JOURNAL OF INFORMATICS AND COMPUTING (RANDOM)
Vol. 3, No. 1 (Tahun 2024), Hal. 41 - 48

PERBANDINGAN ALGORITMA *NAÏVE BAYES* DAN C4.5 PADA KLASIFIKASI PENYAKIT *TUBERCULOSIS*

Putri Nur Isnaeni [1], Hidayatur Rakhmawati [2], Fitri Ayuning Tyas [3]

[1,2,3]STMIK MUHAMMADIYAH PAGUYANGAN BREBES
Putrinurg11ti@gmail.com

INFORMASI ARTIKEL

Diajukan:
25 September 2023
Direvisi:
13 Januari 2024
Diterima:
14 April 2024

Kata kunci:

Tuberculosis
Naïve Bayes
C4.5
Perbandingan

Abstrak

Bakteri *Mycobacterium Tuberculosis* menyebabkan *tuberculosis*. Penyakit ini biasanya menyebar melalui percikan cairan seperti batuk atau bersin. *Tuberculosis* Paru dan *Tuberculosis* Ekstra Paru adalah dua klasifikasi *tuberculosis*. Indonesia adalah Negara ketiga dengan jumlah kasus *tuberculosis* tertinggi. Oleh karena itu, perlu dilakukan upaya untuk mengurangi jumlah kasus *tuberculosis* di Indonesia. Salah satu upaya yang dapat dilakukan adalah dengan mengklasifikasikan data *tuberculosis* pada medis. Ini dapat membantu dokter dalam diagnosis penyakit dan mendeteksi dini *tuberculosis*, yang memungkinkan penyembuhan lebih cepat bagi mereka yang menderita penyakit ini. Salah satu jenis algoritma klasifikasi adalah *Naive Bayes* dan C4.5. Yang pertama digunakan untuk memprediksi peluang di masa depan berdasarkan pengalaman sebelumnya dengan mencari peluang terbesar dari beberapa kemungkinan klasifikasi dan melihat frekuensi tiap klasifikasi pada data latih. C4.5 dipilih karena merupakan tolak ukur yang sering digunakan dalam perbandingan dengan algoritma *Naive Bayes*. Hasil dari penelitian ini adalah perbandingan algoritma *Naïve Bayes* dan C4.5 menggunakan model *10-fold cross validation* dan model pengukuran akurasi *confusion matrix*. Hasil yang didapat algoritma C4.5 memiliki tingkat akurasi lebih tinggi dengan presentase sebanyak 85% dari algoritma *Naïve Bayes* yang memiliki akurasi 83% yang berarti algoritma C4.5 lebih baik saat diterapkan pada proses klasifikasi penyakit *tuberculosis*.

COMPARISON OF NAÏVE BAYES AND C4.5 ALGORITHMS IN THE CLASSIFICATION OF TUBERCULOSIS

ARTICLE INFORMATION

Submitted:
25 September 2023
Received:
13 January 2024
Accepted:
14 April 2024

Keywords:

Tuberculosis
Naïve Bayes
C4.5
Comparison

Abtract

Mycobacterium Tuberculosis bacteria cause tuberculosis. This disease is usually spread through splashes such as coughing or sneezing. Pulmonary Tuberculosis and Extra Pulmonary Tuberculosis are two classifications of tuberculosis. Indonesia is the third country with the highest number of tuberculosis cases. Therefore, efforts need to be made to reduce the number of tuberculosis cases in Indonesia. One effort that can be made is to classify tuberculosis data in medical settings. This can help doctors in diagnosing the disease and early detection of tuberculosis, allowing faster healing for those suffering from this disease. One type of classification algorithm is Naive Bayes and C4.5. The first is used to predict future opportunities based on previous experience by looking for the largest opportunity from several possible classifications and looking at the frequency of each classification in the training data. C4.5 was chosen because it is a benchmark that is often used in comparison with the Naive Bayes

p-ISSN 2986-1942
e-ISSN 2986-1187

algorithm. The results obtained by the C4.5 algorithm have a higher level of accuracy with a percentage of 85% than the naïve bayes algorithm which has an accuracy of 83%, which means the C4.5 algoritm is better when applied to the tuberculosis disease classification process.

