The Indonesian Journal of Computer Science
www.ijcs.net
Volume 13, Issue 5, October 2024
https://doi.org/10.33022/ijcs.v13i5.4354

Akurasi Metode Mesin Pembelajaran Dalam Analisis Variabel Penting Faktor Risiko Sindrom Down

Oscar Oleta Palit[1],Rafi Prayoga Dhenanta[2], Agnes Indarwati Susanto[3], Adzky Mathla Syawly[4], Aditya Purwa Santika[5], Mochamad Ikbal Arifyanto[6], Atthar Luqman Ivansyah[7], Fahdzi Muttaqien[8]
20923001@mahasiswa.itb.ac.id[1], fahdzi@itb.ac.id[8]
[1,2,3,4,5,6,7,8] Program Studi Magister Sains Komputasi, Fakultas Matematika dan Ilmu Pengetahuan Alam, Institut Teknologi Bandung, Jalan Ganesa 10, Bandung 40321

Informasi Artikel

Diterima : 2 Agu 2024
Direview : 20 Sep 2024
Disetujui : 1 Okt 2024

Kata Kunci

sindrom Down, kasus-kontrol, mesin pembelajaran, akurasi, variabel penting.

Abstrak

Penelitian ini bertujuan untuk menentukan faktor risiko sindrom Down dari data kasus-kontrol menggunakan mesin pembelajaran. Metode yang digunakan meliputi *random forest*, k-*nearest neighbors*, *support vector machine* (SVM), n*aive bayes*, *k-means*, dan a*rtificial neural network* (ANN). Data memiliki tiga subjek kelompok, sindrom Down (41 baris), tuna grahita (41 baris), dan tanpa kebutuhan khusus (27 baris) dengan total 37 variabel. Tingkat akurasi setiap metode ditentukan oleh jenis subjek dan jumlah variabel yang digunakan. Kelompok yang menggunakan seluruh subjek dan variabel menghasilkan akurasi yang rendah. Metode yang memberikan akurasi tertinggi adalah *naive bayes* sebesar 86% dengan menggunakan data yang nilai fitur pentingnya tinggi (variabel penting) pada subjek sindrom Down (kasus) dan tanpa kebutuhan khusus (kontrol). Hasil menunjukkan bahwa variabel usia ibu, usia ayah, dan jarak waktu bekerja orang tua sebelum anak lahir merupakan faktor yang paling berpengaruh terhadap sindrom Down. Kategori *supervised learning* bekerja lebih baik dibanding *unsupervised learning* pada klasifikasi data kasus-kontrol.

Keywords

Down syndrome, case control, machine learning, accuracy, important variables.

Abstract

This study aims to determine the risk factors for Down syndrome from case-control data using machine learning. The methods used include random forest, k-nearest neighbors, support vector machine (SVM), naive bayes, k-means, and artificial neural network (ANN). The data has three subject groups, Down syndrome (41 rows), mentally retarded (41 rows), and without special needs (27 rows) with a total of 37 variables. The level of accuracy of each method is determined by the type of subject and the number of variables used. The group that uses all subjects and variables produces low accuracy. The method that provides the highest accuracy is naive bayes at 86% using data with high feature importance values (important variables) in Down syndrome subjects (cases) and without special needs (controls). The results show that the variables of maternal age, paternal age, and the distance of parents' work before the child was born are the high-risk factors for Down syndrome. The supervised learning category performs better than unsupervised learning in classifying case-control data.

A. Pendahuluan

Sindrom Down merupakan salah satu penyakit yang disebabkan oleh kesalahan genetik atau biasa disebut sebagai penyakit autosomal. Kesalahan yang menyebabkan sindrom Down terjadi ketika ada pembelahan tidak sempurna dari kromosom pada fase meiosis saat proses pembentukan sel telur (oogenesis) [1]. Pembelahan tidak sempurna tersebut dinamakan dengan istilah *nondisjunction*. Penyakit yang disebabkan oleh *nondisjunction* bukan hanya sindrom Down, karena bisa terjadi di berbagai kromosom yang berpasang-pasangan. Sindrom Down terjadi pada kromosom 21, sehingga sering disebut trisomi 21, sedangkan contoh penyakit lain yang terjadi karena *nondisjucntion* adalah sindrom Edwards (trisomi 18) dan sindrom Patau (trisomi 13) [2]. Beberapa kelainan genetik ini dinamakan sesuai dengan ilmuan yang pertama mendeskripsikannya, seperti sindrom Down diambil dari nama dokter John Langdon Down setelah dideskripsikan pada tahun 1866 dan John H. Edwards untuk sindrom Edwards pada tahun 1960 [3], [4].

Secara epidemiologis, penyakit sindrom Down merupakan penyakit autosomal terbesar di dunia yang disebaban oleh *nondisjuction*. Perkiraan jumlah kelahiran sindrom Down adalah satu banding tujuh ratus sampai seribu kelahiran[5]Sampai saat ini belum diketahui secara spesifik apa yang menyebabkan terjadinya sindrom Down, namun banyak penelitian yang telah dilakukan untuk mengetahui faktor risikonya. Berdasarkan sejarah, faktor risiko yang pertama kali dianggap menyebabkan sindrom Down adalah umur ayah, namun penemuan tersebut diperbaiki ketika umur ibu melahirkan menjadi faktor yang lebih signifikan [6]. Selain faktor dari orang tua yang dimiliki seperti umur ketika anak lahir, faktor internal lain yang pernah diteliti adalah pemakaian kontrasepsi, riwayat keguguran, dan riwayat penyakit autosomal yang terjadi di keluarga. Sejalan dengan hal tersebut, beberapa penelitian juga mencari tahu faktor eksternal seperti keberadaan tempat tinggal dengan industri, riwayat pekerjaan orang tua, perilaku merokok, dan sumber air minum. Faktor-faktor tersebut ada yang terlihat signifikan pada beberapa penelitian, meskipun belum dapat menjelaskan secara spesifik seberapa berpengaruh terhadap kejadian sindrom Down [7].

