Jurnal Sistem Komputer dan Kecerdasan Buatan
Volume IV – Nomor 1 – September 2020

Penerapan Algoritma Naïve Bayes Untuk Klasifikasi Penyakit Diabetes Mellitus

Achmad Ridwan

Program Studi Sistem Informasi, Universitas Muhammadiyah Kudus, Jalan Ganesha 1 Purwosari Kudus
Email : achmadridwan@umkudus.ac.id

Diterima : 30 Agustus 2020
Disetujui : 30 September 2020

***Abstract*— Diabetes Mellitus atau kencing manis adalah penyakit metabolisme disebabkan oleh kadar gula tinggi didalam darah. Gula darah disimpan atau digunakan untuk energi yang berasal dari darah yang dipindahkan ke sel manusia oleh hormon insulin . ketika terserang Diabetes, pada tubuh manusia insulin tidak biasa dihasilkan secara cukup bahkan tubuh tidak dapat menggunakan insulin tersebut secara benar sesuai kebutuhan. Diabetes Mellitus terdaftar sebagai penyakit penyumbang kematian terbesar terbesar didunia. Diabetes Mellitus dapat diklasifikasikan berdasarkan kemungkinan terkenanya dari atribut gejala diawal fasenya. penyakit ini bisa dideteksi karena banyak gejala yang terdeteksi. Data yang digunakan pada analisis ini merupakan data dari dataset UCI Machine Learning yaitu Early Stage Diabetes Risk tahun 2020 dan terdiri 17 attribut. Analisis yang dilakukan meliputi data preprocessing, model, dan evaluasi. Pengujian Metode klasifikasi pada riset adalah Naïve Bayes Classification. Hasil klasifikasi menunjukkan akurasi sebesar 90.20% dan nilai AUCnya yaitu 0,95**

Keywords : Data Mining, Klasifikasi, Naïve Bayes, Diabetes Mellitus

I. PENDAHULUAN

Diabetes adalah penyakit dimana insulin tidak dapat diproduksi oleh pankreas dan digunakan oleh tubuh atau ketika tubuh manusia tidak dapat menggunakan insulin yang telah dibuat oleh pankreas dengan baik. Insulin merupakan hormon yang diciptakan oleh pankreas, yang berfungsi seperti kunci agar glukosa dari makanan yang manusia makan mengalir ke sel-sel dari darah dalam tubuh yang yang kemudian menghasilkan energi.

Makanan yang mengandung karbohidrat diproses menjadi glukosa dalam darah. glukosa dibantu insulin masuk ke dalam sel. Ketika tubuh tidak bisa memproduksi insulin dan bahkan menggunkan dengan benar menyebabkan meningkatnya kadar glukosa dalam darah. Sehingga pada waktujangka panjang, kadar glukosa tinggi dalam darah dapat merusak organ pada tubuh dan kegagalan fungsi organ dan jaringan.Beberapa peneliti membagi diabetes menjadi diabetes tipe 1, tipe 2, dan diabetes gestasional [1]. Diabetes gestasional adalah jenis diabetes yang hanya terjadi pada kehamilan karena perubahan hormonal. Gejala umum diabetes adalah poliuria, polidipsia, polifagia, penurunan berat badan mendadak (biasanya tipe 1), kelemahan, obesitas (biasanya tipe 2), penyembuhan tertunda, penglihatan kabur, gatal, iritabilitas, sariawan genital, paresis parsial, otot kaku, alopecia , dll. [2].

Di era 4.0 , teknologi komputer dapat membantu kita untuk mendeteksi penyakit secara akurat dan dapat menghemat waktu. Penambangan data adalah bidang penting dalam ilmu komputer yang digunakan untuk prediksi. Ini adalah proses menemukan data baru dari data

yang sebelumnya diketahui melalui analisis data [3]. Untuk memprediksi suatu penyakit dengan pendekatan data mining diperlukan gejala yang disertai dengan data klinis. Gejala merupakan faktor yang sangat penting untuk pasien baru dan prediksi tahap awal dengan data gejala. Kami juga membutuhkan data klinis untuk menganalisisnya

Salah satu algoritma *data mining* adalah Naïve Bayes. Metode Naïve Bayes digunakan mengklasifikasikan penyakit *Diabetes Mellitus*

Research ini menganalisis klasifikasi Naïve Bayes untuk mengklasifikasi gejala awal penyakit *Diabetes Mellitus* sehingga mendapatkan akurasi yamg di dapat dari hasil proses evaluasi.

