UNIVERSITAS
MUHAMMADIYAH
SIDOARJO

UMSIDA
DARI SINI PENCERAHAN BERSEMI

Prediksi *Early-Stage* Diabetes Mellitus menggunakan Pendekatan *Ensemble Learning*

Oleh:

Syaikhina Usabili

Uce Indahyanti

Progam Studi Manajemen Informasi Keshatan

Universitas Muhammadiyah Sidoarjo

Agustus 2023

Pendahuluan

Ciri khas Diabetes Mellitus ditandai dengan hiperglikemia yang disebabkan oleh pankreas yang tidak dapat memproduksi insulin dengan baik. Diabetes memiliki gejala tahap awal yang dapat dijadikan sebagai tolak ukur seseorang terprediksi Diabetes Mellitus atau tidak, Diabetes Mellitus juga dikenal sebagai Diabetes tipe 1, dimana gangguan metabolism kronis dapat ditandai dengan hiperglikemia yang berkembang Ketika pancreas tidak dapat memproduksi insulin dengan baik. Adapun Diabetes Mellitus tipe 2 yang disebabkan oleh pola makan yang tidak teratur sehingga dapat menghasilkan kadar gula darah yang tinggi dan mempengaruhi seberapa baik tubuh menggunakan insulin. Tahap awal diabetes meliputi berbagai gejala seperti sering buang air kecil, rasa haus yang berlebihan, kadar gula darah tinggi, penurunan berat badan, pemulihan tubuh yang buruk. berdasarkan data Rumah Sakit Umum Daerah Sidoarjo, diabetes menempati urutan keempat dari 10 penyakit terbesar di Rumah Sakit Umum Daerah Sidoarjo.

Rumusan Masalah

Bagaimana cara memprediksi *early-stage* penyakit Diabetes Mellitus menggunakan metode *Ensemble Learning*?

Bagaimana cara menguji permodelan prediksi menggunakan Teknik *Confusion Matrix*?

Manfaat penelitian

Manfaat penelitian ini sebagai informasi mengenai prediksi gejala awal dari penyakit diabetes mellitus dan memberikan informasi terkait tingkat keakuratan prediksi gejala awal menggunakan *Ensemble Learning*

Research Gap

berdasarkan penelitian dari Widya Apriliah dan kawan – kawan pada tahun 2021 mengenai Klasifikasi data mining untuk diagnosa Diabetes Mellitus menggunakan *Random Forest* dan mengambil dataset berupa data sekunder atau data public dari UCI *Machine Learning* yang bersumber dari *Hospital in Sylhet*, Bangladesh menunjukkan bahwa algoritma terbaik dihasilkan dari *Random Forest* dengan hasil 97.88% dan temuan tersebut dikonfirmasi menggunakan kurva *Receiver Operating Characteristic* (ROC)

Penelitian dari Wahyu Nugraha dan Raja Sabaruddin pada tahun 2021 mengenai klasifikasi data mining untuk diagnosa Diabetes menggunakan C45, *Random Forest*, dan (*Support Vector Machines*) SVM. dataset yang digunakan berupa data sekunder atau data public yang diambil dari *Kaggle* Dataset Repository (UCI Pima Indians Diabetes Datasets) menunjukkan bahwa model klasifikasi SVM memiliki performa terbaik dengan nilai sebesar 80%

Metode

Jenis penelitian

Jenis penelitian berupa penelitian Kuantitatif

Data Penelitian

Data penelitian yang digunakan berupa data yang diperoleh dari Rumah Sakit Umum Daerah Sidoarjo dan data bersifat privat