Metode yang dilakukan untuk mencari faktor risiko sindrom Down biasanya menggunakan metode penelitian kesehatan yang dilakukan secara epidemiologis atau eksperimental. Metode epidemiologis biasanya membandingkan dua populasi berbeda seperti yang sakit dengan yang sehat lalu mencari perbedaannya, sedangkan metode ekperimental dengan memberikan perlakuan khusus terhadap subjek eksperimen di laboratorium[6]Tingkat faktor risiko kemudian diukur menggunakan nilai statistik yang dihitung berdasarkan temuan pada populasi atau kelompok subjek yang berbeda [8]. Berlandaskan pendekatan tersebut, penelitian ini bertujuan memanfaatkan mesin pembelajaran untuk melihat variabel yang berpengaruh terhadap faktor risiko sindrom Down dengan mengaplikasikan berbagai metode yang ada sebagai pengganti perhitungan statistik yang biasa digunakan pada penelitian epidemilogis. Metode yang digunakan seperti *random forest, k-nearest neighors, support vector machine,* dan *naive bayes* untuk *supervised learning*, serta menggunakan metode *k-means* dan *artificial neural network* untuk *unsupervised learning.* Perbedaan *supervised* dan *unsupervised learning* terletak pada label yang diberikan pada setiap data. S*upervised learning* memberikan label pada setiap data, sehingga mengetahui kelompok yang sindrom Down dan bukan,

sedangkan *unsupervised learning* tidak memberikan label sehingga data belum diketahui sebagai kelompok apa [9].

Fungsi digunakannya perbedaan metode untuk melihat apakah pada proses klusterisasi data akan terbagi sesuai dengan kelompoknya, seperti kelompok dengan sindrom Down dengan yang tidak. Selain itu, penggunaan berbagai metode mesin pembelajaran dapat menyimpulkan metode yang tepat untuk menganalisis data penelitian dari studi epidemiologis yang ada. Data yang digunakan adalah data pada penelitian studi epidemiologis yang dilakukan di Kota dan Kabupaten Tangerang [7]. Tujuan penelitian ini adalah mencari faktor risiko (variabel) yang berpengaruh terhadap kejadian penyakit, dalam hal ini adalah sindrom Down, menggukan mesin pembelajaran dan mencari tahu metode mesin pembelajaran yang tepat dengan melihat nilai akurasi untuk menganalisis data epidemiologis dengan metode kasus-kontrol dengan jumlah data yang terbatas.

B. Metode Penelitian

Data yang digunakan pada penelitian ini bersumber dari studi epidemiologis dengan metode kasus-kontrol yang dilakukan di Sekolah Kebutuhan Khusus yang berada di Kota dan Kabupaten Tangerang. Metode kasus-kontrol adalah cara pengumpulan data dengan melihat perbedaan dari dua subjek responden. Subjek kasus biasanya adalah orang dengan penyakit yang ingin diamati, sedangkan subjek kontrol adalah pembanding dari orang yang tidak mengalami peyakit tersebut. Subjek kasus pada data yang digunakan adalah responden dengan sindrom Down, sedangkan subjek kontrolnya adalah responden yang mengalami tuna grahita dan tidak mengalami kebutuhan khusus. Data epidemiologis tersebut dikumpulkan dengan wawancara menggunakan kuesioner. Data yang telah diperoleh diinput pada aplikasi pengolahan data statistik dan dilakukan pembersihan data dengan menghapus beberapa variabel yang tidak relevan. Setelah melakukan pembersihan data, ada 37 variabel yang digunakan untuk dilakukan analisis dengan targetnya adalah tipe subjek. Berikut adalah variabel yang digunakan pada penelitian, dengan tipe data dan kode yang dipakai.

Tabel 1. Variabel data penelitian.

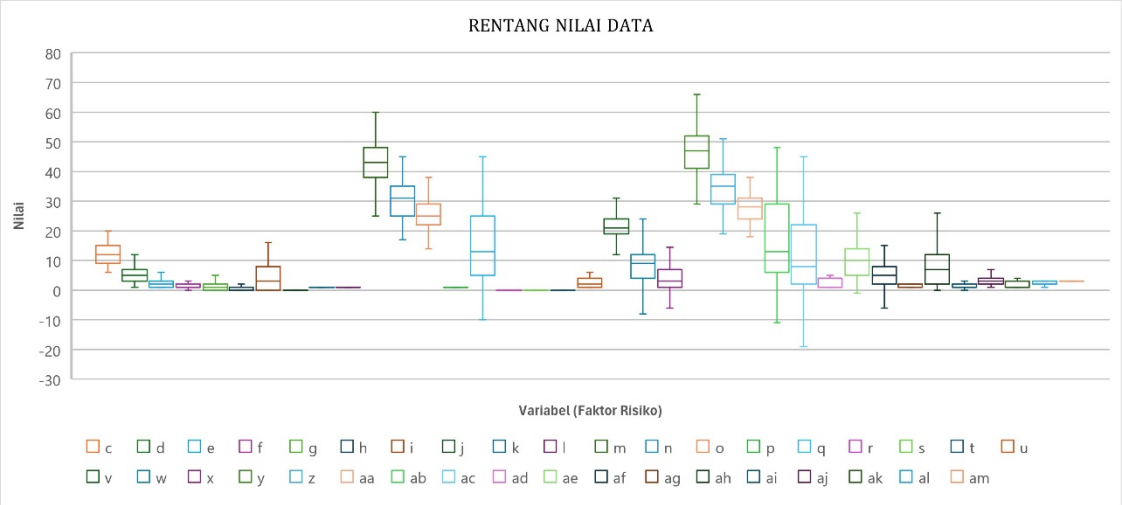
Variabel	Tipe Data	Kode
Target	ordinal	b
Usia	numerik	c
Jenjang pendidikan	numerik	d
Urutan kelahiran	numerk	e
Jumlah Saudara Kandung	numerik	f
Jumlah Kakak	numerik	g
Jumlah Adik	numerik	h
Selisih umur dengan Kakak	numerik	i
Selisih umur dengan Adik	numerik	j
Saudara dengan DS	ordinal	k
Orang tua kerabat atau tidak	ordinal	l
Usia Ibu saat ini	numerik	m
Usia Ibu saat melahirkan responden	numerik	n
Usia Ibu saat lahir anak pertama	numerik	o

Variabel	Tipe Data	Kode
Hubungan darah antar orang tua Ibu	ordinal	p
Jarak Tahun Tinggal - Melahirkan subjek	numerik	q
Keguguran	numerik	r
Aborsi	numerik	s
Lahir Meninggal	numerik	t
Pekerjaan Ibu sebelum anak lahir	nominal	u
Usia Ibu saat mulai bekerja	numerik	v
Jarak tahun Ibu bekerja dengan melahirkan subjek	numerik	w
Jarak tahun Ibu bekerja dengan lahir anak pertama	numerik	x
Usia Ayah saat ini	numerik	y
Usia Ayah saat responden lahir	numerik	z
Usia Ayah saat lahir anak pertama	numerik	aa
Jarak Tahun Tinggal - Subjek lahir	numerik	ab
Jarak Tahun Tinggal - anak pertama lahir	numerik	ac
Pekerjaan Ayah	nominal	ad
Jarak tahun Ayah bekerja dengan subjek lahir	numerik	ae
Jarak tahun Ayah bekerja dengan anak pertama lahir	numerik	af
Masih bekerja atau tidak	nominal	ag
Jangka waktu pernikahan - kelahiran subjek	numerik	ah
Jangka waktu pernikahan - Lahir anak pertama	numerik	ai
Jumlah kehamilan	numerik	aj
Sumber air minum	nominal	ak
Pengeluaran sebelum anak lahir	nominal	al
Pengeluaran setelah anak lahir	nominal	am

