Indonesian Journal of Computer Science
ISSN 2302-4364 (*print*) dan 2549-7286 (*online*)
Jln. Khatib Sulaiman Dalam, No. 1, Padang, Indonesia, Telp. (0751) 7056199, 7058325
Website: ijcs.stmikindonesia.ac.id | E-mail: ijcs@stmikindonesia.ac.id

Meningkatkan Kinerja Decision Tree C4.5 Dengan Seleksi Fitur Korelasi Pearson Pada Deteksi Penyakit Diabetes

Mohammad Burhan Hanif[1], Galet Guntoro Setiaji[2]
hanifburhan@usm.ac.id, gallet@usm.ac.id
Universitas Semarang

Informasi Artikel

Diterima : 20 Aug 2022
Direview : 25 Aug 2022
Disetujui : 30 Aug 2022

Kata Kunci

Algoritma C4.5, Seleksi Fitur, Korelasi Pearson, Diabetes

Abstrak

Diabetes sebuah penyakit yang menjadi momok seluruh dunia. Kerugianya tidak hanya pada penderita sendiri tetapi juga merambah ke banyak sektor. Baik di sektor pelayanan kesehatan dan sektor financial yang sangat menjadi beban tinggi yang perlu ditangani dengan baik dengan jalan pendeteksian penyakit diabetes sejak dini. Salah satu pendeteksian dini penyakit diabetes dapat memanfaatkan algoritma machine learning pada bidang data mining. Algoritma C4.5 merupakan algoritma machine learning yang memiliki tingkat akurasi dan kecepatan perhitungan tinggi dalam klasifikasi. Namun demikian algoritma C4.5 terganggu dengan data tak seimbang dan fitur data berdimensi tinggi. Pemanfaatan seleksi fitur menjadi salah satu penyelesain masalah data berdimensi tinggi. Algoritma Korelasi Pearson memiliki kemampuan dalam mengukur informasi antar fitur dan diterapkan dalam penelitian ini. Penggunaan Korelasi Pearson dianggap berhasil dalam meningkatkan kinerja algoritma C4.5 dalam deteksi awal penyakit diabetes. Keberhasilan ini terlihat pada hasil akurasi sebesar 95.31% tanpa korelasi pearson menjadi 96.16% dengan pemanfaatan korelasi pearson.

Keywords

C4.5 Algorithm, Feature Selection, Pearson Correlation, Diabetes.

Abstrak

Diabetes is a disease that is a scourge of the whole world. The loss is not only in the sufferers themselves but also penetrates into many sectors. Both in the health service sector and the financial sector, which is a high burden that needs to be handled properly by detecting diabetes early. One of the early detection of diabetes can utilize machine learning algorithms in the field of data mining. The C4.5 algorithm is a machine learning algorithm that has a high level of accuracy and calculation speed in classification. However, the C4.5 algorithm is compromised with unbalanced data and high-dimensional data features. The use of feature selection is one of the solutions to high-dimensional data problems.Pearson's Correlation Algorithm has the ability to measure information between features and is applied in this study. The use of Pearson Correlation is considered successful in improving the performance of the C4.5 algorithm in the early detection of diabetic diseases. This success was seen in the accuracy results of 95.31% without pearson correlation to 96.16% with pearson correlation utilization.

A. Pendahuluan

Diabetes penyakit yang terjadi ketika insulin dalam pankreas tubuh manusia mengalami penurunan dibawah batas rendah minimal. Pengurangan ini dikarenakan sel – sel penghasil insulin telah hancur atau rusak. Gejala umum penyakit ini biasanya dapat dilihat pada menurunnya berat badan, haus yang berlebihan, polyuria, kaburnya penglihatan dan sebagainya [1].

Diabetes sebuah momok penyakit menakutkan yang menjadi masalah dunia karena cakupan sebaran penyakit sudah hampir merata di dunia. Tidak hanya merugikan kesehatan tubuh manusia saja hal ini juga berdampak pada beban keuangan pelayanan kesehatan negara. Dimana pada tahun 2019 tercatat di asia timur dan asia tenggara menghabiskan USD 162 Milliar hanya unutuk penanganan pelayanan kesehatan penyakit diabetes [2]. Oleh karena itu perlu adanya deteksi penyakit yang akurat sebagai analisa primer yang dapat membantu meringankan pekerja medis dalam deteksi awal penyakit diabestes yang nantinya juga akan berimbas pada penurunan beban keuangan pelayanan medis [3].