II. KAJIAN LITERATUR

A. Data Mining

Data mining adalah sebuah ilmu yang mempelajari alur kerja pengalian data atributnya yang saling berkaitan untuk menemukan pengetahuan atau menemukan sebuah pola dari suatu data yang besar. penggalian dari data ke pengetahuan dapat disimpulkan mempunyai 3 kunci yaitu :

Data karena data tersebut adalah fakta yang terekam dan tidak membawa arti Informasi merupakan rangkuman penjelasan dan statistik dari data. Pengetahuan ini merupakan hasil dari penggalian dari data Nama lain dari data mining adalah Knowledge Discovery in Data, ada juga tentang Big data atau bussines intelijen, Knowledge Extraction dan pattern analisis atau information harvesting .

Konsep dari data mining yaitu himpunan data yang telah terhimpun banyak sekali kemudian dengan metode data mining yang ada diproses dengan rumus/algoritma data-data tersebut diambil atau diekstraksi untuk menjadi di sebuah pengetahuan. Pengetahuan Itulah bisa untuk diambil untuk sebuah tujuan tertentu. Pengaturan tersebut sangat berguna untuk berbagai keperluan .

B. Algoritma Naïve Bayes

Algoritma naïve bayes adalah algoritma pembelajaran mesin untuk masalah klasifikasi yang terutama digunakan untuk klasifikasi teks yang melibatkan kumpulan data pelatihan berdimensi tinggi. Beberapa contohnya adalah analisis sentimental penyaringan spam dan mengklasifikasikan tidak hanya dikenal karena Kesederhanaannya tetapi juga untuk Efektivitasnya. Dengan algoritma Naive bayes dapat membangun model dengan cepat dan menjadikannya algoritma prediksi yang paling cepat untuk dipelajari. Algoritme ini menggunakan probabilitas suatu objek . Mengapa disebut algoritma naïve bayes karena membuat asumsi bahwa kemunculan fitur tertentu tidak tergantung pada kemunculan fitur lain bahkan jika ciri-ciri ini bergantung satu sama lain atau pada keberadaan ciri-ciri lainnya, semua sifat ini secara individual berkontribusi pada probabilitas dan itulah mengapa disebut naif Algoritma ini mengacu pada ahli statistik dan filosof Thomas Bayes. Dasar dari algoritma naïve bayes adalah teorema dasar yang secara alternatif dikenal sebagai aturan Bayes atau Hukum bayes algoritma ini adalah metode untuk menghitung probabilitas kondisional yaitu probabilitas suatu peristiwa berdasarkan pengetahuan sebelumnya yang tersedia [4] .

Naïve Bayes memiliki kemampuan yang cepat dalam membuat model, mempunyai kemampuan memprediksi dan juga menyediakan metode baru dalam mengeksplor dan memahami data. Algoritma Naïve Bayes hanya mendukung pada atribut yang bertipe data discrete atau discretized, atau tidak mendukung atribut yang bernilai continuous (numerik) dan semua atribut dapat menjadi independen, menjadi atribut yang memberi kontribusi kepada atribut yang diprediksi.

Secara singkat algoritma naïve bayes classification adalah pengklasifikasi kumpulan data statistika yang mana untuk memprediksi semua probabilitas tiap anggota suatu class. Neural Network Dan Decision Tree memiliki persamaan kekuatan klasifikasi dengan Naïve Bayes yang didasarkan pada teorema Bayes.

Naïve Bayes memiliki bukti kecepatan dan akurasi yang tinggi saat digunakan ke dalam kumpulan data data yang besar[5].

$$P(H \mid X) = \frac{P(X \mid H)P(H)}{P(X)} \qquad (1)$$

Rumus teorema Naïve bayes :

X = Data dengan class yang belum diketahui

H = Hipotesis data X merupakan suatu class spesifik

P(H|X) = Probabilitas hipotesis H berdasarkan kondisi x (posteriori prob.)

P(H) = Probabilitas hipotesis H (prior prob.)