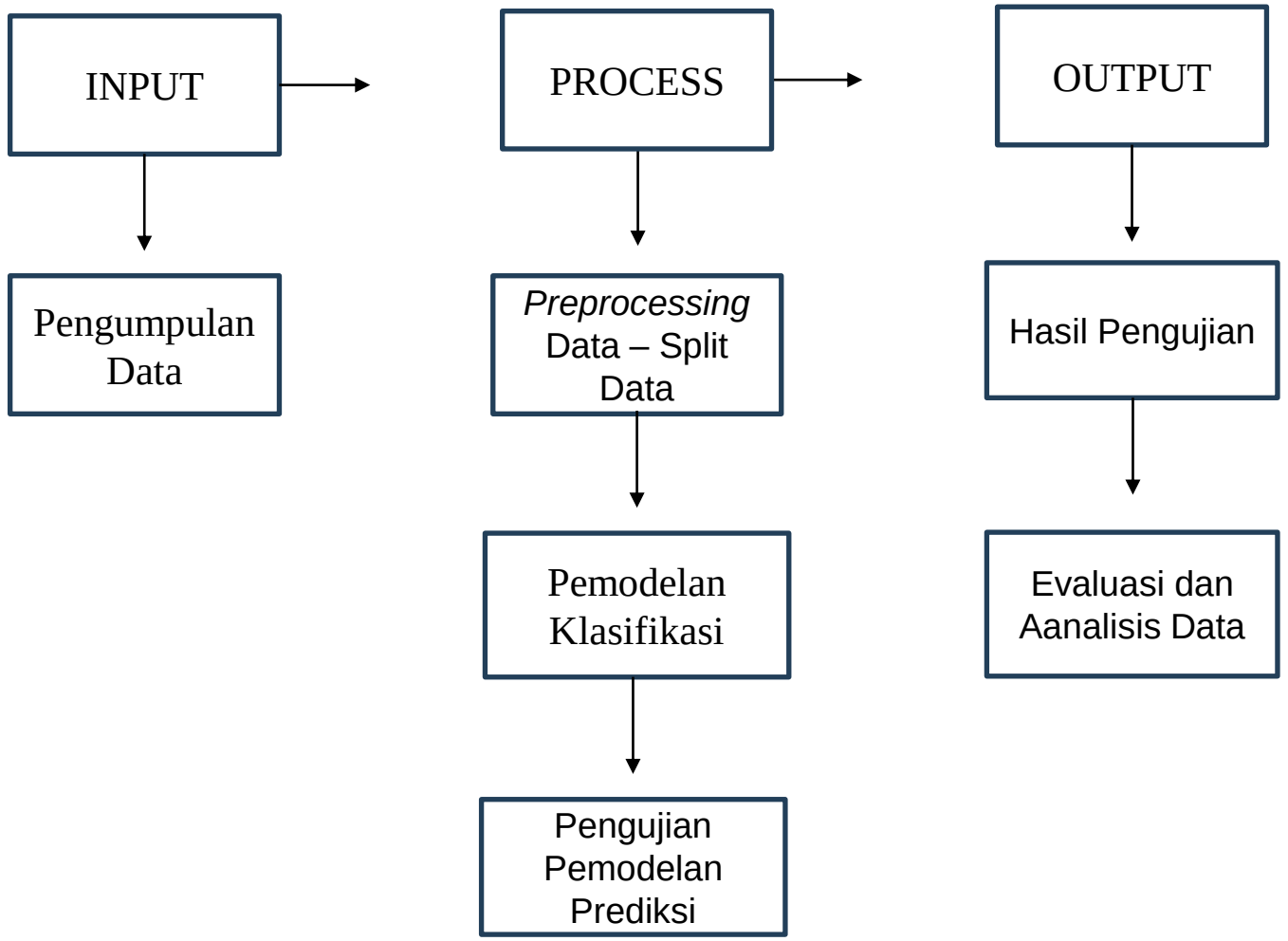
Variable Penelitian

Umur, Jenis Kelamin, Sering Haus, Sering Kencing, Luka Lama Sembuh, Berat Badan Menurun, Kadar Gula Meningkat, Obesitas, Penglihatan Kabur, Lemas/ Mudah Lelah, Class (pelabelan/target)