PENDAHULUAN

Tuberculosis adalah penyakit penyebab kematian utama di seluruh dunia menurut WHO, dan merupakan salah satu penyakit manusia tertua [1]. Menurut Kementrian Kesehatan Republik Indonesia, Indonesia adalah negara ketiga dengan pasien *tuberculosis* terbanyak, diikuti oleh India dan Cina [2]. Menurut Kementrian Kesehatan Republik Indonesia tiga provinsi pada tahun 2021 memiliki kasus *Tuberculosis* terbanyak di Indonesia, paling banyak ditemukan di Provinsi Jawa Barat dengan jumlah 91.368 kasus, kemudian di Provinsi Jawa Tengah dengan jumlah 43.121 kasus, dan ketiga Provinsi Jawa Timur dengan jumlah 42.193 kasus [3]. Salah satu Kabupaten di Jawa Tengah sebagai provinsi dengan kasus *Tuberculosis* terbanyak kedua adalah Kabupaten Banyumas.

Puskesmas Kedungbanteng yang berada di wilayah Kabupaten Banyumas adalah salah satu layanan kesehatan masyarakat yang turut serta dalam penanganan dan pengobatan *tuberculosis*. Dari hasil wawancara dengan Bapak Pramono selaku penanggung jawab penanganan penyakit *tuberculosis* di Puskesmas Kedungbanteng Kabupaten Banyumas diketahui jumlah kasus penderita *tuberculosis* paru positif di wilayah kerja Puskesmas Kedungbanteng Kabupaten Banyumas pada tahun 2022 sebanyak 656 kasus dari perkiraan 405 kasus atau sebanyak (162%). Dengan klasifikasi data penyakit *tuberculosis* pada medis dapat membantu dokter dalam mengambil keputusan diagnosis penyakit [4]. Selain itu, deteksi dini terhadap penyakit *tuberculosis* memungkinkan penderita penyakit ini dapat disembuhkan lebih cepat. Algoritma yang dapat diterapkan untuk klasifikasi antara lain algoritma *Naïve Bayes*, *Decision Tree, Logistic Regresion, Random Forest,* KNN dan ANN dan beberapa algoritma klasifikasi yang lain. Algoritma yang digunakan pada penelitian ini adalah *Naïve Bayes* dan *Decision Tree* C4.5.

LANDASAN TEORI

Naïve Bayes

Naive Bayes Classifier merupakan, pengklasifikasi probabilitas sederhana berdasarkan pada *Teorema Bayes*. *Teorema Bayes* akan dikombinasikan dengan "*Naïve*" yang artinya pada setiap atribut/variabel bersifat bebas (independent) [5]. *Naïve Bayes* didasarkan pada asumsi penyederhanaan bahwa nilai atribut secara kondisional saling bebas jika diberikan nilai *output* [6]. 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. Dalam Deshpande menyatakan formula yang digunakan pada *teorema bayes* adalah sebagai berikut [7]:

$$P(C|X) = \frac{P(X|C)P(C)}{P(X)} \quad (1)$$

Keterangan:

X	: Data dengan class yang belum diketahui.
C	: Hipotesis data E merupakan suatu class spesifik
P(C|X)	: Peluang hipotesis berdasar kondisi (*posterior probability*)
P(C)	: Peluang Hipotesis (*Prior probability*)
P(X|C)	: Peluang berdasarkan kondisi pada hipotesis
P(X)	: Probabilitas C.

C4.5

Algoritma C4.5 yaitu algoritma yang digunakan untuk membentuk sebuah pohon keputusan. Beberapa elemen pembentuk pohon keputusan yaitu root, node dan relationship [8]. Nilai gain tertinggi dari atribut digunakan sebagai penentu atribut yang akan dijadikan akar. Cara menghitung gain digunakan persamaan berikut:

$$\text{Gain}(S, A) = \text{Entropy}(S) - \sum_{i=1}^{n} \left(\frac{|Si|}{|S|} * Entropy(Si) \right) \quad (2)$$

Keterangan:
S :Himpunan Kasus |Si| :Jumlah kasus pada partisi ke-i
A :Atribut |S| :Jumlah kasus dalam S
n :Jumlah partisi atribut A
Sementara itu, untuk mencari nilai entropi digunakan persamaan berikut:

$$\text{Entropy}(S) = \sum_{i=1}^{n} -\text{pi} * \log 2\text{pi} \quad (3)$$

Keterangan:
S :Himpunan Kasus
A :Atribut
Si :Jumlah sampel untuk atribut i

$$\text{Gain Ratio (a)} = \frac{\text{gain(a)}}{\text{split(a)}} \quad (4)$$

Keterangan:
a :Atribut
Gain :Nilai gain pada atribut a
Split :Nilai split pada atribut a