Penelitan terdahulu tentang penyakit diabetes telah banyak dilaksanakan salah satunya oleh shin jye lee et all. Penelitianya berkaitan dengan algortima C4.5 yang akan digunakan untuk melakukan deteksi awal penyakit diabetes. Namun algortima C4.5 ini mengalami gangguan terhadap data penyakit diabetes yang berdimensi tinggi. Untuk mengatasi hal tersebut peneliti menggunakan metode seleksi fitur hingga mendapat peningkatan akurasi yang cukup signifikan [4].

Penelitan lainya tentang penyakit diabetes juga dilakukan oleh zhaozhao xu et all. Penelitiannya berkaitan dengan data medis salah satunya data penyakit diabetes yang akan di klasifikasi menggunakan algoritma C4.5 yang terkenal dengan ketepatan akursi klasifikasi yang tinggi. Akan tetapi algoritma C4.5 ini memiliki kendala jika berhadapan dengan data tidak seimbang pada dataset diabetes [5].

Penyakit diabetes juga menjadi topik penelitian dari nazin ahmed et all. Menggunakan data diabetes para peneliti ini memanfaatkan metode machine learning salah satu nya menggunakan algortima decision tree C4.5 untuk melakukan deteksi penyakit diabetes. Untuk meningkatkan kinerja algoritma ini digunakan rangkaian tahapan pre-processing yang efesien yaitu melaui pendekatan metode seleksi fitur. Dengan menggunakan pendekatan seleksi fitur dihasilkan peningkatan akurasi dari 2.71% menjadi 13.13% [6].

Selanjutnya deteksi awal penyakit diabetes dilakukan oleh gaurav tripathi dan rakesh kumar. Penelitiannya berkisar tentang penggunaan machine learning untuk deteksi awal penyakit diabetes dengan data yang diambil dari pima indian dataset. Dari berbagai macam algoritma clasifikasi di dapatkan hasil akurasi yang cukup tinggi sebesar 87.66% [7].

Peneliti lain yang meneliti tentang diabetes yaitu d vigneswati et all. Didalam penelitianya tentang prediksi penyakit diabetes digunakan metode machine learning untuk mendapatkan hasil perhitungan yang terbaik. Metode mechine learning yang dipakai salah satunya menggunakan algortima C4.5 yang menghasilkan nilai akurasi sebesar 76.25 % [8].

Algoritma pohon keputusam C4.5 sudah terkenal dengan kelebihanya yang memiliki akurasi yang tinggi dalam hal klasifikasi, juga memiliki kecepatan dalam hal perhitungan. Walaupun demikian algoritma pohon keputusan C4.5 ini memiliki

kelemahan pada data tidak seimbang dan berdimensi tinggi yang memiliki fitur banyak sehingga dapat menurunkan kinerja dari algoritma ini [9]. Untuk mengurai fitur fitur penting dan relevan pemilihan fitur adalah cara yang bisa dipilih dalam meningkatkan kinerja algoritma pohon keputusan C4.5 [10].

Algoritma yang di manfatkan untuk mengukur informasi antar fitur dan fitur dengan label nya salah satunya adalah korelasi pearson. Korelasi pearson hadir juga secara efesien mampu mengatisi fitur bertipe campuran [11]. Dengan demikian algoritma korelasi pearson dapat dimanfaatkan untuk pemilihan fitur data diabetes yang nantinya akan meningkatkan kinerja dari algoritma c4.5.