Dataset yang digunakan setelah dilakukan pembersihan data berjumah 111 baris, terdiri dari 41 baris data sindrom Down, 43 baris data tuna grahita, dan 27 baris data subjek tanpa kebutuhan khusus. Persebaran data setelah dilakukan pembersihan terlihat pada Gambar 1, yang dapat menampilkan rentang data numerik. Rentang tersebut memperlihatkan adanya data pencilan pada beberapa variabel dan memperlihatkan standar deviasi dari beberapa data numerik. Data yang memiliki standar deviasi yang tinggi dapat berupa data numerik seperti umur, jarak melahirkan dengan waktu bekerja dalam hitungan tahun, dan beberapa variabel numerik lainnya yang dihitung dalam angka tahun, contohnya seperti jarak bekerja dengan memiliki anak pertama adalah 5 tahun.

Setelah data siap diolah selanjutnya adalah proses analisis data menggunakan mesin pembelajaran. Metode mesin pembelajaran yang digunakan pada penelitian ini terbagi menjadi dua kategori yakni *supervised learning* dan *unsupervised learning*. Perbedaan metode tersebut dilakukan untuk melihat apakah ada kesamaan hasil dari dua kategori tersebut, karena data yang didapatkan telah memiliki label. Metode-metode yang digunakan juga telah dipakai dalam bidang kesehatan. Pada kategori *supervised learning*, subjek kasus dan kontrol sudah diberikan label, sedangkan pada *unsupervised learning*, label tersebut tidak diberikan. Tujuannya untuk mengetahui apakah mesin pembelajaran dapat membaca data sesuai dengan kategori yang telah dikelompokkan atau tidak. Selain

itu untuk melihat seberapa baik data apabila diklasterisasi mengunakan mesin pembelajaran. Berikut adalah penjelasan setiap metode yang digunakan.



Gambar 1. Visualisasi persebaran nilai data menggunakan *box* plot

1. *Supervised learning*

Supervised learning adalah salah satu metode dalam mesin pembelajaran di mana model dilatih menggunakan data yang sudah diberi label. Dalam konteks ini, "label" berarti data tersebut sudah memiliki output yang diketahui. Prosesnya melibatkan dua komponen utama: data *input* (*features*) dan label *output* (target). Tujuan dari *supervised learning* adalah untuk mempelajari hubungan atau pola antara input dan output sehingga model dapat memprediksi output yang benar untuk input yang belum pernah dilihat sebelumnya atau melakukan prediksi berdasarkan data yang pernah ada [10].

1.1 *Random forest*

Random Forest adalah salah satu algoritma yang digunakan untuk tugas klasifikasi dan regresi. Algoritma ini merupakan bentuk pengembangan dari *decision tree* (pohon keputusan), dan termasuk dalam teknik *ensemble learning*, yaitu metode yang menggabungkan beberapa model pembelajaran untuk mendapatkan hasil yang lebih baik [11]. Konsepnya adalah memiliki banyak pohon keputusan yang masing-masing memiliki hasil lalu mendapatkan akurasi dari seluruh hasil tersebut. Setiap pohon dibuat dari sampel acak data pelatihan, yang dimulai dari akar lalu bertahap membagi node (poin cabang) sampai mencapai batas tertentu. Node dibagi dari variabel-variabel yang dipilih secara acak dengan tujuan memilih variabel terbaik untuk mendapatkan aturan pemisahan terbaik. Parameter yang digunakan untuk mengukur hasil pembagian node tersebut adalah indeks Gini, yaitu ukuran yang digunakan pada statistik dan mesin pembelajaran untuk mengukur tingkat kemurnian suatu data. Pada konteks *random forest*, indeks Gini mengukur seberapa baik pohon keputusan memisahkan data ke dalam kelompok yang berbeda [12].

Alasan lain penggunaan *random forest* adalah fungsinya untuk menentukan fitur penting dengan cara menghitung perubahan akurasi dengan mengacak nilai variabel tertentu. Apabila perubahan nilai membuat penurunan akurasi yang

signifikan maka variabel tersebut dianggap berkontribusi besar pada model. Hal ini menjadi salah satu kelebihan *random forest* karena dapat menggabungkan informasi semua pohon keputusan untuk menghitung seberapa penting setiap variabel dalam memprediksi hasil, sehingga dapat mencari variabel apa yang berpengaruh terhadap faktor risiko sindrom Down dari data yang dianalisis [13].

1.2 *K-nearest Neighbors* (KNN)

K-Nearest Neighbor (KNN) adalah salah satu algoritma pembelajaran mesin yang sederhana namun kuat, digunakan baik untuk tugas klasifikasi maupun regresi. Algoritma ini termasuk dalam kategori *lazy learning* karena tidak melakukan generalisasi apa pun pada data pelatihan dan hanya menyimpan data tersebut untuk kemudian digunakan pada saat prediksi [14]. Inti dari KNN adalah memprediksi kelas atau nilai dari sebuah sampel berdasarkan kedekatan atau kemiripan dengan sejumlah *k* tetangga terdekat dalam ruang fitur. Hal ini dapat dilakukan dengan menghitung jarak Euclidean antara pusat kluster dan titik data uji. Selain itu dapat pula menghitung jarak data uji dengan batas kluster, bukan hanya dengan pusat kluster, dengan tujuan mengetahui penyebaran kluster pada berbagai dimensi ruang data. Kepadatan kluster juga dihitung dalam mencari kemiripan pada data. KNN kemudian mempertimbangkan faktor jarak dan kepadatan tersebut untuk meningkatkan akurasi prediksi. Dengan kata lain, tantangan dalam algoritma ini adalah pemilihan jumlah tetangga yang optimal untuk proses klasifikasi [15]. Algoritma KNN telah digunakan untuk berbagai model penelitian seperti dalam bidang kesehatan dalam melihat perbedaan dua kelompok kasus dan kontrol [16].