Evaluasi Kinerja Klasifikasi Data Mining

1. *K-fold Cross Validation*
   K-fold cross validation merupakan sebuah metode yang digunakan untuk mengetahui rata-rata keberhasilan pengklasifikasian dengan melakukan pembagian dataset secara acak menjadi k himpunan bagian [9]. Untuk 10-*fold cross validation*, dataset dibagi menjadi 10 lipatan yang saling terpisah dengan ukuran yang hampir sama. Dalam setiap *run*, 9 subset digunakan untuk pelatihan dan sisanya untuk validasi [10].
2. *Confusion Matrix*
   Confusion Matrix merupakan salah satu cara untuk menganalisis kinerja model klasifikasi [11]. *Confusion matrix* diterapkan dengan tujuan untuk melakukan analisa tentang sebaik apa model klasifikasi yang digunakan dalam mengetahui data yang berbeda [12].

Tabel 1.*Confusion Matrix*

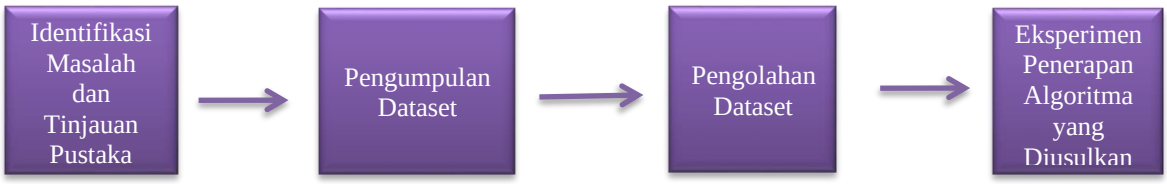
		Prediksi Kelas	
		Yes	***No***
Kelas Sebenarnya	***Yes***	TP	FN
	No	FP	TN

True positif dan *true negatif* menunjukan bahwa model klasifikasi mengkategorikan dengan benar, *false positif* dan *false negatif* menunjukan bahwa model klasifikasi mengkategorikan dengan salah [13]. Nilai akurasi dapat dihitung dengan rumus.

$$Akurasi = \frac{TP+TN}{TP+TN+FN+FP} \quad (5)$$

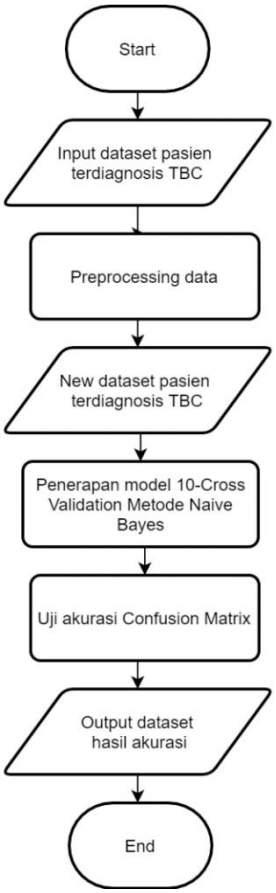
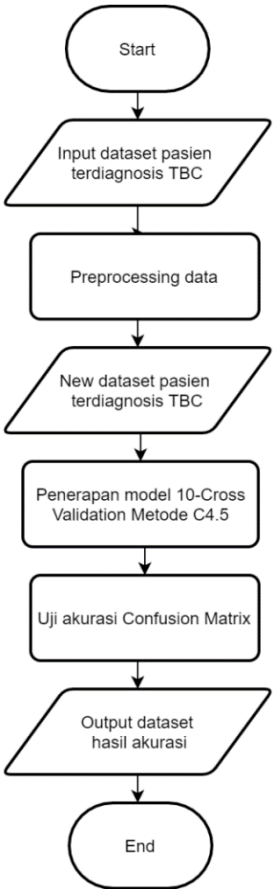
METODE PENELITIAN

Penelitian ini menggunakan metode *eksperimen*. Penelitian *eksperimen* merupakan penelitian yang digunakan untuk mencari pengaruh perlakukan tertentu terhadap yang lain dalam kondisi yang terkendalikan [14] Penelitian dilakukan di Puskesmas Kedungbanteng Kabupaten Banyumas Provinsi Jawa Tengah. Alat yang digunakan pada penelitian ini yaitu tools aplikasi rapidminer. Adapun tahapan penelitian yang dilakukan sebagai berikut:



Gambar 1. Tahapan Penelitian

1. Identifikasi Masalah dan Tinjauan Pustaka
   Mengidentifikasi masalah berupa penentuan klasifikasi penyakit *tuberculosis* dan metode *naïve* bayes dengan C4.5 dapat digunakan sebagai metode klasifikasi.
2. Pengumpulan Dataset
   Dataset dalam penelitian ini diambil dari data rekap medik Puskesmas Kedungbanteng pasien yang terdiagnosis *tuberculosis* paru dan pasien terdiagnosis *tuberculosis* ekstra paru dengan jumlah 200 *record* yang terdiri dari 142 *record* (71%) pasien *tuberculosis* paru dan 58 *record* (29%) pasien *tuberculosis* ekstra paru.
3. Pengolahan Dataset
   Pada tahap pengolahan dataset ini dilakukan *transformation data*, tahap ini dilakukan untuk mengubah atau mentransformasikan data menjadi simbol atau *coding* agar sesuai dengan algoritma yang digunakan untuk mempermudah proses penambangan informasi.
4. Eksperimen Penerapan Algoritma yang Diusulkan
   Penelitian ini menggunakan algoritma *Naïve Bayes* dan algoritma C4.5 dengan proses eksperimen sebagai berikut:

Langkah-langkah eksperimen Metode *Naïve Bayes*.	Langkah-langkah eksperimen Metode C4.5.
	

Gambar 2. Tahapan Eksperimen

HASIL DAN PEMBAHASAN

1. Dataset

Pengolahan data awal terhadap row data menghasilkan dataset dengan 11 atribut faktor penentu seperti pada Tabel 2.

Tabel 2. Struktur Data Penyakit *Tuberculosis*

No.	Nama Variabel	Keterangan	Proses
1.	Nama	Atribut	Tidak Digunakan
2.	Alamat	Atribut	Tidak Digunakan
3.	Kategori Usia	Atribut	Digunakan
4.	Jenis Kelamin	Atribut	Digunakan
5.	Jenis Batuk	Atribut	Digunakan
6.	Sakit Dinding Dada	Atribut	Digunakan
7.	Demam	Atribut	Digunakan
8.	Penurunan Nafsu Makan	Atribut	Digunakan
9.	Muncul Benjolan	Atribut	Digunakan
10.	Kondisi benjolan	Atribut	Digunakan
11.	Kekakuan bagian belakang	Atribut	Digunakan
12.	Jenis Penyakit	Label	Digunakan

Pada tahap pengolahan data ini juga dilakukan *transformation data.* pada algoritma klasifikasi atribut kategorikal dianggap lebih bermanfaat pada tahap penambangan [14]. Oleh karena itu, pada atribut usia yang memiliki tipe numerik diubah menjadi 6 interval kategori.

Tabel 3. Tabel Kategori usia

No	Usia	Kategori Usia
1	0-5 Tahun	Balita
2	6-10 Tahun	Anak-anak
3	11-19 Tahun	Remaja
4	20-44 Tahun	Dewasa
5	45-59 Tahun	Par-lansia
6	≤60 Tahun	Lansia

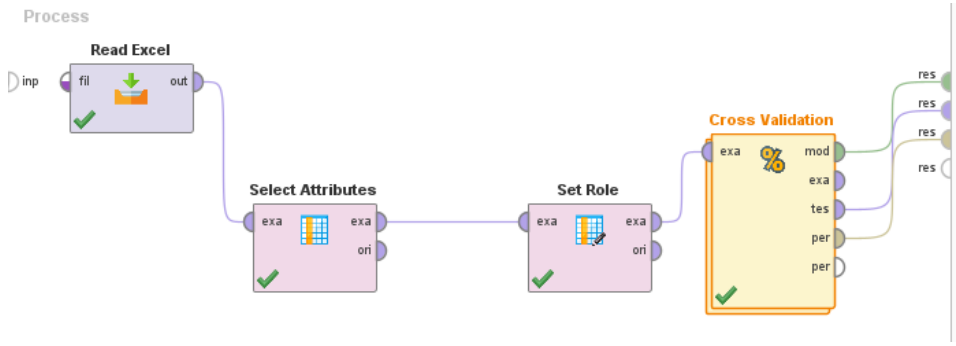
Berikut data dan digunakan dalam penelitian seperti yang terlihat pada Gambar 3