B. Metode Penelitian

Tahapan penelitian ini memiliki beberapa langkah mulai dari proses pengumpulan data. Dilanjutkan dengan proses seleksi fitur terbaik menggunakan algoritma korelasi pearson. Lalu proses selanjutnya menggunakan algortima pohon keputusan C4.5 untuk klasifikasi data penyakit diabetes. Dan terakhir mengukur evaluasi kinerja algoritma tersebut. Langkah – langkah yang dilakukan dalam penelitian ini akan dijelaskan sebagai berikut :

1. Pengumpulan Data

Data yang digunakan pada penelitian ini adalah data sekunder yaitu data publik pasien penyakit diabetes hospital banglades. Data sekunder penyakit diabetes ini diambil dari UCI machine learning repositori datasate [12]. Dataset ini memiliki 17 atribut dengan 1 label dan memiliki 520 baris data. Attribut yang age, sex, polyuria, polydipsia, sudden weight loss, weakness, polyphagia, genital thrush, visual blurring, itching, irritability, delayed healing, partial paresis, muscle stiffness, alopecia, dan obesity. Sedangkan label nya adalah class positive atau negative.

2. Metode yang diusulkan

Didalam metode penelitian yang diusulkan ini peneliti menggunakan algortima klasifikasi untuk mengatasi deteksi penyakit diabetes secara awal tanpa pasien harus bersusah payah antri sangat lama di tempat pelayanan keehatan. Model klasifikasi berdasarkan pada metode algoritma decesion tree c4.5. Untuk mendapatkan nilai akurasi terbaik maka algoritma C4.5 akan dioptimalkan sehingga dapat meningkatkan kinerja dengan baik.

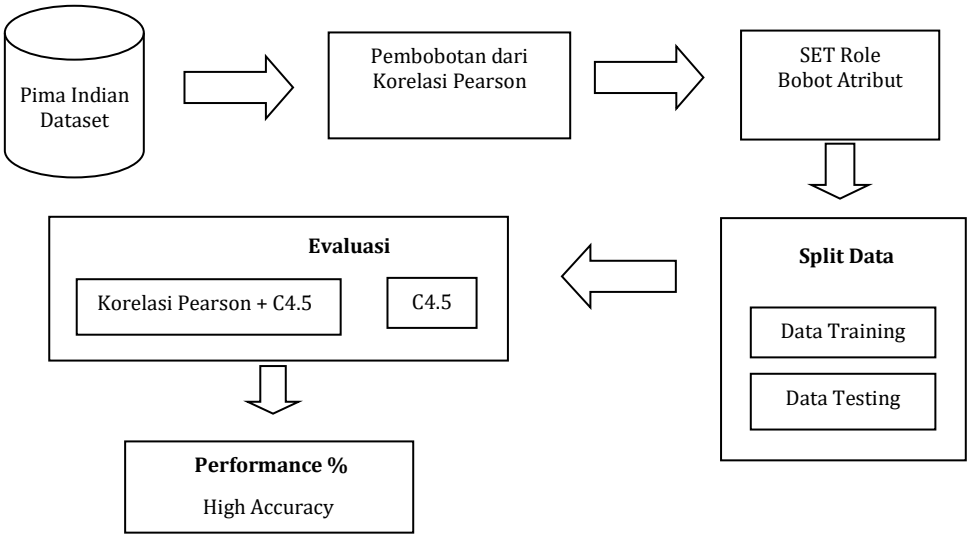
Langkah awal yang akan dilakukan pada percobaan kali ini adalah melakukan pre-proceesing data. Didalam tahapan pre-processing ini akan di lakukan langkah pertama yaitu pelabelan atribut dimana penetapan atau pemilihan atribut akan di pilih. Tahapan kedua yaitu menetapkan kelas label yaitu pada atribut bernama outcome 0 atau 1 dimana 0 adalah tidak mengidap diabetes dan 1 mengidap diabetes.

Tahapan ketiga dalam proses pre-processing selanjutnya adalah menerapkan dan tanpa menerapkan algoritma Korelasi Pearson sebagai pembanding pengaruh terhadap algoritma klasifikasi C4.5. Algoritma Korelasi Pearson digunakan untuk menentukan bobot pada atribut atribut penting. Setelahnya ditentukan pemilihan bobot atribut penting berdasarkan besaran bobot dari bobot terbesar hingga terkecil yang telah dihasilkan oleh algoritma korelasi pearson tersebut. Hal ini

didasarkan pada matriks kovarians data untuk mengevaluasi kekuatan hubungan antara atribut satu dengan atribut lainya [13].