1.3 *Support Vector Machine* (SVM)

Support Vector Machine (SVM) adalah salah satu algoritma pembelajaran mesin yang digunakan untuk tugas klasifikasi dan regresi. Algoritma ini sangat efektif dalam ruang fitur tinggi dan dikenal karena kemampuannya untuk menghasilkan margin klasifikasi yang maksimal. Margin ini adalah jarak terdekat antara titik data mana pun dan *hyperplane* yang memisahkan kelas-kelas yang berbeda. Inti dari SVM adalah menemukan *hyperplane* terbaik yang memisahkan data ke dalam kelas-kelas berbeda. *Hyperplane* adalah sebuah garis dalam dua dimensi, atau lebih umum, sebuah subruang dalam dimensi lebih tinggi, yang memisahkan titik data berdasarkan kelasnya. Margin adalah jarak antara *hyperplane* dan titik data terdekat dari setiap kelas. Titik data yang paling dekat dengan *hyperplane* disebut *support vectors*. Mereka adalah elemen kunci dalam menentukan posisi dan orientasi *hyperplane*. SVM berusaha memaksimalkan margin antara *support vectors* dari dua kelas. Ketika data tidak dapat dipisahkan secara linear dalam ruang fitur asli, SVM menggunakan trik kernel untuk memetakan data ke dalam ruang fitur yang lebih tinggi di mana data menjadi lebih mudah dipisahkan secara linear. Fungsi kernel umum yang digunakan termasuk *linear kernel*, *polynomial kernel*, *radial basis function* (RBF) atau *gaussian kernel,* dan *sigmoid kernel* [17].

1.4 *Naive Bayes* (NB)

Naive Bayes adalah kelompok algoritma pembelajaran mesin berbasis teorema Bayes dengan asumsi independensi kuat antar fitur. Algoritma ini sering digunakan untuk klasifikasi teks dan berbagai tugas klasifikasi lainnya dan cocok dipakai untuk

data yang berjumlah kecil serta memerlukan model yang simpel. Meskipun asumsi independensinya sederhana dan mungkin tidak selalu valid, namun dapat memberikan hasil yang baik dalam berbagai aplikasi [18]Teorema Bayes menghubungkan probabilitas suatu peristiwa dengan probabilitas peristiwa lainnya, seperti menghitung besar kemungkinan sebuah data masuk ke dalam kategori tertentu. Istilah *naive* (naif) yang dipakai mempunyai maksud bahwa semua variabel atau atribut pada data dianggap saling independen, sehingga tidak mempengaruhi variabel lainnya. Meskipun tidak selalu tepat, namun algoritma ini dapat bekerja dengan baik pada berbagai macam kasus. Cara algoritma ini belajar dengan menentukan kelas atau target, mengumpulka petunjuk dari fitur (variabel) yang ada sebagai ciri-ciri dari target, lalu menghitung probabilitas. Probabilitas tersebut terbagi menjadi probabilitas kelas (*prior*), probabilitas fitur pada kelas (*likelihood*), dan probabilitas global (*evidence*). *Prior* adalah peluang awal suatu sampel akan masuk ke dalam kelas tertentu, *likelihood* adalah peluang melihat ciri-ciri tertentu jika mengetahui sampel termasuk ke dalam suatu kelas tertentu, dan *evidence* adalah peluang umum melihat ciri-ciri tertentu di seluruh data. Konsep ini sangat membantu karena dapat memberikan independensi pada tipe data numerik dan kategorik yang sering dipakai untuk prediksi persebaran penyakit [19].

2. *Unsupervised learning*

Unsupervised learning adalah metode dalam mesin pembelajaran yang hanya mempunyai input pada data untuk diproses. Tujuannya agar mesin pembelajaran dapat menganalisis data tanpa adanya campur tangan manusia untuk kemudian mengidentifikasi pola penting, membuat grup, dan mengeksplorasi hasil yang didapatkan. Fungsi *unsupervised learning* sangat luas pada penggunaan mesin pembelajaran, seperti untuk klusterisasi, deteksi anomali, reduksi dimensi, dan pembelajaran fitur [14].

2.1 *K-Means*

K-Means adalah salah satu algoritma klasterisasi yang paling sederhana dan paling populer dalam mesin pembelajaran. Algoritma ini digunakan untuk membagi kumpulan data yang tidak berlabel ke dalam sejumlah *k* klaster (kelompok) berdasarkan kesamaan di antara data tersebut. Setiap klaster ditentukan oleh pusat atau *centroid*, yang merupakan rata-rata dari semua titik data dalam klaster tersebut. *K-Means* berusaha untuk mempartisi data ke dalam *k* klaster sedemikian rupa sehingga setiap titik data termasuk dalam klaster dengan centroid terdekat [20]. Cara kerja algoritma ini adalah dengan menentukan terlebih dahulu jumlah klaster. Perbedaan jumlah kluster akan memberikan perbedaan hasil. Setelah menentukan *k* kluster, kemudian adalah menentukan titik pusat awal yang dimiliki oleh setiap kluster. Titik pusat berguna dalam penentuan jarak setiap titik data yang ada pada data ke titik pusat atau kluster terdekatnya. Kedekatan objek kepada objek lain dan kepada suatu kluster dihitung menggunakan jarak Euclidian. Jarak ini dihitung berdasarkan jarak koordinat sebuah data dengan titik data lainnya, dengan menggunakan akar kuadrat perbedaan antara koordinat x dan y dari kedua titik data tersebut [21]. Algoritma ini telah dipakai untuk berbagai keperluan, termasuk ke dalam bidang kesehatan untuk mencari klusterisasi pada balita yang mengalami gizi buruk dan stunting [21], [22].

2.2 *Artificial Neural Network* (ANN)

Artificial neural networks (ANN) atau jaringan saraf tiruan adalah model komputasi yang terinspirasi oleh cara kerja otak manusia. Mereka terdiri dari sejumlah unit yang disebut neuron atau node, yang dihubungkan satu sama lain dalam beberapa lapisan. Apabila neuron manusia terdiri dari empat bagian, maka algoritma ANN terdiri dari tiga lapisan atau *layer*, yaitu *input layer, hidden layer,* dan *output layer*. Cara kerjanya dengan mempelajari dan memberikan nilai atau bobot pada data dengan melalukan pemrosesan terdistribusi secara paralel, sehingga algoritma ini cocok untuk pekerjaan yang membutuhkan paralelisasi. Hal ini membuat ANN dapat mengenali pola serta membuat prediksi berdasarkan data yang diberikan dan cocok untuk digunakan dalam berbagai aplikasi seperti pengenalan gambar, pengenalan suara, dan prediksi pola [10] Algortima yang dipakai adalah m*ultilayer perceptron* (MLP), merupakan salah satu arsitektur ANN yang paling dasar dan umum digunakan dalam berbagai aplikasi. MLP juga bisa digabunkan dengan *principle component analysis* (PCA) yang berfungsi untuk mereduksi variabel-variabel yang ada pada data lalu dipakai nilai yang paling berpengaruh. Secara singkatnya metode kerja algoritma ini dengan menerima sinyal dari setiap lapisan yang ada. Node dari *hidden layer* menerima sinyal dari *input layer*, kemudian menghitung luaran dari fungsi aktivasinya, kemudian hasilnya dikirim kepada node pada *output layer.* Metode ini juga telah dipakai dalam bidang kesehatan seperti untuk mengukur tingkat kesembuhan dari pasien kanker [23]

Tujuan menggunakan beragam metode agar dapat mengetahui akurasi yang paling tinggi dari model data yang diberikan. Seluruh metode menggunakan data latih sebesar 80% dan data uji sebesar 20% dengan memanfaatkan kode yang tersedia pada *library scikit-learn* untuk bahasa pemrograman *python*. Berikut adalah tabel yang menjelaskan *library* yang digunakan pada penelitian dan diagram alir sederhana sebagai alur berpikir analisis data pada penelitian. Diagram alir memperlihatkan alur dari pembacaan awal data sampai menampilkan akurasi dari setiap metode yang digunakan.