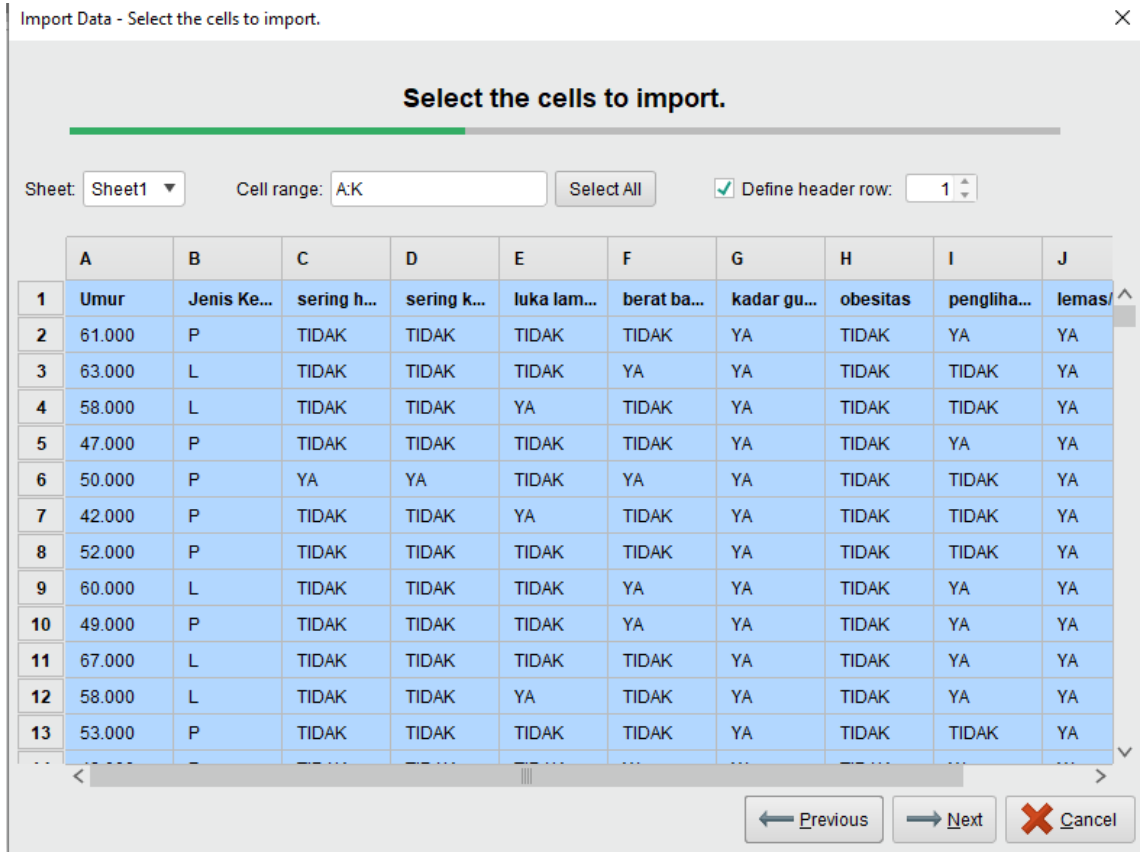
Teknik Analisa Data

Ensemble Learning berupa algoritma *Decision Tree* dan *Random Forest* dengan Hasil *Confusion Matrix*