Kategori Usia	Jenis Kelamin	Jenis Batuk	Sakit dinding dada	ıa Waktu Demam (Pel	Penurunan Nafsu Makan	Munculnya Benjolan	Bobot Tubuh
Dewasa	L	Batuk Berdahak	Sakit dada terus	1	Iya	Leher,Ketiak,Sela Paha	Kurus
Anak-anak	L	Batuk Biasa	Sakit dada terus	2	Iya	Leher,Ketiak,Sela Paha	Kurus
Dewasa	P	Batuk Berdahak	Sakit dada terus	Kurang dari	Iya	Payudara	Kurus
Dewasa	L	Batuk Berdahak	Sakit setiap batuk	2	Iya	Punggung	Kurus
Remaja	L	Batuk Darah	Sakit setiap batuk	3	Iya	Leher,Ketiak,Sela Paha	Kurus
Dewasa	P	Batuk Biasa	Sakit dada terus	Lebih dari 3	Iya	Leher,Ketiak,Sela Paha	Kurus
Dewasa	L	Tidak Batuk	Tidak sakit dada	1	Iya	Punggung	Kurus
Dewasa	L	Batuk Biasa	Sakit dada terus	1	Tidak	Leher,Ketiak,Sela Paha	Gendut
Dewasa	P	Batuk Berdahak	Sakit setiap batuk	2	Iya	Leher,Ketiak,Sela Paha	Kurus
Dewasa	L	Batuk Berdahak	Sakit dada terus	2	Iya	Punggung	Kurus
Lansia	L	Batuk Berdahak	Sakit setiap batuk	1	Tidak	Leher,Ketiak,Sela Paha	Kurus
Dewasa	L	Batuk Darah	Sakit dada terus	lebih dari 3	Iya	Leher,Ketiak,Sela Paha	Gendut
Dewasa	P	Batuk Biasa	Sakit dada terus	2	Iya	Punggung	Kurus
Dewasa	P	Batuk Berdahak	Sakit setiap batuk	1	Iya	Leher,Ketiak,Sela Paha	Kurus
Dewasa	P	Batuk Berdahak	Sakit setiap batuk	lebih dari 3	Iya	Payudara	Kurus
Remaja	P	Batuk Biasa	Sakit dada terus	2	Iya	Punggung	Kurus
Dewasa	L	Tidak Batuk	Tidak sakit dada	kurang dari 1	Iya	Punggung	Kurus
Dewasa	P	Tidak Batuk	Tidak sakit dada	lebih dari 3	Iya	Leher,Ketiak,Sela Paha	Ideal
Dewasa	P	Tidak Batuk	Tidak sakit dada	2	Iya	Leher,Ketiak,Sela Paha	Kurus
Dewasa	P	Batuk Biasa	Sakit dada terus	kurang dari 1	Iya	Punggung	Kurus
Remaja	P	Batuk Biasa	Sakit setiap batuk	lebih dari 3	Iya	Leher,Ketiak,Sela Paha	Kurus
Remaja	P	Batuk Berdahak	Sakit dada terus	lebih dari 3	Iya	Leher,Ketiak,Sela Paha	Kurus
Dewasa	P	Batuk Berdahak	Sakit setiap batuk	2	Iya	Leher,Ketiak,Sela Paha	Kurus