P(X|H) = Probabilitas X berdasarkan kondisi tersebut

P(X) = Probabilitas dari X

C. *Rapid Miner*

Rapidminer adalah sebuah paket tools machine learning praktis yang berfungsi untuk penelitian pendidikan dan berbagai aplikasi. Rapidminer mampu menyelesaikan masalah data mining di dunia nyata khususnya klasifikasi yang mendasari pendekatan mesin learning. perangkat lunak ini ditulis dalam hierarki Java dengan metode berorientasi objek dan dapat berjalan hampir di semua platform sistem operasi. Rapidminer sangat mudah digunakan serta diterapkan pada tingkatan yang berbeda. Rapidminer mengimplementasi algoritma pembelajaran yang dapat diterapkan pada data sheet dari command Line. Rapidminer mengandung tools untuk preprocessing data, klasifikasi, klastering, Regresi, Asosiasi dan Visualisasi . Rapidminer dapat melakukan Processing pada data, memasukkannya dalam skema pembelajaran dan menganalisa classifier yang dihasilkan dan performansinya. semua itu itu tanpa menulis kode program sama sekali. Contoh penggunaan Rapidminer adalah dengan menerapkan sebuah metode pembelajaran ke dataset dan menganalisa hasilnya untuk memperoleh informasi tentang data atau menerapkan beberapa metode dan membandingkannya performansi untuk dipilih.tools yang dapat digunakan untuk preprocessing dataset membuat pengguna berfokus pada algoritma yang digunakan tanpa terlalu memperhatikan detail seperti pembacaan data dari file implementasi algoritma filtering dan penyediaan kode untuk evaluasi hasil mengikuti model rilis Linux.

D. *Evaluasi Algoritma Klasifikasi Data Mining*

D.1 Evaluasi Confusion Matrix

Untuk melakukan evaluasi terhadap model klasifikasi berdasarkan perhitungan objek testing mana yang diprediksi benar dan tidak benar. Confusion Matrix berisi informasi tentang aktual (actual) dan prediksi (predicted) pada sistem klasifikasi. Kinerja sistem seperti ini biasanya dievaluasi dengan menggunakan data pada matriks. Perhitungan dimasukan kedalam tabel Confusion Matrix [6]. Gambaran tentang Confusion Matrix di Tabel berikut :

Tabel 2.1 Confusion Matrix

Klasifikasi yang memprediksi kelas		
Kelas :Ya	Kelas: tidak	
OBSERVED	Kelas : ya	
	a	b
KELAS	(Benar Positive-TP)	(Salah Negative-FN)
Kelas: Tidak	c	d
(salah Positif -Fp)	(Ya negative-TN)	

Pada Tabel 2.1 Prediksi Salah positif adalah tuple positif di data set yang mengklasifikasikan negatif salah. Negatif adalah positif mengklasfikasikan jumlah tuple negatif . Dan untuk benar Positif adalah positif di data set mengklasifikasikan kolom positif didata set , True negatives merupakan negatif di data set mengklasifiksikan tupel negatif

Proses berikutnya akan dihitung akurasi, spesifik, PPV, NPV. Sensitivity adalah proses perhitungan pembandingankan jumlah tuple yang positif. Sedangkan specificity adalah proses

perhitungan pembandingankan negatif terhadap jumlah tuple negatif.

Selanjutnya PPV adalah perbandingan kasus dengan hasil diagnosa positif, Negatives Predictive adalah perbandingan dimana kasusyang sama dengan hasil diagnosis negatif.

Confusion matrix berfungsi untuk membuat penilaian kerja model klasifikasi yang mempunyai jumlah objek yang diramal dengan benar dan salah [7]. Akurasi kelas minoritas dapat menggunakan metrik recall Rumus-rumus yang digunakan untuk melakukan penghitungannya adalah:

Akurasi adalah perbandingan jumlah prediksi yang benar. Semua ditentukan dengan mengimplementasikan rumus akurasi berikut:

$$Accuracy = \frac{a+b}{a+b+c+d} = \frac{TP+TN}{TP+TN+FP} \quad (2)$$

Sensitivity disebut juga dengan recall. Jika sensitivity 100% sama artinya dengan pengklasifikasian menganggap kasus yang diamati positif. Sebai contoh semua orang memiliki tumor ganas dianggap sakit.

$$Recall = \frac{number\ of\ True\ Possitive}{number\ of\ True\ Possitive + number\ of\ False\ Negative} \quad (3)$$

Presisi adalah tingkat positif salah adalah perbandingan nilai positif yang salah diklasifisikan pada kasus negatif , yang perhitungannya menggunakan Rumus :

$$Precision = \frac{number\ of\ true\ possitive}{number\ of\ true\ possitive + number\ of\ false\ possitive} \quad (4)$$