Alur penelitian Data Mining

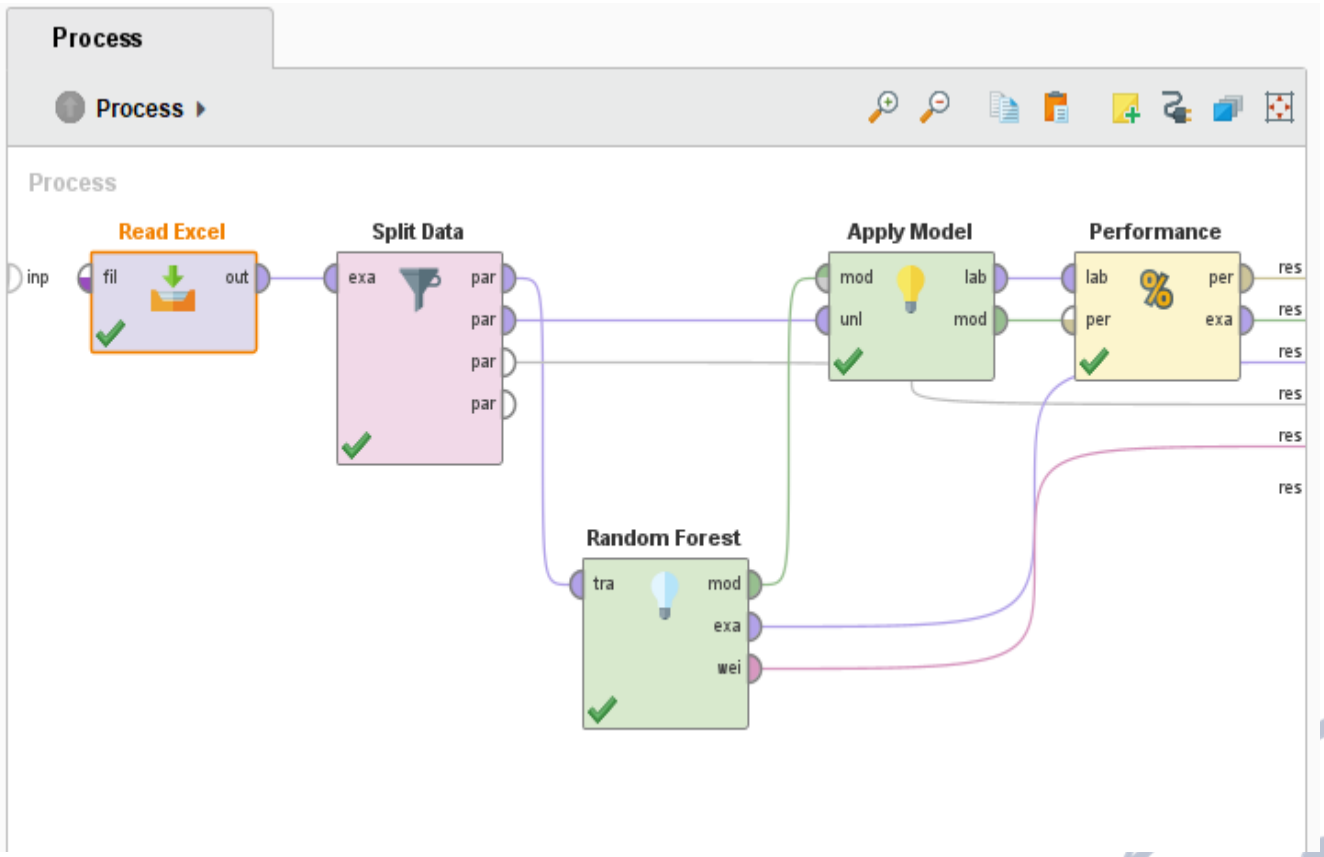


Pembahasan



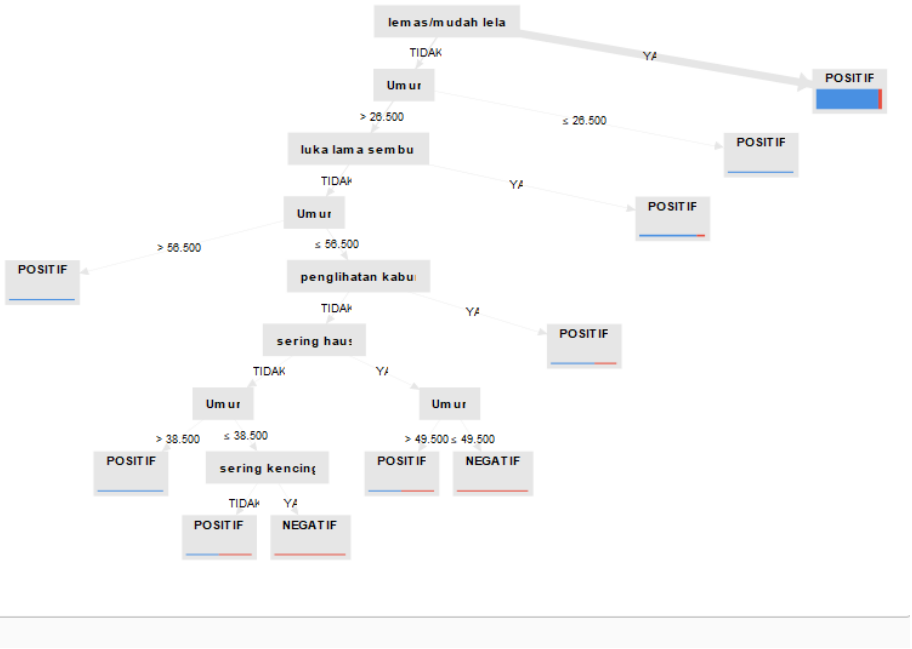
	A	B	C	D	E	F	G	H	I	J
1	Umur	Jenis Ke...	sering h...	sering k...	luka lam...	berat ba...	kadar gu...	obesitas	pengliha...	lemas/
2	61.000	P	TIDAK	TIDAK	TIDAK	TIDAK	YA	TIDAK	YA	YA
3	63.000	L	TIDAK	TIDAK	TIDAK	YA	YA	TIDAK	TIDAK	YA
4	58.000	L	TIDAK	TIDAK	YA	TIDAK	YA	TIDAK	TIDAK	YA
5	47.000	P	TIDAK	TIDAK	TIDAK	TIDAK	YA	TIDAK	YA	YA
6	50.000	P	YA	YA	TIDAK	YA	YA	TIDAK	TIDAK	YA
7	42.000	P	TIDAK	TIDAK	YA	TIDAK	YA	TIDAK	TIDAK	YA
8	52.000	P	TIDAK	TIDAK	TIDAK	TIDAK	YA	TIDAK	TIDAK	YA
9	60.000	L	TIDAK	TIDAK	TIDAK	YA	YA	TIDAK	YA	YA
10	49.000	P	TIDAK	TIDAK	TIDAK	YA	YA	TIDAK	YA	YA
11	67.000	L	TIDAK	TIDAK	TIDAK	TIDAK	YA	TIDAK	YA	YA
12	58.000	L	TIDAK	TIDAK	YA	TIDAK	YA	TIDAK	YA	YA
13	53.000	P	TIDAK	TIDAK	TIDAK	TIDAK	YA	TIDAK	TIDAK	YA