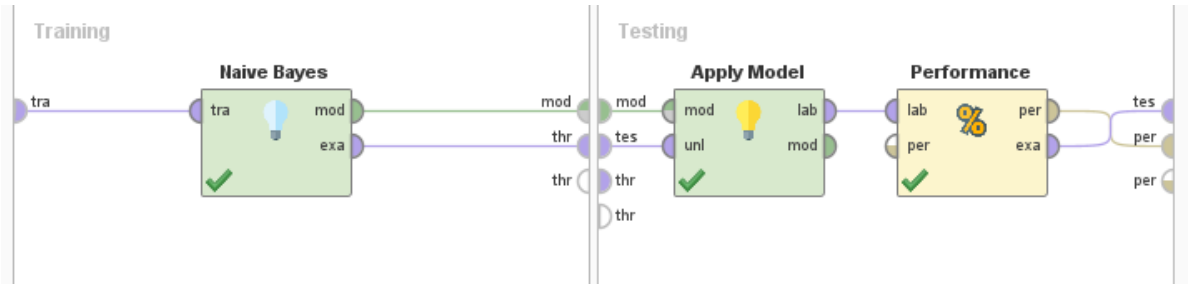
Gambar 3.Potongan Dataset

2. Proses Eksperimen menggunakan Rapidminer

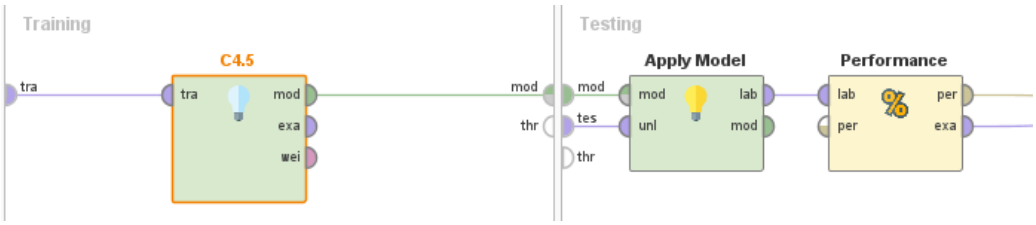
Proses Eksperimen yang dilakukan menggunakan tools aplikasi rapidminer dengan metode validasi 10-*fold cross validation* dan pengukuran akurasi menggunakan *confusion matrix* dilakukan terhadap *dataset* dengan algoritma *Naïve Bayes* dan algoritma C4.5. Berikut merupakan proses eksperimen yang dilakukan.



Gambar 4.Proses Eksperimen



Gambar 5.Penerapan Metode *Naïve Bayes*



Gambar 6. Penerapan Metode C4.5

3. Hasil Akurasi Confusion *Matrix*

accuracy: 83.00% +/- 11.11% (micro average: 83.00%)

	true Tuberculosis Paru	true Tuberculosis Ekstra Paru	class precision
pred. Tuberculosis Paru	134	26	83.75%
pred. Tuberculosis Ekstra Paru	8	32	80.00%
class recall	94.37%	55.17%	

Gambar 7. Hasil Akurasi *Naïve Bayes*

accuracy: 85.00% +/- 8.50% (micro average: 85.00%)

	true Tuberculosis Paru	true Tuberculosis Ekstra Paru	class precision
pred. Tuberculosis Paru	132	20	86.84%
pred. Tuberculosis Ekstra Paru	10	38	79.17%
class recall	92.96%	65.52%	

Gambar 8. Hasil Akurasi C4.5

4. Analisis Hasil
   Dari eksperimen tersebut didapatkan hasil yang dapat dilihat pada Tabel 4

Tabel 4.Tabel hasil eksperimen

No	Metode Eksperimen	Model pengujian Eksperimen	Metode Pengukuran akurasi	Hasil Akurasi
1	*Naïve Bayes*	10-Fold Cross Validation	*Confusion Matrix*	83%
2	C4.5	10-Fold Cross Validation	*Confusion Matrix*	85%

Berdasarkan hasil tersebut dapat disimpulkan algoritma C4.5 memiliki tingkat akurasi lebih tinggi dengan selisih 2% dari algoritma *naïve bayes* yang berarti algoritma C4.5 lebih baik saat diterapakan pada klasifikasi penyakit *tuberculosis*.

KESIMPULAN

Berdasarkan penelitian yang telah dilakukan dengan 200 sampel data pasien yang terdiagnosa *tuberculosis* di Puskesmas kedungbanteng, dengan dua kelas yaitu kelas "TBC Paru" dan "TBC Ekstra Paru" dan 11 atribut. Penelitian ini menggunakan aplikasi *rapidminer* dengan model *Cross Valid*ation dan pengujian akurasi menggunakan *confusion matrix.* pada eksperimen yang telah dilakukan didapatkan algoritma C4.5 memiliki akurasi tertinggi dengan rata-rata akurasi yang didapat sebesar 85% dengan selisih akurasi sebesar 2% dari algoritma *naïve bayes* dengam demikian algoritma algoritma C4.5 lebih baik saat diterapakan pada klasifikasi penyakit *tuberculosis.*

UCAPAN TERIMA KASIH

Kepada Puskesmas Kedungbanteng yang telah mengijinkan penulis melakukan penelitian.