D. 2 (Dua) Kurva ROC

Kurva ROC digunakan untuk menilai hasil prediksi, kurva ROC adalah teknik untuk menggambarkan , mengatur, dan memilih pengklasifikasian berdasarkan kinerja mereka[8]. Kurva ROC adalah tool dua dimensi di dunia data mining yang digunakan untuk menilai kinerja klasifikasi yang menggunakan dua class keputusan, tiap objek dipetakan ke salah satu elemen dari himpunan pasangan, positif atau negatif. Pada gambar kurva ROC, True Positif rate diplot pada sumbu Y dan False Positif rate diplot pada sumbu X. Menurut Gorunescu Dalam metode klasifikasi data mining , hasil AUC dapat dibagi menjadi beberapa kelompok [9]:

1. 0,90 - 1,00 = Klasifikasi Sangat Baik
2. 0,80 - 0,90 = Klasifikasi Baik
3. 0,70 - 0,80 = Klasifikasi Sedang
4. 0,60 - 0,70 = Klasifikasi Buruk
5. 0,50 - 0,60 = Kegagalan

III. METODOLOGI PENELITIAN

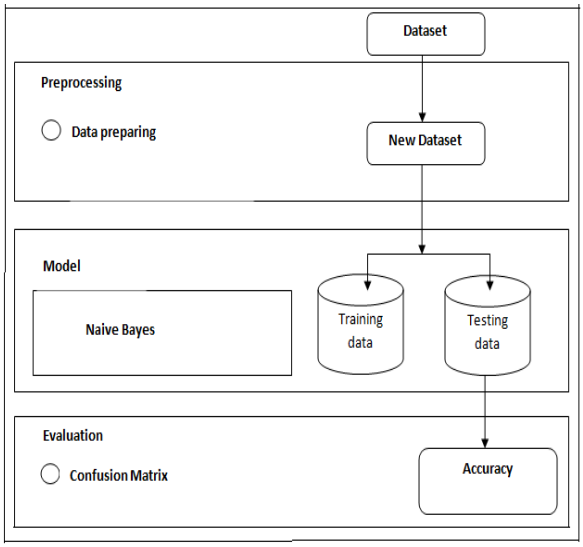
Pendekatan utama dalam penelitian ini yaitu pendekatan kualitatif dan pendekatan kuantitatif. Dalam penelitian ini metode yang digunakan yaitu metode penelitian kuantitatif. Tujuan dari penelitian ini adalah melakukan klasifikasi dan evaluasi model algoritma Naïve Bayes untuk mengetahui akurasi algoritma Naïve Bayes dalam mengklasifikasikan penyakit Diabetes Mellitus.

A. *Sumber Data*

Sumber data yang digunakan pada penelitian ini adalah data dari dataset UCI Machine Learning Repository .Dataset yang digunakan adalah Early stage diabetes risk prediction dataset dimana file tersebut bernama diabetes_data_upload.csv. Variabel yang digunakan pada penelitian ini adalah sebanyak 17 variabel dengan jumlah data sebanyak 520. Ini termasuk data tentang orang-orang termasuk gejala yang dapat menyebabkan diabetes. Kumpulan data ini dibuat dari kuesioner langsung kepada orang-orang yang baru saja menjadi penderita diabetes, atau yang masih nondiabetes tetapi memiliki sedikit atau lebih gejala.

B. Kerangka Pemikiran

Kerangka penelitian dari penelitian bisa kita lihat pada Gambar dibawah ini :



Gambar 3.1 Kerangka Pemikiran

Adapun kerangka pemikiran pada Gambar 3.1 dapat dijelaskan sebagai berikut :

1. Preprocessing

Preprocessing data dilakukan dengan cara menangani nilai yang hilang mengikuti teknik mengabaikan tupel dengan nilai yang tidak lengkap. Setelah praproses, total 500 instans telah tersisa. Diantaranya, 314 adalah nilai positif dan 186 adalah nilai negatif. Deskripsi detail dari dataset dan atributnya ditunjukkan pada Tabel 4.1. Dua variabel kelas digunakan untuk mengetahui apakah pasien memiliki risiko diabetes (positif) atau tidak (negatif).

2. Model

Tool RapidMiner diterapkan pada Dataset yang baru untuk di Training maupun di Testing pada Algoritma Naïve Bayes, serta untuk menganilisis hasilnya antara lain Kesalahan klasifikasi (Classification Error), nilai akurasi Probabilitas maksimal tiap kelas, Recall dan Presisinya

3. Evaluasi

Kemudian Dataset diuji/dievaluasi dengan Confusion Matrix serta diukur tingkat akurasinya.