Gambar *Preprocessing* data dengan *software Rapidminer*



Gambar desain Proses data yang akan diolah menggunakan operator tertentu

Pembahasan



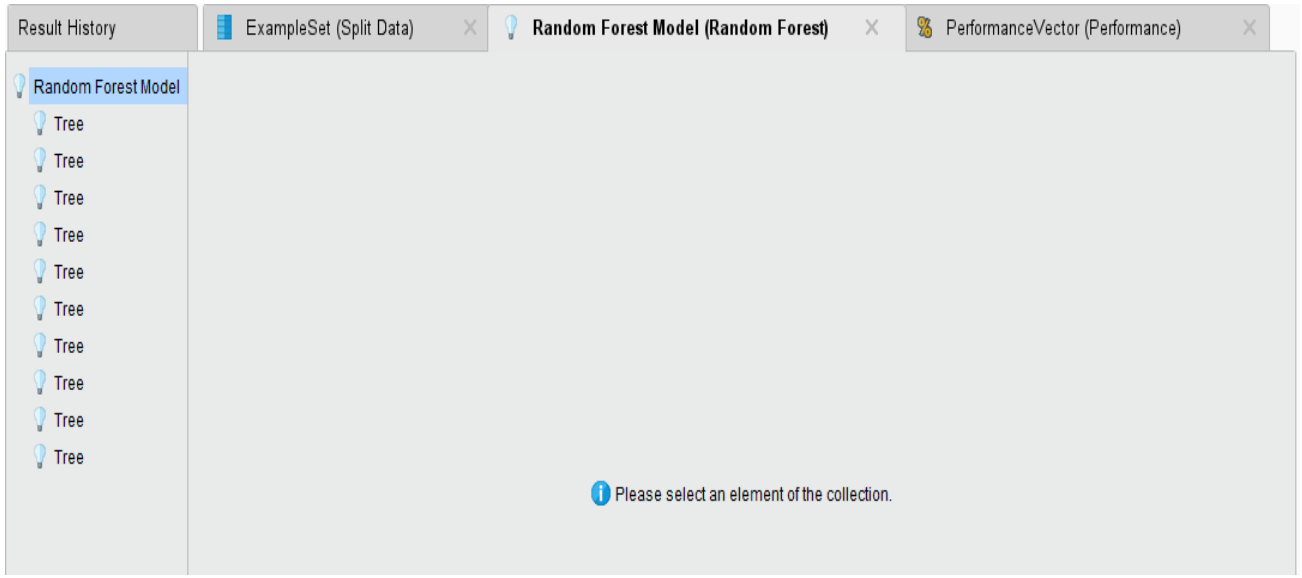
Berdasarkan pohon keputusan diatas, akar dari *Decision Tree* berupa lemas atau mudah Lelah dengan deskripsi :