Tahapan selanjutnya adalah proses split data menentukan data latih dan data uji. Perbandingan data latih dan data uji yang di pakai yaitu sebanya 90% untuk besaran data latih dan sebanyak 10% yang akan digunakan untuk data uji. Setelah didapati data latih dan data uji langkah selanjutnya yaitu melakukan klasifikasi menggunakan algoritma decission tree C4.5.

Didalam tahapan akhir pada penelitian ini yaitu berupa perbandingan antara tingkat akurasi dari model yang diusulkan. Model yang di maksud yaitu model antara sebelum di tingkatkan dengan model yang sudah di ditingkatkan. Hal ini di harapkan dapat memperoleh gambaran tingkat keberhasilan yang dicapai. Kerangka kerja tahapan pemodelan yang di usulkan ditunjukan pada gambar 1.



Gambar 1. Kerangka Model kerja yang diusulkan

3. Pengujian Model

Dalam pengujian model ini akan dilakukan optimasi terhadap model C4.5 melalui fitur seleksi metode korelasi pearson. Proses pengujian kinerja dilihat dari sejauhmana hasil akurasi dari klasifikasi yang diperoleh. Tahapan akhir yaitu tentang uji prediksi pada data testing.

4. Evaluasi dan Validasi

Tahapan validasi merupakan tahapan yang sangat penting dalam pemodelan metode karena dari hasil inilah dapat dilihat sejauh mana kehandalan model yang akan digunakan dalam pengambilan keputusan [14]. Confusion matrix untuk melihat evaluasi dan validasi hasil dari klasifikasi metode yang mengahasilkan performa akurasi, presisi, dan recall terbaik [15].

C. Hasil dan Pembahasan

Hasil dan pembahasan pada penelitian ini merupakan runtutan dari metode peneliat dimaulai dari :

1. Data penelitian

Kumpulan data yang digunakan sebanyak 520 baris data penyakit diabetes banglades yang akan di bagi menjadi data training dan data testing yang nampak pade tabel 1 dan tabel 2 sebagai berikut :

Tabel 1. Data Training

No	Age	Gender	Polyuria	Polydipsia	.....	Obesity	class
1	40	Male	No	Yes	....	Yes	Positive
2	58	Male	No	No	....	No	Positive
3	41	Male	Yes	No	....	No	Positive
5	45	Male	No	No	....	No	Positive
6	60	Male	Yes	Yes	....	Yes	Positive
7	55	Male	Yes	Yes	....	Yes	Positive
8	57	Male	Yes	Yes	....	No	Positive
9	66	Male	Yes	Yes	....	No	Positive
10	67	Male	Yes	Yes	....	Yes	Positive
....	....	....	....	....	....	....	....
....	....	....	....	....	....	....	....
398	46	Male	No	No	....	No	Negative
399	53	Male	No	No	....	No	Negative
400	64	Male	No	No	....	No	Negative

Tabel 2. Data Testing

No	Age	Gender	Polyuria	Polydipsia	.....	Obesity	class
1	44	Male	Yes	No	....	Yes	Negative
2	36	Male	No	No	....	No	Negative
3	43	Male	No	No	....	No	Negative
5	53	Male	No	No	....	No	Negative
6	47	Male	No	No	....	Yes	Negative
7	58	Male	No	Yes	....	No	Negative
8	56	Male	No	No	....	No	Negative
9	51	Female	No	No	....	Yes	Negative
10	59	Female	No	No	....	No	Negative
....	....	....	....	....	....	....	....
....	....	....	....	....	....	....	....
98	58	Female	No	No	....	No	Negative
99	32	Female	No	No	....	No	Negative
100	42	Male	No	No	....	No	Negative

2. Fitur seleksi

Pada bagian ini perhitungan metode korelasi pearson akan digunankan untuk mengeleminasi fitur yang tidak penting. Penentuan fitur penting didapatkan dari hasil bobot yang di hasilkan oleh metode ini sperti terlihat pada tabel 3 berikut ini :

Tabel 3. Bobot Fitur

No	Fitur	Bobot
1	Polyuria	0.666
2	Polydipsia	0.649
3	Gender	0.449
4	sudden weight loss	0.437
5	partial paresis	0.432
6	Polyphagia	0.343
7	Irritability	0.299
8	Alopecia	0.268
9	visual blurring	0.251
10	weakness	0.243
11	muscle stiffness.	0.122
12	Genital thrush	0.110
13	Age	0.109
14	Obesity	0.072
15	Delayed healing	0.047
16	Itching	0.013