IV. HASIL DAN PEMBAHASAN

Variabel data penelitian yang digunakan pada penelitian ini disajikan pada Tabel 4.1 yakni sebagai berikut.

Tabel 4.1 Deskripsi Variable Dataset

No.	Atribut	Value
1	Age	1.20–35, 2.36–45, 3.46–55,4.56–65, 6.above 65
2	Sex	1.Male, 2.Female
3	Polyuria	1.Yes, 2.No.
4	Polydipsia	1.Yes, 2.No.
5	Sudden weight loss	1.Yes, 2.No.
6	Weakness	1.Yes, 2.No.
7	Polyphagia	1.Yes, 2.No.
8	Genital thrush	1.Yes, 2.No.
9	Visual blurring	1.Yes, 2.No.
10	Itching	1.Yes, 2.No.
11	Irritability	1.Yes, 2.No.
12	Delayed healing	1.Yes, 2.No.
13	Partial paresis	1.Yes, 2.No.
14	Muscle stiffness	1.Yes, 2.No.
15	Alopecia	1.Yes, 2.No.
16	Obesity	1.Yes, 2.No.
17	Class	1.Positive, 2.Negative.

Dari Tabel 4.1 Ada 16 variabel dataset gejala dan 1 variabel class penentu klasifikasi.

A. Preprocessing

Dari dataset yang akan di analisis dijadikan sebagai Data Training dan Data Testing yang ada di klasifikasikan oleh algoritma Naïve Bayes

Adapun contoh data training dan data testing dapat di lihat pada tabel 4.2 :

Tabel 4.2 Data Training dan Testing

No.	Atribut	Value	
		1	2
1	Age	40	58
2	Sex	male	female
3	Polyuria	no	no
4	Polydipsia	yes	yes
5	Sudden weight loss	yes	yes
6	Weakness	yes	yes
7	Polyphagia	no	no

8	Genital thrush	no	no
9	Visual blurring	yes	yes
10	Itching	yes	yes
11	Irritability	no	no
12	Delayed healing	no	no
13	Partial paresis	yes	yes
14	Muscle stiffness	yes	yes
15	Alopecia	no	no
16	Obesity	no	no
17	Class	negatif	positif

Setelah praproses, total 500 data yang telah tersisa. Diantaranya, 314 adalah Class nilai positif dan 186 adalah Class nilai negatif.

B. Model

Setelah Praprocessing, data Training dan Data Testing akan kita proses klasifikasi menggunakan aplikasi Rapidminer Adapun hasil dari Confunsion Matrix.nya :

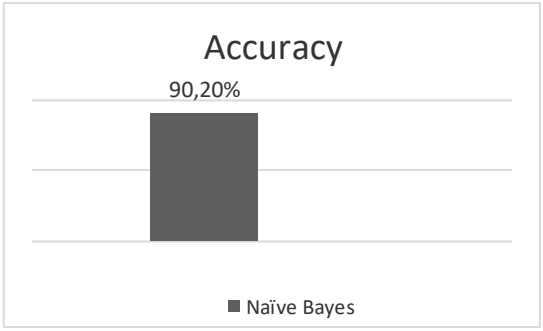
Tabel 4.3 Tabel Hasil Class Recall dan Precision

	true Positive	true Negative	class precision
pred. Positive	32	2	94.12%
pred. Negative	3	14	82.35%
class recall	91.43%	87.50%	

Dari Tabel 4.3 didapatkan Class Precision = 82.35%, dan Class Recall: 87.50%

C. Evaluasi

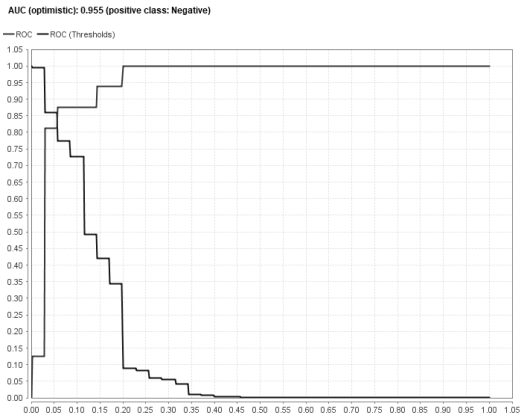
Hasil dari evaluasi pengklasifikasian dengan Naïve Bayes menghasilkan :