1. Jika lemas atau mudah lelah = ya, luka lama sembuh = ya, maka class menunjukkan positif
2. Jika lemas atau mudah Lelah = ya, umur >56, maka class menunjukkan hasil positif

	True Positive	True Negative	Class Precision
Pred. Positive	68	6	91.89%
Pred. Negative	0	1	100.00%
Class Recall	100.00%	14.29%	

Adapun hasil dari *confusion matrix* atau uji tingkat akurasi dari klasifikasi *Decision tree* dapat dilihat bahwa nilai akurasi sebesar 92.00%, yang mana jumlah data benar atau *True* positif pada prediksi positif menghasilkan sebanyak 68 data dan 6 data lainnya yang tidak sesuai atau *true negative*. Sedangkan pada prediksi negative memiliki data benar sebesar 0 dan data yang tidak sesuai sebesar 1 data yang menghasilkan *class recall* 100% data benar dan 14.29% data yang tidak sesuai.

Pembahasan



Gambar diatas merupakan model dari *random forest* dengan 10 kali *maximal depth* atau 10 kali perputaran untuk *variable* secara acak, dan berikut adalah salah satu hasil terbaik dari 10 model *random forest*

Berdasarkan gambar pada pohon keputusan di atas, akar dari *Random Forest* berupa lemas atau mudah Lelah dengan deskripsi :

1. jika lemas atau mudah lelah = Ya, kadar gula meningkat = ya, maka class menunjukkan hasil Positif
2. Jika lemas atau mudah lelah= Ya, sering kencing = ya, umur >58, maka class menunjukkan hasil positif

Pembahasan

	True Positive	True Negative	Class Precision
Pred. Positive	67	6	91.78%
Pred. Negative	1	1	50.00 %
Class Recall	98.53%	14.29%	

Tabel diatas adalah confusion matrix yang menunjukkan bahwa nilai akurasi sebesar 90.67%, yang dimana jumlah data benar atau *true* positif pada prediksi positif sebesar 67 dan data yang tidak sesuai sebanyak 6 data pada *true negative*, sedangkan pada prediksi *negative* terdapat data benar sebanyak 1 data dan juga 1 data lainnya pada data yang tidak sesuai sehingga menghasilkan *class recall* pada data benar yaitu 98.53% dan data tidak sesuai sebesar 14.29%.

Tingkat akurasi yang ditujukan untuk menilai hasil terhadap klasifikasi bernilai 0.80- 0.90 = *good classification*, 0.90 – 1.00 = *excellent classification*

Kesimpulan

klasifikasi menjadi salah satu metode alternatif yang mudah digunakan dan mudah dipahami untuk membuat sebuah informasi dengan beberapa model yang ada dalam berbagai bidang, salah satunya bidang Kesehatan. Hasil analisis klasifikasi *Decision Tree* Diabetes Mellitus yang ada di Rumah Sakit Umum Daerah Sidoarjo berdasarkan *Confusion Matrix* menghasilkan nilai sebesar 92.00% sedangkan klasifikasi *Random Forest* pada penyakit Diabetes Mellitus berdasarkan Confusion Matrix didapatkan hasil sebesar 90.67% yang dimana kedua nilai yang dihasilkan dari 2 klasifikasi tersebut dikategorikan sebagai kategori klasifikasi sangat baik atau *excellent classification.* Adapun hasil dari klasifikasi *Decision Tree* dan *Random Forest* tidak terlalu signifikan sehingga menghasilkan hasil yang berbeda namun masih dengan kategori yang sama.