Setelah bobot fitur didapatkan menggunakan korelasi pearson maka yang perlu di kerjakan adalah mengurutkan nilai bobot dari bobot tertinggi ke bobot terendah. Kemudian tentukan tingkat kepentingan ambang batas dari masing-masing atribut ini. Untuk selanjutnya, atribut yang memiliki bobot kepentingan yang sama dengan ambang batas atau lebih besar akan tetap digunakan atau dipertahankan, namun untuk atribut yang memiliki tingkat kepentingan atau nilai bobot yang lebih kecil atau di bawah nilai ambang batas, mereka akan diabaikan atau tidak akan digunakan dalam proses perhitungan selanjutnya.

Penentuan ambang batas dilakukan dengan mengerjakan percoban berkali kali hingga didapatkan hasil ambang batas yang di pakai. Hasil pengujian untuk mencari ambang batas dapat dilihat pada tabel 4 berikut ini :

Tabel 4. Bobot Fitur

No	Fitur	persentase Bobot
1	16	95.31%
2	15	96.16%
3	14	94.66%
4	13	94.44%
5	12	95.09%
6	11	93.59%

Dari data eksperimen di atas, akurasi tertinggi terletak pada jumlah fitur sebanyak 15 fitur dengan nilai akurasi 96.16%. Sehingga ambang batas yang akan digunakan adalah ambang atribut ke 15, yang bernilai 0.047.

Setelah nilai ambang batas di dapatkan selanjutnya adalah melakukan fitur seleksi dengan acuan ambang batas dari hasil perhitungan korelasi pearson. Nilai ambang batas yang di dapat yaitu sebesar 0.047, maka fitur yang mempunyai bobot kurang dari nilai tersebut tidak akan digunakan. Sehingga fitur yang di pakai setelah melalui proses seleksi fitur nampak pada tabel 5 berikut :

Tabel 5. Bobot Fitur

No	Fitur	Bobot
1	Polyuria	0.666
2	Polydipsia	0.649
3	Gender	0.449
4	sudden weight loss	0.437
5	partial paresis	0.432
6	Polyphagia	0.343
7	Irritability	0.299
8	Alopecia	0.268
9	visual blurring	0.251
10	weakness	0.243
11	muscle stiffness.	0.122
12	Genital thrush	0.110
13	Age	0.109
14	Obesity	0.072
15	Delayed healing	0.047

3. Evaluasi dan Validasi

Evaluasi dan Validasi digunakan untuk mengukur seberapa jauh kinerja model yang dihasilkan. Penelitian ini akan menghitung sejauh mana tingkat prediksi menggunakan model validasi. Dengan memahami kinerja model yang dipakai dapat membantu mengoptimalkan parameter dan memilih algoritma yang optimal. Cross-Validation adalah salah satu dari sekian banyak model validasi yang sering banyak dipakai oleh para peneliti maka pada penelitian ini juga akan menggunakan model tersebut. Teknik ini digunakan untuk memvalidasi banyak data subset pelatihan dan pengujian berulang kali. Setiap iterasi menguji data subset dengan data sisa sebagai data pelatihan.

Hasil dari model validasi cross-validation disajikan dalam tabel confusion matrix. Matriks ini berisi hasil prediksi dari sistem dan kondisi aktual. Tabel ini juga berisi nilai akurasi, presisi, penarikan kembali, dan UAC. Hasil pengukuran confusion matrix tanpa ada fitur seleksi korelasi pearson disajikan pada tabel 6 berikut ini:

Tabel 6. Confusion Matrix C4.5

		Actual Class	
		+ (Positive)	**- (Negative)**
Prediction class	+ (positive)	270	4
	- (negative)	18	176

Sedangkan untuk hasil pengukuran confusion matrix dari algoritma Korelasi Pearson + C4.5 nampak pada tabel 7 berikut ini :

Tabel 7. Confusion Matrix C4.5 + Korelasi Pearson

		Actual Class	
		+ (Positive)	**- (Negative)**
Prediction class	+ (positive)	274	4
	- (negative)	14	176