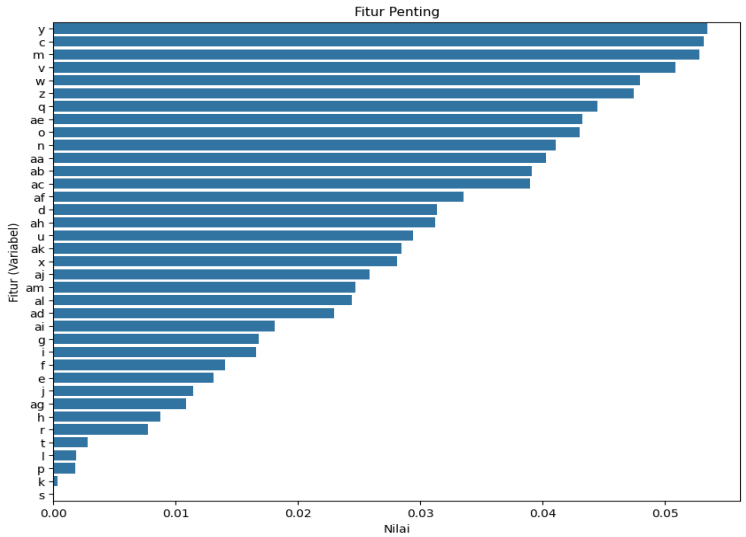
Tabel 2. *Library* yang dipakai pada metode mesin pembelajaran

Metode	**Nama *library***
Random Forest (RF)	*sklearn.ensemble import RandomForestClassifier* [24]
K-nearest Neighbors (KNN)	*sklearn.neighbors import KNeighborsClassifier* [25]
Support Vector Machine (SVM)	*sklearn.svm import SVC* [26]
Naive Bayes (NB)	*sklearn.naive_bayes import GaussianNB* [27]
K-Means	*sklearn.cluster import KMeans* [28]
Artificial Neural Network (ANN)	*sklearn.neural_network import MLPClassifier* [29]

Gambar 2. Diagram alir analisis data pada mesin pembelajaran

C. Hasil dan Pembahasan

Percobaan awal analisis adalah menggunakan seluruh data yang ada tanpa ada variasi data. Nilai akurasi yang didapatkan dari seluruh metode tidak ada yang mencapai 60%. Nilai akurasi yang paling tinggi didapat dari metode *random forest* sebesar 0,52 atau 52%. Meskipun begitu, seluruh metode menunjukkan secara konsisten variabel berpengaruh dan sesuai dengan landasan teori, seperti umur ibu dan umur ayah, lama bekerja orang tua, dan jarak bekerja dengan kelahiran subjek. Berikut adalah nilai fitur penting setiap variabel yang berhasil didapatkan menggunakan mesin pembelajaran.



Gambar 3. Nilai fitur penting dari mesin pembelajaran

Hasil tersebut memperlihatkan bahwa data mungkin harus diperbaiki sebelum dianalisis menggunakan algoritma mesin pembelajaran untuk meningkatkan nilai akurasi. Berdasarkan nilai fitur penting yang didapatkan, rentang nilai setiap variabel berkisar dari mendekati 0 untuk yang paling rendah sampai 0,05 untuk yang paling tinggi. Hasil ini menunjukkan bahwa tidak harus semua variabel diperhitungkan dalam pencarian faktor risiko. Variabel yang memiliki nilai paling tinggi perlu diprioritaskan untuk dianalisis agar mendapatkan akurasi yang maksimal. Beberapa variabel ini seperti usia anak (y), usia ayah (c), usia ibu (m), usia ibu saat mulai bekerja (v) sampai variabel jarak tahun ayah bekerja sampai melahirkan subjek (ac) yang memiliki nilai fitur penting di atas 0,03. Setelah mengetahui variabel yang lebih berpengaruh dari fitur penting, selanjutnya adalah menganalisis variabel tersebut untuk memaksimalkan akurasi dari setiap metode mesin pembelajaran terhadap faktor risiko sindrom Down.

Hal lain yang dilakukan untuk meningkatkan akurasi adalah mencoba membuat perbedaan kelompok dalam data yang dipakai. Kelompok yang dimasud berdasarkan subjek kasus-kontrol pada sumber data. Kelompok pertama dengan menggunakan seluruh subjek yaitu sindrom Down (SD), tuna grahita (TGH), dan tanpa kelainan khusus (TK). Kelompok kedua adalah SD dan TGH, sedangkan kelompok ketiga adalah antara SD dan TK. Setelah pembuatan tiga kelompok, langkah selanjutnya adalah membuat perbedaan pada variabel yang diuji. Ada tiga tipe jenis kelompok variabel yang dibuat, yang pertama adalah menggunakan seluruh variabel, yang kedua menggunakan variabel penting atau variabel dengan nilai fitur penting yang tinggi, sedangkan yang ketiga menggunakan variabel penting dan variabel yang memiliki nilai fitur penting yang rendah (variabel kurang penting). Namun, untuk kelompok ketiga yaitu SD dan TK memiliki satu variabel penting yang berbeda dengan kelompok lainnya setelah dianalisis, yaitu sumber air minum. Tujuan membuat tiga kelompok data adalah untuk melihat akurasi kinerja mesin pembelajaran dalam menganalisis perbedaan subjek kasus dan kontrol, sedangkan membuat sub kelompok berdasarkan jumlah variabel adalah untuk menentukan variabel yang sebaiknya digunakan untuk menganalisis data dalam meningkatkan akurasi. Berikut adalah tabel yang memperlihatkan variabel yang dipakai berdasarkan nilai dari fitur penting.