Adapun bagi Kesehatan, pada penelitian ini yang menjadikan pasien terdiagnosis Diabetes Mellitus dengan gejala tahap awal berupa lemas / mudah Lelah, umur, dan kadar gula meningkat

Referensi

- [1] B. E. Nyarko, R. S. Amoah, and A. Crimi, “Boosting diabetes and pre-diabetes detection in rural Ghana [version 2; peer review: 2 approved],” *F1000 Res.*, vol. 8, p. 19, Aug. 2019, doi: https://doi.org/10.12688/f1000research.18497.2.
- [2] W. Yusnaeni and W. Widiarina, “Penerapan Algoritma C4.5 Dalam Prediksi Resiko Diabetes Tahap Awal (Early Stage Diabetes),” *J. Tek. Komput.*, vol. 8, no. 1, pp. 56–60, Jan. 2022, doi: 10.31294/jtk.v8i1.11566.
- [3] A. M. Argina, “Penerapan Metode Klasifikasi K-Nearest Neigbor pada Dataset Penderita Penyakit Diabetes,” *Indones. J. Data Sci.*, vol. 1, no. 2, pp. 29–33, Jul. 2020, doi: 10.33096/ijodas.v1i2.11.
- [4] D. Magliano and E. J. Boyko, *IDF diabetes atlas*, 10th edition. Brussels: International Diabetes Federation, 2021.
- [5] A. Ridwan, “Penerapan Algoritma Naïve Bayes Untuk Klasifikasi Penyakit Diabetes Mellitus,” *J. SISKOM-KB Sist. Komput. Dan Kecerdasan Buatan*, vol. 4, no. 1, pp. 15–21, Oct. 2020, doi: 10.47970/siskom-kb.v4i1.169.
- [6] P. Subarkah, “Penerapan Algoritme Klasifikasi Classification And Regression Trees (Cart) Untuk Diagnosis Penyakit Diabetes Retinopathy,” *MATRIK J. Manaj. Tek. Inform. Dan Rekayasa Komput.*, vol. 19, no. 2, pp. 294–301, May 2020, doi: 10.30812/matrik.v19i2.676.
- [7] A. I. Kusumarini, P. A. Hogantara, M. Fadhlurohman, and N. Chamidah, “Perbandingan Algoritma Random Forest, Naïve Bayes, Dan Decision Tree Dengan Oversampling Untuk Klasifikasi Bakteri E. Coli,” 2021.
- [8] W. Apriliah, I. Kurniawan, M. Baydhowi, and T. Haryati, “Prediksi Kemungkinan Diabetes pada Tahap Awal Menggunakan Algoritma Klasifikasi Random Forest,” *SISTEMASI*, vol. 10, no. 1, p. 163, Jan. 2021, doi: 10.32520/stmsi.v10i1.1129.
- [9] W. Nugraha and R. Sabaruddin, “Teknik Resampling untuk Mengatasi Ketidakseimbangan Kelas pada Klasifikasi Penyakit Diabetes Menggunakan C4.5, Random Forest, dan SVM,” *Techno.Com*, vol. 20, no. 3, pp. 352–361, Aug. 2021, doi: 10.33633/tc.v20i3.4762.
- [10] B. T. R. Doni, S. Susanti, and A. Mubarok, “PENERAPAN DATA MINING UNTUK KLASIFIKASI PENYAKIT HEPATOCELLULAR CARCINOMA MENGGUNAKAN ALGORITMA NAÏVE BAYES,” *J. Responsif Ris. Sains Dan Inform.*, vol. 3, no. 1, pp. 12–19, Feb. 2021, doi: 10.51977/jti.v3i1.403.
- [11] A. K. F. Aidia, P. J. Amelia, and V. R. Setyaning Nastiti, “Prediksi Jumlah Pasien Covid-19 Dengan Menggunakan Klasifikasi Algoritma Machine Learning,” *SINTECH Sci. Inf. Technol. J.*, vol. 5, no. 2, pp. 165–172, Oct. 2022, doi: 10.31598/sintechjournal.v5i2.1163.
- [12] R. Ghorbani and R. Ghousi, “Predictive data mining approaches in medical diagnosis: A review of some diseases prediction,” *Int. J. Data Netw. Sci.*, pp. 47–70, 2019, doi: 10.5267/j.ijdns.2019.1.003.
- [13] F. Aris, “Penerapan Data Mining untuk Identifikasi Penyakit Diabetes Melitus dengan Menggunakan Metode Klasifikasi,” vol. 1, no. 1, 2019.
- [14] M. Azhari, Z. Situmorang, and R. Rosnelly, “Perbandingan Akurasi, Recall, dan Presisi Klasifikasi pada Algoritma C4.5, Random Forest, SVM dan Naive Bayes,” *J. MEDIA Inform. BUDIDARMA*, vol. 5, no. 2, p. 640, Apr. 2021, doi: 10.30865/mib.v5i2.2937.
- [15] F. Elfaladonna and A. Rahmadani, “ANALISA METODE CLASSIFICATION-DECISSION TREE DAN ALGORITMA C.45 UNTUK MEMPREDIKSI PENYAKIT DIABETES DENGAN MENGGUNAKAN APLIKASI RAPID MINER,” *SINTECH Sci. Inf. Technol. J.*, vol. 2, no. 1, pp. 10–17, Apr. 2019, doi: 10.31598/sintechjournal.v2i1.293.