DAFTAR PUSTAKA

[1] Yarmaliza., Zakiyudin. (2019). Pencegahan dini terhadap penyakit tidak menular (PTM) melalui germas. Pengabdian masyarakat multidisiplin.3(2): 93-100.
[2] Simanjuntak, D., Sindar, W. (2019). Sistem pakar deteksi gizi buruk balita dengan metode *naïve bayes classifier*. Infokar 1(2):54-60.
[3] Nikmatun, I. A., U., Waspada, I. U. (2019). Implementasi data mining untuk klasifikasi masa studi mahasiswa menggunakan algoritma k-nearest neighbor. 10(2):421–432.
[4] Nurlia E, Jajaluli M & Purnamasari I (2021). Penerapan *Naïve Bayes* untuk klasifikasi ringkat risiko diagnosis gigi di UPTD Puskesmas Cingambul.4(2):127-132.
[5] Syahirul, A., Peni, P. L., Rusliyawati. (2020). Sistem pakar diagnosa penyakit tanaman kakao menggunakan metode certainty factor pada kelompok tani PT olam indonesia (cocoa) cabang lampung. JDMSI. 1,No.4:26-31.
[6] S. Dutalia, A. K. Lalo, P. Batarius, Y. Carmeneja, and H. Siki, "Implementasi Algoritma C4 . 5 Untuk Klasifikasi Penjualan," vol. 06, pp. 1–12, 2021.
[7] S. F. Wulandari, W. Handoko, and U. B. B. F. Agis, "Perbandingan Naive Bayes Dan C45 Dalam Klasifikasi Tes Kesehatan Mahasiswa Baru Akbid AsSyifa," vol. 2, no. 3, pp. 167–176, 2022.
[8] S. Dutalia, A. K. Lalo, P. Batarius, Y. Carmeneja, and H. Siki, "Implementasi Algoritma C4 . 5 Untuk Klasifikasi Penjualan," vol. 06, pp. 1–12, 2021.
[9] K. R. Dewi, K. F. Mauladi, T. Informatika, F. Teknik, and U. I. Lamongan, "Analisa Algoritma C4 . 5 untuk Prediksi Penjualan Obat Pertanian di Toko Dewi Sri," pp. 109– 114, 2020.
[10] Kevin Merico Setiawan, F. Santi Wahyuni, and A. Faisol, "Perbandingan Algoritma C4.5 Dan Danive Bayes Untuk Menentukan Karyawan Berprestasi," JATI (Jurnal Mhs. Tek. Inform., vol. 5, no. 1, pp. 235–245, 2021, doi: 10.36040/jati.v5i1.3311.
[11] R. R. R. Arisandi, B. Warsito, and A. R. Hakim, "Aplikasi Naïve Bayes Classifier (Nbc) Pada Klasifikasi Status Gizi Balita Stunting Dengan Pengujian K-Fold Cross Validation," J. Gaussian, vol. 11, no. 1, pp. 130–139, 2022, doi: 10.14710/j.gauss.v11i1.33991.

[12] Nainggolan, D. S. Prasvita, and D. S. Bukit, "Klasifikasi Informasi Kesehatan Pada Data Media Sosial Menggunakan Support Vector Machine dan K-Fold Cross Validation," Malikussaleh J. Mech. Sci. Technol., vol. 5, no. 2, pp. 34–38, 2021, [Online]. Available: https://ojs.unimal.ac.id/mjmst/article/v iew/6317%0Ahttps://ojs.unimal.ac.id/ mjmst/article/download/6317/3169

[13] R. S. Putra, E. D. Putra, M. H. Rifqo, and H. Witriyono, "Klasifikasi Penyebaran Covid-19 Menggunakan Algoritma C4.5 Kota Pagar Alam," JUKOMIKA (Jurnal Ilmu Komput. dan Inform., vol. 4, no. 1, pp. 23–35, 2021, doi: 10.54650/jukomika.v4i1.346.

[14] M. R. Matondang, M. R. Lubis, and H. S. Tambunan, "Analisis Data mining dengan Metode C.45 pada Klasifikasi Kenaikan Rata-Rata Volume Perikanan Tangkap," Brahmana J. Penerapan Kecerdasan Buatan, vol. 2, no. 2, pp. 74–81, 2021, doi: 10.30645/brahmana.v2i2.68.