Tabel 3. Variabel penting dan kurang penting

Variabel penting	Variabel kurang penting
Usia	Jenjang pendidikan
Usia Ibu saat ini	Urutan kelahiran
Usia Ibu saat melahirkan responden	Jumlah Saudara Kandung
Usia Ibu saat lahir anak pertama	Jumlah Kakak
Usia Ibu saat mulai bekerja	Jumlah Adik
Jarak tahun Ibu bekerja dengan melahirkan subjek	Selisih umur dengan Kakak
Usia Ayah saat ini	Selisih umur dengan Adik
Usia Ayah saat responden lahir	Saudara dengan DS
Jarak tahun Ayah bekerja dengan subjek lahir (kelompok 1 dan 2)	Orang tua kerabat atau tidak
Sumber air minum (kelompok 3)	

Setelah data dikelompokan dan variabel yang digunakan telah divariasikan pada analisis mesin pembelajaran, hasil akurasi mengalami perubahan bergantung pada kelompok dan variabel yang digunakan. Hasil akurasi tersebut ditampilkan pada Tabel 4.

Tabel 4. Hasil akurasi pada penelitian

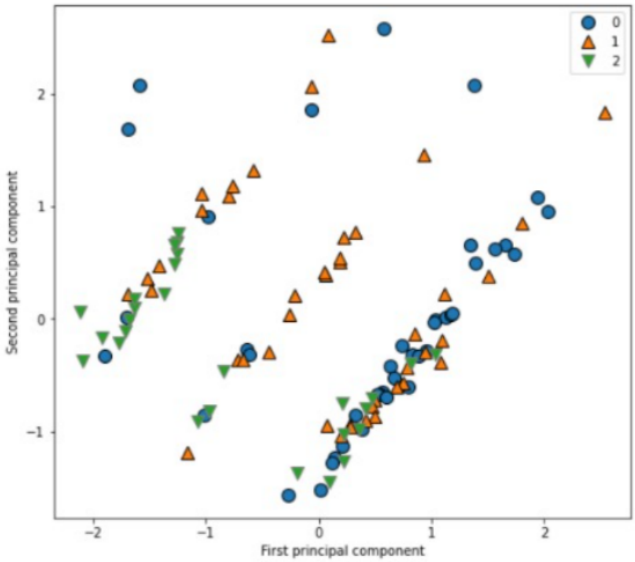
Kelompok Data		**Metode**					
Subjek*	**Variabel**	**RF**	**KNN**	**SVM**	**NB**	**K-Means**	**ANN**
Kelompok 1 (SD + TGH +TK)	Semua	0,52	0.26	0.44	0,43	0.09	0.35
	Penting	0,48	0,48	0,43	0,57	0,16	0,43
	Penting dan Kurang	0,61	0,43	0,52	0,43	0,21	0,48
Rata-rata		0,53	0,39	0,46	0,48	0,15	0,42
Kelompok 2 (SD + TGH)	Semua	0,53	0,41	0,59	0,59	0,09	0,53
	Penting	0,59	0,47	0,76	0,71	0,18	0,71
	Penting dan Kurang	0,68	0,53	0,65	0,59	0,24	0,59
Rata-rata		0,6	0,47	0,67	0,63	0,17	0,61
Kelompok 3 (SD + TK)	Semua	0.71	0,57	0,79	0,79	0,15	0,57
	Penting	0,71	0,79	0,71	0,86	0,28	0,71
	Penting dan Kurang	0,71	0,71	0,71	0,79	0,24	0,71
Rata-rata		0,71	0,69	0,74	0,81	0,22	0,66
Rata-rata total		**0,62**	**0,52**	**0,62**	**0,64**	**0,18**	**0,56**

*keterangan: SD (sindrom Down), TGH (tuna grahita), TK (tanpa kebutuhan khusus)

Random forest (RF) secara konsisten menunjukkan akurasi yang tinggi di berbagai kelompok dan jumlah variabel yang digunkan, menjadikannya metode yang mampu digunakan untuk menganalisis data kasus-kontrol dengan jumlah sampel yang sedikit. Namun, *random forest* dari segi akurasi tidak lebih tinggi dari metode *supervised* lainnya yang mendapatkan nilai di atas 75% ketika menganalisis data kelompok 3. Secara akurasi, *naive bayes* (NB) merupakan metode yang mendapatkan nilai paling tinggi dan satu-satunya yang berhasil mencapai angka di atas 80%, meskipun hanya untuk sekali pada analisis kelompok 3 menggunakan variabel penting saja. *Support vector machine* (SVM) dan *k-nearest neighbors* sama-sama memiliki akurasi yang cukup tinggi sampai 79% pada analisis kelompok 3 meskipun pada analisis di jenis variabel yang berbeda.

Metode yang paling tidak cocok digunakan adalah *k-means clustering* karena secara konsisten menunjukkan akurasi terendah, diikuti dengan metode *artificial neural network* (ANN). Meskipun begitu, rata-rata total ANN memiliki nilai yang sedikit lebih tinggi dari KNN, karena lebih baik dalam menganalisis data kelompok 1 dan kelompok 2, sehingga memperlihatkan keunggulan yang berbeda pada dua metode tersebut untuk kelompok data yang berbeda. Rendahnya nilai akurasi pada metode *unsupervised learning* dapat dijelaskan ketika melakukan visualisasi titik data untuk seluruh subjek data, yang memperlihatkan penumpukan di lokasi yang sama pada subjek yang berbeda. Penumpukan ini dapat dilihat ketika data dianalisis menggunakan seluruh subjek dan variabel dengan *principal component analysis*

(PCA). Teknik PCA dipakai karena dapat mempermudah klusterisasi hanya dengan menggunakan komponen atau variabel penting berdasarkan hasil reduksi dari seluruh variabel yang digunakan sehingga dapat mengubah data yang rumit menjadi bentuk yang lebih sederhana tanpa kehilangan terlalu banyak informasi penting. Metode ini menggunakan transformasi matematika yang disebut transformasi ortogonal untuk mengubah variabel-variabel yang saling berkorelasi (saling terkait) menjadi satu set variabel baru yang tidak berkorelasi satu sama lain yang disebut sebagai komponen utama [10]. Hasil tersebut dapat dilihat pada Gambar 4.



Gambar 4. Visualisasi klusterisasi data menggunakan PCA

Memasukkan variabel penting dan kurang penting secara umum meningkatkan kinerja model, namun ketika hanya menggunakan variabel penting maka akurasi dapat meningkat lebih baik. Hasil ini dapat mengindikasikan bahwa analisis pada sembarang variabel dapat membuat model bekerja kurang baik dalam memprediksi data. Hanya metode RF pada kelompok 3 yang nilai akurasinya sama pada seluruh sub kelompok variabel, yang mengindikasikan bahwa metode ini konsisten bekerja secara baik pada data dengan jumlah baris sedikit dengan banyak kelompok dan menggunakan variabel yang banyak. Hal ini dapat terjadi karena metode RF mengambil data dengan acak lalu menggabungkan seluruh nilai dari pohon keputusan secara independen, sehingga baik dalam mengatasi *overfitting*. Begitu pula dengan metode NB yang menghitung peluang antar variabel secara independen, yang memungkinkan model bekerja dengan baik untuk mendapatkan akurasi ketika hanya menganalisis variabel penting. Selain itu akan bekerja lebih baik jika ada perbedaan signifikan antara kelompok kasus dengan kontrol.