Dari tabel Confusion matrix itu menghasilkan nilai akurasi, presisi, recall dan AUC dengan hasil sebagai berikut:

Akurasi

Berikut ini adalah nilai akurasi antara algoritma decision tree C4.5 saja dan metode algoritma decision tree C4.5 + Korelasi Pearson

a. Akurasi dari algoritma decision tree C4.5 saja

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} = \frac{270 + 176}{270 + 176 + 4 + 18} * 100 = 95.31$$

b. Akurasi dari algoritma decision tree C4.5 + Korelasi Pearson

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} = \frac{274 + 176}{274 + 176 + 4 + 14} * 100 = 96.16$$

Presisi

Berikut ini adalah nilai presisi antara algoritma decision tree C4.5 saja dan metode algoritma decision tree C4.5 + Korelasi Pearson

a. Presisi dari algoritma decision tree C4.5 saja

$$Precision = \frac{TP}{TP + FP} = \frac{270}{270 + 4} = 0.9101$$

b. Presisi dari algoritma decision tree C4.5 + Korelasi Pearson

$$Precision = \frac{TP}{TP + FP} = \frac{274}{274 + 4} = 0.9296$$

Recall

Berikut ini adalah nilai recall antara algoritma decision tree C4.5 saja dan metode algoritma decision tree C4.5 + Korelasi Pearson

a. Recall dari algoritma decision tree C4.5 saja

$$Recall = \frac{TP}{TP + FN} = \frac{270}{270 + 18} = 0.9778$$

b. Recall dari algoritma decision tree C4.5 + Korelasi Pearson

$$Recall = \frac{TP}{TP + FN} = \frac{274}{274 + 14} = 0.9778$$

AUC (Area Under Curve)

Berikut ini adalah nilai AUC antara algoritma decision tree C4.5 saja dan metode algoritma decision tree C4.5 + Korelasi Pearson

a. AUC dari algoritma decision tree C4.5 saja

$$AUC = \frac{1 + TPrate - FPrate}{2} = 0.936$$

b. AUC dari algoritma decision tree C4.5 + Korelasi Pearson

$$AUC = \frac{1 + TPrate - FPrate}{2} = 0.949$$

Perbandingan hasil metode algoritma decision tree C4.5 tanpa metode Korelasi Pearson dengan metode algoritma decision tree C4.5 + metode korelasi pearson nampak pada tabel 8 berikut ini :

Tabel 8. Hasil Perbandingan Antara Sekenario

Metode	Akurasi	AUC
C4.5	95.31 %	0.936
C4.5 + Korelasi Pearson	96.16 %	0.949

Dari tabel 8 dapat dilihat bahwa hasil uji metode algoritma decision tree C4.5 memiliki nilai akurasi sebesar 95.31% dengan nilai AUC sebesar 0.936. Sedangkan hasil pengujian metode algoritma decision tree C4.5 menggunakan metode Korelasi Pearson sebagai pemilihan fitur mendapatkan hasil nilai akurasi sebesar 96.16% dengan nilai AUC sebesar 0.949.

D. Simpulan

Dalam penelitian ini dapat nampak keberhasilan peningkatan nilai akurasi dan nilai AUC dari algoritma C4.5 melalui penerapan metode seleksi fitur korelasi pearson. Peningkatan nilai besaran akurasi dari 95.31% menjadi 96.16% dengan nilai AUC sebesar 0.936 menjadi 0.949. Untuk penelitian selanjutnya dapat menerapkan model menggunakan algoritma lain, sehingga tingkat akurasi yang berbeda akan diperoleh. Pencarian fitur-fitur penting dalam penelitian ini memainkan peran penting. Sehingga optimasi berfokus pada optimalisasi seleksi fitur dari data untuk model algoritma C4.5 dan model lain sangat diperlukan.

E. Referensi

[1] B. Jain, N. Ranawat, P. Chittora, P. Chakrabarti, and S. Poddar, "A machine learning perspective: To analyze diabetes," *Mater. Today Proc.*, no. xxxx, 2021, doi: 10.1016/j.matpr.2020.12.445.

