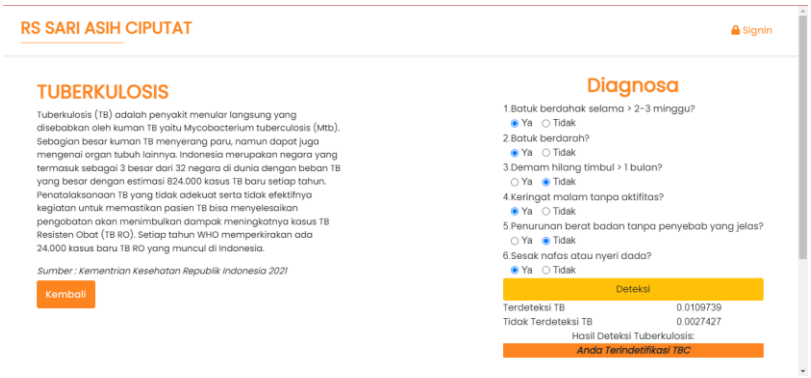
1. *Navbar* terdapat Nama RS Sari Asih Ciputat.
2. Dan di bawah nya terdapat infromasi terkait penyakit tuberkulosis.

Biner : Jurnal Ilmu Komputer , Teknik dan Multimedia
Volume 1, No. 3, Agustus 2023
ISSN 2988-3814 (media online)
Hal 745-759

3. Pada form halaman utama terdapat form konsultasi dan *signin.*
4. *Content* menampilkan informasi yang diberikan di halaman tersebut.
5. *Footer* berisi informasi pengetahuan metode naïve bayes.

4.7 Form Konsultasi

Form konsultasi adalah form yang digunakan pasien untuk mengisi gejala-gejala penyakit dan menghitung nilai probabilitas penyakit tuberkulosis. Form konsultasi yang terdiri dari form halaman utama dan *signin*. Gambar form konsultasi sebagai berikut:



Gambar 10. Form Diagnosa

Keterangan gambar 10:

1. *Navvar* terdapat nama RS Sari Asih Ciputat dan *signin*.
2. Dan di bawah nya terdapat infromasi terkait penyakit tuberkulosis
3. Pada form konsultasi terdapat pertanyaan untuk mendiagnosa penyakit tuberkulosis.
4. *Content* menampilkan hasil perhitungan diagnosa.
5. *Footer* berisi informasi pengetahuan metode naïve bayes.

5. KESIMPULAN

Berdasarkan dari penulisan yang telah diuraikan, maka dapat dibuat kesimpulan sebagai berikut:

1. Implementasi data mining dengan metode klasifikasi dan algoritma naive bayes dapat memprediksi diagnosa penyakit tuberculosis dengan lebih cepat dan mudah.
2. Implementasi data mining dengan metode klasifikasi dan algoritma naive bayes dapat memprediksi diagnosa penyakit tuberculosis dengan cukup akurat yaitu dengan akurasi 85,00%.

REFERENCES

Amalia, R. (2020). Penerapan Data Mining untuk Memprediksi Hasil Kelulusan Siswa Menggunakan Metode Naive Bayes. *Jurnal Sistem Informasi Fakultas Ilmu Komputer Universitas Darwan Ali*, Vol. 06, No. 01, Hal 33-34.

Apriyanto. (2021). SiDedi (Sistem Informasi Deteksi Diabetes) : Sistem Pendukung Keputusan Deteksi Dini Diabetes. *Jsika Jurnal*, Hal 3.

Asfi, M. (2020). Implementasi Algoritma Naive Bayes Classifier sebagai Sistem Rekomendasi Pembimbing Skripsi. *InfoTekJar : Jurnal Nasional Informatika Dan Teknologi Jaringan*, Vol, 5. No, 1. Hal 45.

Indonesia, K. K. R. (2021). *Perubahan Alur Diagnosis dan Pengobatan Tuberkulosis di Indonesia*.

Irmayani, W. (2021). Visualisasi Data Pada Data Mining Menggunakan Metode Klasifikasi Naive Bayes. *Jurnal Khatulistiwa Informatika*, Hal 69.

Isbaniyah. (2021). Tuberkulosis: Pedoman Diagnosis dan Penatalaksanaan di Indonesia (Perhimpunan Dokter Paru Indonesia). In *Tuberkulosis: Pedoman Diagnosis dan Penatalaksanaan di Indonesia (Perhimpunan Dokter Paru Indonesia)*. Perhimpunan Dokter Paru Indonesia.

Nurdiansyah, & Vidia, V. (n.d.). Klasifikasi Penyakit Tuberkulosis (TB) menggunakan Metode Extreme Learning Machine (ELM). *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*, 2020.

Rifai, M. F. (2019). Penerapan Algoritma Naïve Bayes Pada Sistem Prediksi Tingkat Kelulusan Peserta Sertifikasi Microsoft Office Specialist (MOS). *Jurnal Pengkajian Dan Penerapan Teknik Informatika*, Hal 133.

Saputro, I. W. (2019). Uji Performa Algoritma Naïve Bayes untuk Prediksi Masa Studi Mahasiswa. *Citec Journal*, Hal 5.

Suntoro, J. (2019). *Data Mining: Algoritma dan Implementasi dengan Pemrograman PHP*. PT. Elex Media Komputindo.