- [16] D. R. Ente, S. A. Thamrin, S. Arifin, H. Kuswanto, and A. Andreza, “KLASIFIKASI FAKTOR-FAKTOR PENYEBAB PENYAKIT DIABETES MELITUS DI RUMAH SAKIT UNHAS MENGGUNAKAN ALGORITMA C4.5,” *Indones. J. Stat. Its Appl.*, vol. 4, no. 1, pp. 80–88, Feb. 2020, doi: 10.29244/ijsa.v4i1.330.
- [17] M. Tarawneh and O. Embarak, “Hybrid Approach for Heart Disease Prediction Using Data Mining Techniques,” in *Advances in Internet, Data and Web Technologies*, L. Barolli, F. Xhafa, Z. A. Khan, and H. Odhabi, Eds., in Lecture Notes on Data Engineering and Communications Technologies, vol. 29. Cham: Springer International Publishing, 2019, pp. 447–454. doi: 10.1007/978-3-030-12839-5_41.
- [18] D. H. Depari, Y. Widiastiwi, and M. M. Santoni, “Perbandingan Model Decision Tree, Naive Bayes dan Random Forest untuk Prediksi Klasifikasi Penyakit Jantung,” *Inform. J. Ilmu Komput.*, vol. 18, no. 3, p. 239, Dec. 2022, doi: 10.52958/iftk.v18i3.4694.
- [19] Y. Widiastiwi and I. Ernawati, “Klasifikasi Penyakit Batu Ginjal Menggunakan Algoritma Decision Tree C4.5 Dengan Membandingkan Hasil Uji Akurasi,” vol. 5, no. 2, 2021.
- [20] S. Rahayu and J. J. Purnama, “KLASIFIKASI KONSUMSI ENERGI INDUSTRI BAJA MENGGUNAKAN TEKNIK DATA MINING,” *J. Teknoinfo*, vol. 16, no. 2, p. 395, Jul. 2022, doi: 10.33365/jti.v16i2.1984.
- [21] M. Syukron, R. Santoso, and T. Widiharih, “PERBANDINGAN METODE SMOTE RANDOM FOREST DAN SMOTE XGBOOST UNTUK KLASIFIKASI TINGKAT PENYAKIT HEPATITIS C PADA IMBALANCE CLASS DATA,” *J. Gaussian*, vol. 9, no. 3, pp. 227–236, Aug. 2020, doi: 10.14710/j.gauss.v9i3.28915.
- [22] C. C. Aggarwal, *Data Mining: The Textbook*. Cham: Springer International Publishing, 2015. doi: 10.1007/978-3-319-14142-8.
- [23] D. Krstinić, M. Braović, L. Šerić, and D. Božić-Štulić, “Multi-label Classifier Performance Evaluation with Confusion Matrix,” *Comput. Sci.*.
- [24] P. Cavalin and L. Oliveira, “Confusion Matrix-Based Building of Hierarchical Classification,” in *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, R. Vera-Rodriguez, J. Fierrez, and A. Morales, Eds., in Lecture Notes in Computer Science, vol. 11401. Cham: Springer International Publishing, 2019, pp. 271–278. doi: 10.1007/978-3-030-13469-3_32.
- [25] F. Gorunescu, *Data Mining*, vol. 12. in Intelligent Systems Reference Library, vol. 12. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011. doi: 10.1007/978-3-642-19721-5.


UMSIDA
DARI SINI PENCERAHAN BERSEMI