[2] C. H. Mok, H. H. Y. Kwok, C. S. Ng, G. M. Leung, and J. Quan, "Health State Utility Values for Type 2 Diabetes and Related Complications in East and Southeast Asia: A Systematic Review and Meta-Analysis," *Value Heal.*, vol. 24, no. 7, pp. 1059–1067, 2021, doi: 10.1016/j.jval.2020.12.019.

[3] A. F. Daru, M. B. Hanif, and E. Widodo, "Improving Neural Network Performance with Feature Selection Using Pearson Correlation Method for Diabetes Disease Detection," *JUITA J. Inform.*, vol. 9, no. 1, p. 123, 2021, doi: 10.30595/juita.v9i1.9941.

[4] S. J. Lee, Z. Xu, T. Li, and Y. Yang, "A novel bagging C4.5 algorithm based on wrapper feature selection for supporting wise clinical decision making," *J. Biomed. Inform.*, vol. 78, pp. 144–155, 2018, doi: 10.1016/j.jbi.2017.11.005.

[5] Z. Xu, D. Shen, T. Nie, Y. Kou, N. Yin, and X. Han, "A cluster-based oversampling algorithm combining SMOTE and k-means for imbalanced medical data," *Inf. Sci. (Ny).*, vol. 572, pp. 574–589, 2021, doi: 10.1016/j.ins.2021.02.056.

[6] N. Ahmed *et al.*, "Machine learning based diabetes prediction and development of smart web application," *Int. J. Cogn. Comput. Eng.*, vol. 2, no. December, pp. 229–241, 2021, doi: 10.1016/j.ijcce.2021.12.001.

[7] G. Tripathi and R. Kumar, "Early Prediction of Diabetes Mellitus Using Machine Learning," *ICRITO 2020 - IEEE 8th Int. Conf. Reliab. Infocom Technol. Optim. (Trends Futur. Dir.*, pp. 1009–1014, 2020, doi: 10.1109/ICRITO48877.2020.9197832.

[8] D. Vigneswari, N. K. Kumar, V. Ganesh Raj, A. Gugan, and S. R. Vikash, "Machine Learning Tree Classifiers in Predicting Diabetes Mellitus," *2019 5th Int. Conf. Adv. Comput. Commun. Syst. ICACCS 2019*, pp. 84–87, 2019, doi: 10.1109/ICACCS.2019.8728388.

[9] H. Hairani, K. E. Saputro, and S. Fadli, "K-means-SMOTE for handling class imbalance in the classification of diabetes with C4.5, SVM, and naive Bayes," *J. Teknol. dan Sist. Komput.*, vol. 8, no. 2, pp. 89–93, 2020, doi: 10.14710/jtsiskom.8.2.2020.89-93.

[10] G. Manikandan and S. Abirami, "An efficient feature selection framework based on information theory for high dimensional data," *Appl. Soft Comput.*, vol. 111, p. 107729, 2021, doi: 10.1016/j.asoc.2021.107729.

[11] Y. Mu, X. Liu, and L. Wang, "A Pearson's correlation coefficient based decision tree and its parallel implementation," *Inf. Sci. (Ny).*, vol. 435, pp. 40–58, 2018.
[12] M M Faniqul Islam, Rahatara Ferdousi, Sadikur Rahman, and Yasmin Bushra, "Early stage diabetes risk prediction dataset," 2022. https://archive.ics.uci.edu/ml/datasets/Early stage diabetes risk prediction dataset. (accessed Aug. 15, 2022).
[13] M. B. Hanif and Khoirudin, "Sistem Aplikasi Prediksi Penyakit Diabetes Menggunakan Fiture Selection Korelasi Pearson Dan Klasifikasi Naive Bayes," *Pengemb. Rekayasa dan Teknol.*, vol. 16, no. 2, pp. 199–205, 2020.
[14] S. Eker, E. Rovenskaya, S. Langan, and M. Obersteiner, "Model validation: A bibliometric analysis of the literature," *Environ. Model. Softw.*, vol. 117, no. December 2018, pp. 43–54, 2019, doi: 10.1016/j.envsoft.2019.03.009.
[15] V. Kotu and B. Deshpande, "Model Evaluation," *Data Sci.*, pp. 263–279, 2019, doi: 10.1016/b978-0-12-814761-0.00008-3.