Perbedaan kelompok 2 dan kelompok 3 adalah kemiripinnya dengan kelompok kasus. Subjek TGH memiliki kemiripan yang lebih tinggi dengan subjek DS daripada subjek TK. Kemiripan tersebut karena subjek TGH disurvei dari Sekolah Kebutuhan Khusus yang sama, sedangkan subjek TK disurvei di luar lingkungan sekolah. Peningkatan akurasi ini juga terjadi secara signifikan pada seluruh metode lain. Dengan menganalisis dua kelompok berbeda secara signifikan dapat membuat

model mesin pembelajaran bekerja jauh lebih baik. Ketika model dapat memisahkan data kepada kelompok yang tepat, maka akurasi akan meningkat. Data yang memiliki kemiripan bekerja lebih baik pada model *supervised learning*, karena label output dapat membantu memisahkan data. Seperti pada metode *k-means* yang tidak menggunakan label kelas sehingga kurang cocok untuk tugas klasifikasi kasus-kontrol dan mendapatkan nilai akurasi paling rendah. Perbedaan kategori dan variabel tidak banyak mempengaruhi *K-Means* karena metode ini tidak dirancang untuk menangani tugas klasifikasi secara efektif, meskipun akurasinya masih tetap meningkat pada kelompok 3.

Metode ANN yang masuk dalam kategori *unsupervised learning* dapat bekerja lebih daripada *k-means* karena menggunakan fungsi aktivasi untuk memodelkan hubungan kompleks dalam data. Namun, akurasinya belum sebaik *supervised learning* karena memerlukan tuning dan jumlah data yang besar untuk mencapai kinerja optimal. Meskipun begitu, metode ANN dapat dapat menganalisis kelompok 3 secara lebih baik khususnya ketika hanya memakai sub kelompok variabel penting dan kurang penting. Hal ini karena variabel penting membuat model mampu mempelajari fitur-fitur yang relevan lebih efektif sehingga dapat meningkatkan akurasi. Alasan ini juga berlaku kepada metode KNN yang mendapatkan nilai akurasi paling rendah pada semua kelompok dalam kategori *supervised learning*. Metode KNN yang menggunakan jarak Euclidian seperti *k-means* cenderung tidak bisa membedakan data ketika menumpuk pada lokasi yang sama, meskipun merupakan subjek yang berbeda. Hal ini terlihat dari rendahnya akurasi KNN ketika menganalisis data kelompok 1 dan 2, namun meningkat tajam ketika menanalisis kelompok 3 dengan sub kelompok variabel yang penting dan kurang penting.

Secara keseluruhan, perbedaan akurasi antara metode-metode ini sebagian besar disebabkan oleh cara masing-masing metode memproses dan memanfaatkan informasi dari variabel-variabel yang diberikan. Metode yang lebih kompleks seperti RF, SVM, dan NB menunjukkan kinerja yang baik pada dataset menggunakan seluruh variabel karena ada metode independensinya yang memproses informasi relevan secara lebih efektif. Metode yang kurang cocok dipakai adalah *k-means* dan KNN, karena menggunakan perhitungan jarak dalam analisisnya sehingga tidak mampu untuk klasifikasi secara optimal pada data dengan subjek mirip. Meskipun begitu, seluruh metode bekerja dengan baik ketika menganalisis kelompok 3, yang memperlihatkan bahwa akurasi dapat meningkat apabila kelompok subjek berbeda secara signifikan. Selain itu, mayoritas metode bekerja lebih baik ketika hanya menggunakan variabel penting, memperlihatkan bahwa jumlah variabel dan tingkat nilai pentingnya berpengaruh dalam meningkatkan akurasi model.

D. Simpulan

Penelitian ini dapat menentukan variabel penting yang berpengaruh terhadap sindrom Down menggunakan analisis mesin pembelajaran sesuai dengan landasan teori. Faktor risiko yang berpengaruh adalah usia ibu, usia ayah, dan rentang waktu orang tua bekerja sampai anaknya lahir, dan sumber air minum (pada kelompok 3). Seluruh faktor risiko tersebut dilihat dari nilai fitur penting, yang selalu mendapatkan nilai yang tinggi pada seluruh model. Hasil akurasi paling tinggi yang didapat pada penelitian sebesar 86% dengan metode *naïve bayes* (NB) pada analisis kelompok 3 dengan menggunakan variabel penting. Data yang menghasilkan

akurasi paling tinggi adalah kelompok 3, yang menganalisis subjek sindrom Down (SD) menjadi subjek kasus dan subjek tanpa kebutuhan khusus (TK) menjadi subjek kontrol. Metode yang bekerja secara konsisten pada seluruh kelompok adalah *random forest* (RF) dan metode yang mendapatkan akurasi paling rendah adalah *k-means.* Secara umum, kategori *supervised learning* bekerja lebih baik dibanding *unsupervised learning* pada klasifikasi data kasus-kontrol.

E. Ucapan Terima Kasih

Ucapan terima kasih kepada Sekolah Kebutuhan Khusus yang telah bekerja sama untuk pengumpulan data penelitian serta pemerintah Kota dan Kabupaten Tangerang yang telah memberikan izin penelitian. Sekolah Kebutuhan Khusus tersebut adalah SKN Negeri 01 Kab. Tangerang, SKH YKDW 01, SKH Pelangi Anakku, SKH Darmawati Arief, SKH Syahida Harapan Bunda, dan SKH Nurayan 01. Kepada Departemen Kesehatan Lingkungan Fakultas Kesehatan Masyarakat Universitas Indonesia dan Program Studi Sains Komputasi Fakultas Matematika dan Ilmu Pengetahuan Alam Institut Teknologi Bandung yang telah memberikan panduan dalam pengambilan dan pengolahan data.

F. Referensi

[1] P. Halder *et al.*, "Genetic aetiology of Down syndrome birth: novel variants of maternal DNMT3B and RFC1 genes increase risk of meiosis II nondisjunction in the oocyte," *Molecular Genetics and Genomics*, vol. 298, no. 1, pp. 293–313, 2023, doi: 10.1007/s00438-022-01981-4.