Gambar 4.1 Akurasi Naïve Bayes

Dari gambar diatas didapatkan akurasi untuk mengklasifisikan Dataset Diabetes yaitu 90,20%

Adapun hasil dari pengujian ini juga menghasilkan gambar 4.2 dibawah ini



Gambar 4.2 Kurva Roc Model Naive Bayes

Kurva ROC (Reciver Operating Characteristic) diatas menunjukkan algoritma Naïve Bayes memiliki nilai AUC sebesar 0.955 yang artinya Excellent Classification (Sangat Bagus) Dalam hasil penelitian, menunjukkan Naïve Bayes memberikan akurasi yang bagus untuk prediksi penyakit diabetes pada UCI dataset Machine Learning Repository.

V. KESIMPULAN

Terdapat 16 atribut yang mempengaruhi Klasifikasi dataset Early stage diabetes risk prediction yaitu Age, Gender, Polyuria, Polydipsia, sudden weight loss, weakness, Polyphagia, Genital thrush, visual blurring, Itching, Irritability, delayed healing, partial paresis, muscle stiffness, Alopecia, Obesity, class.

Penelitian ini menggunakan Algoritma Naïve Bayes untuk pengklasifikasian Dataset Early Stage Diabetes Risk Prediction. Dari 520 data kotor setelah dilakukan pembersihan data tidak lengkap dan melalui proses seleksi sesuai yang dibutuhkan dalam analisa Dataset Stage Diabetes maka didapat data 500 data bersih.

Dari hasil proses Klasifikasi permodelan algoritma naïve bayes terhadap Dataset Stage Diabetes menghasilkan Class Precision = 82.35%, dan Class Recall: 87.50%

Hasil evaluasi klasifikasi ini menghasilkan akurasi untuk mengklasifisikan Dataset Diabetes yaitu 90,20% ini artinya Klasifikasi permodelan algoritma Naïve Bayes terhadap Dataset Stage Diabetes sudah bagus akurasinya, tetapi perlu peningkatan akurasi dengan Assemble atau dengan cara yang lain.

Kurva ROC (Reciver Operating Characteristic) Klasifikasi permodelan algoritma Naïve Bayes terhadap Dataset Stage Diabetes menunjukkan algoritma Naïve Bayes memiliki nilai AUC sebesar 0.955 yang artinya Excellent Classification (Sangat Bagus).

DAFTAR PUSTAKA

[1] World Health Organization, “Global Report on Diabetes,” *Isbn*, 2016, doi: ISBN 978 92 4 156525 7.

[2] American Diabetes Association, “2016 American Diabetes Association (ADA) Diabetes Guidelines Summary Recommendation from NDEI,” *Natl. Diabetes Educ. Initiat.*, 2016.

[3] D. T. Larose, *Data Mining Methods and Models*. 2006.

[4] S. Kusumadewi, “KLASIFIKASI STATUS GIZI MENGGUNAKAN NAIVE BAYESIAN CLASSIFICATION,” *CommIT (Communication Inf. Technol. J.*, 2009, doi: 10.21512/commit.v3i1.506.

[5] A. Naik and L. Samant, “Correlation Review of Classification Algorithm Using Data Mining Tool: WEKA, Rapidminer, Tanagra, Orange and Knime,” *Procedia Comput. Sci.*, vol. 85, pp. 662–668, 2016, doi: 10.1016/j.procs.2016.05.251.

[6] A. Luque, A. Carrasco, A. Martín, and A. de las Heras, “The impact of class imbalance in classification performance metrics based on the binary confusion matrix,” *Pattern Recognit.*, 2019, doi: 10.1016/j.patcog.2019.02.023.

[7] F. Gorunescu, “Data mining: Concepts, models and techniques,” *Intell. Syst. Ref. Libr.*, 2011, doi: 10.1007/978-3-642-19721-5.

[8] Z. H. Hoo, J. Candlish, and D. Teare, “What is an ROC curve?,” *Emerg. Med. J.*, 2017, doi: 10.1136/emermed-2017-206735.

[9] A. Ridwan, P. N. Andono, and C. Supriyanto, “Optimasi Klasifikasi Status Gizi Balita Berdasarkan Indeks Antropometri Menggunakan Algoritma Naive,” *Teknol. Inf.*, 2018.