[2] H. Cuckle and J. Morris, "Maternal age in the epidemiology of common autosomal trisomies," *Prenat Diagn*, vol. 41, no. 5, pp. 573–583, 2021, doi: 10.1002/pd.5840.

[3] F. Tong, "Breaking Down: a critical discourse analysis of John Langdon Down's (1866) classification of people with trisomy 21 (Down syndrome)," *Critical Discourse Studies*, vol. 19, no. 6, pp. 648–666, 2022, doi: 10.1080/17405904.2021.1933117.

[4] S. Tamaki *et al.*, "Improving survival in patients with trisomy 18," *Am J Med Genet A*, vol. 188, no. 4, pp. 1048–1055, 2022, doi: 10.1002/ajmg.a.62605.

[5] C. Mitsuhata, N. Kado, M. Hamada, R. Nomura, and K. Kozai, "Characterization of the unique oral microbiome of children with Down syndrome," *Sci Rep*, vol. 12, no. 1, pp. 1–12, 2022, doi: 10.1038/s41598-022-18409-z.

[6] S. E. Antonarakis *et al.*, "Down Syndrome," *Nat Rev Dis Primers*, vol. 6, no. 1, pp. 1–43, 2020, doi: 10.1038/s41572-019-0143-7.Down.

[7] O. O. Palit and B. Hartono, "Studi Pendahuluan Faktor Risiko Sindrom Down," *Jurnal Nasional Kesehatan Lingkungan Global*, vol. 2, no. 1, pp. 1–9, 2021.

[8] B. Karami Matin, A. Kazemi Karyani, S. Soltani, S. Akbari, S. Toloui Rakhshan, and M. M. Moghadam, "The Relationship Between Health System Functions and the Prevalence of Down Syndrome on a Global Scale," *J Rehabil*, vol. 23, no. 2, pp. 186–203, 2022, doi: 10.32598/rj.23.2.1719.8.

[9] A. Chahal and P. Gulia, "Machine learning and deep learning," *International Journal of Innovative Technology and Exploring Engineering*, vol. 8, no. 12, pp. 4910–4914, 2019, doi: 10.35940/ijitee.L3550.1081219.

[10] J. Alzubi, A. Nayyar, and A. Kumar, "Machine Learning from Theory to Algorithms: An Overview," *J Phys Conf Ser*, vol. 1142, no. 1, pp. 0–15, 2018, doi: 10.1088/1742-6596/1142/1/012012.

[11] G. Louppe, "Understanding Random Forests: From Theory to Practice," no. July, 2014, [Online]. Available: http://arxiv.org/abs/1407.7502

[12] J. Hu and S. Szymczak, "Evaluation of network-guided random forest for disease gene discovery," pp. 1–23, 2023, [Online]. Available: http://arxiv.org/abs/2308.01323

[13] S. Song, R. He, Z. Shi, and W. Zhang, "Variable importance measure system based on advanced random forest," *CMES - Computer Modeling in Engineering and Sciences*, vol. 128, no. 1, pp. 65–85, 2021, doi: 10.32604/cmes.2021.015378.

[14] I. H. Sarker, "Machine Learning: Algorithms, Real-World Applications and Research Directions," *SN Comput Sci*, vol. 2, p. 160, 2021, doi: https://doi.org/10.1007/s42979-021-00592-x.

[15] H. Saadatfar, S. Khosravi, J. H. Joloudari, A. Mosavi, and S. Shamshirband, "A new k-nearest neighbors classifier for big data based on efficient data pruning," *Mathematics*, vol. 8, no. 2, pp. 1–12, 2020, doi: 10.3390/math8020286.

[16] A. Musthofa and M. Rahardi, "Perbandingan Algoritma Support Vector Machine dan K-Nearest Neighbors Pada Sinyal Tubuh Perokok," *Indonesian Journal of Computer Science*, vol. 12, no. 6, pp. 3746–3756, 2023, doi: 10.33022/ijcs.v12i6.3290.

[17] M. A. Hearst, "Support Vector Machines," 1998. [Online]. Available: https://www.researchgate.net/publication/3420408_Support_vector_machines

[18] V. Nakhipova *et al.*, "Use of the Naive Bayes Classifier Algorithm in Machine Learning for Student Performance Prediction," *International Journal of Information and Education Technology*, vol. 14, no. 1, pp. 92–98, 2024, doi: 10.18178/ijiet.2024.14.1.2028.

[19] Rayuwati, Husna Gemasih, and Irma Nizar, "Implementasi Algoritma Naive Bayes Untuk Memprediksi Tingkat Penyebaran Covid-19 Di Indonesia," *Jural Riset Rumpun Ilmu Teknik*, vol. 1, no. 1, pp. 38–46, 2022, doi: 10.55606/jurritek.v1i1.127.

[20] J. Pérez-Ortega, N. Nely Almanza-Ortega, A. Vega-Villalobos, R. Pazos-Rangel, C. Zavala-Díaz, and A. Martínez-Rebollar, " The K -Means Algorithm Evolution ," *Introduction to Data Science and Machine Learning*, 2020, doi: 10.5772/intechopen.85447.

[21] M. D. Chandra, E. Irawan, I. S. Saragih, A. P. Windarto, and D. Suhendro, "Penerapan Algoritma K-Means dalam Mengelompokkan Balita yang Mengalami Gizi Buruk Menurut Provinsi," *BIOS : Jurnal Teknologi Informasi dan Rekayasa Komputer*, vol. 2, no. 1, pp. 30–38, 2021, doi: 10.37148/bios.v2i1.19.

[22] P. Apriyani, A. R. Dikananda, and I. Ali, "Penerapan Algoritma K-Means dalam Klasterisasi Kasus Stunting Balita Desa Tegalwangi," *Hello World Jurnal Ilmu Komputer*, vol. 2, no. 1, pp. 20–33, 2023, doi: 10.56211/helloworld.v2i1.230.

[23] R. Syah, S. Wulandari, Arbansyah, and A. Rezaeipanah, "Design of ensemble classifier model based on mlp neural network for breast cancer diagnosis," *Inteligencia Artificial*, vol. 24, no. 67, pp. 147–156, 2021, doi: 10.4114/intartif.vol24iss67pp147-156.

[24] Lucy Liu and Jérémie du Boisberranger, "RandomForestClassifier." [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

[25] Y. Xiao and Thomas J. Fan, "KNeighborsClassifier." [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html

[26] Bharat Raghunathan, "SVC." [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html

[27] Adrin Jalali, Yao Xiao, Thomas J. Fan, Guillaume Lemaitre, and Jérémie du Boisberranger, "GaussianNB." [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.GaussianNB.html

[28] J. du Boisberranger, "k_means." [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.k_means.html

[29] A. Jalali, "MLPClassifier." [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html