Terbit online pada laman web jurnal: **https://ejurnalunsam.id/index.php/jicom/**

JICOM
Jurnal Informatika dan Teknologi Komputer

Vol. 06 No. 02 (2025) 83 - 88 E-ISSN :2774-7115 P-ISSN: 2775-2089

Klasifikasi Diagnosa Penyakit Tiroid Menggunakan Metode Random Forest

Raja Darmawan[1], Yulison Herry Chrisnanto[2], Gunawan Abdillah[3]
[1,2,3] Informatika, Fakultas Sains dan Informatika, Universitas Jenderal Achmad Yani
[1]rajadarmawan20@if.unjani.ac.id, [2]yhc@if.unjani.ac.id, [3]gunawanabdillah03@gmail.com,

Abstract

The thyroid is a vital gland in the human neck, regulating metabolism through its hormones. Hormonal disorders in the thyroid can significantly affect health. Data mining techniques, such as the random forest algorithm, are used to analyze thyroid disease data. Previous research has used methods such as Decision Tree and Support Vector Machine with high accuracy results. This study aims to apply the random forest method in the classification of thyroid disease diagnoses. Research questions include factors that affect accuracy, the effect of parameter changes on the model, and optimal data sharing. The results show that parameters such as the number of decision trees, maximum depth, and minimum number of samples can affect the accuracy of the model. The evaluation showed that the highest accuracy was obtained in the first test with a data split of 80/20 with an accuracy result of 99%. This study concludes that the random forest method is effective in improving the accuracy of thyroid disease diagnosis and the importance of parameter adjustment for optimal results.

Keywords: Thyroid Disease, Classification, Random Forest

Abstrak

Tiroid merupakan kelenjar vital di leher manusia, mengatur metabolisme melalui hormonnya. Gangguan hormonal pada tiroid dapat mempengaruhi kesehatan secara signifikan. Teknik data mining, seperti algoritma random forest, digunakan untuk menganalisis data penyakit tiroid. Penelitian sebelumnya telah menggunakan metode seperti Decision Tree dan Support Vector Machine dengan hasil akurasi tinggi. Penelitian ini bertujuan untuk menerapkan metode random forest dalam klasifikasi diagnosa penyakit tiroid. Pertanyaan penelitian termasuk faktor-faktor yang memengaruhi akurasi, pengaruh perubahan parameter terhadap model, dan pembagian data yang optimal. Hasil penelitian menunjukkan bahwa parameter seperti jumlah pohon keputusan, kedalaman maksimum, dan jumlah minimum sampel dapat memengaruhi akurasi model. Evaluasi menunjukkan akurasi tertinggi diperoleh pada pengujian pertama dengan pembagian data 80/20 dengan hasil akurasi diperoleh 99%. Penelitian ini menyimpulkan bahwa metode random forest efektif dalam meningkatkan akurasi diagnosa penyakit tiroid dan pentingnya penyesuaian parameter untuk hasil yang optimal.

Kata kunci: Penyakit Tiroid, Klasifikasi, Random Forest

1. Pendahuluan

Bagian tubuh manusia yang paling penting adalah tiroid, yang terletak di bawah jakun pada leher. Kelenjar ini, yang memiliki bentuk seperti kupu-kupu, menghasilkan hormon tiroid dan bertanggung jawab untuk menjaga metabolism [1]. Sangat penting untuk mengamati dan mengukur berbagai gangguan hormonal pada manusia. Faktor penyebab tiroid adalah fleksibilitas hormon manusia yang terlalu tinggi atau rendah. Perawatan kesehatan penting untuk menyeimbangkan variasi hormon yang terlalu tinggi atau rendah. Faktor risiko untuk gangguan hormon cukup banyak, sehingga dokter membutuhkan perhatian yang lebih besar untuk memantau dan menemukan diagnosis yang tepat pada waktu yang tepat [2]. Karena gejala gangguan fungsi tiroid sangat mirip dengan keluhan yang disebabkan oleh gaya hidup modern, gangguan fungsi tiroid sering kali sulit diidentifikasi. Akibatnya, pasien sering mengabaikan masalahnya dan tidak memeriksanya ke dokter [3].

Oleh karena itu, Teknik data mining diperlukan untuk menganalisa data penyakit ini [4]. Dalam Teknik Data mining terdapat berbagai algoritma klasifikasi seperti naive bayes, decision tree, neural network, k-nn, random forest, dan sebagainya. untuk menggambarkan dan membedakan kelas data atau konsep. [5]. Teknik data mining yang digunakan adalah algoritma klasifikasi

random forest. Random forest adalah konsep pembelajaran kelompok yang menggunakan hasil dari beberapa pohon keputusan untuk meningkatkan akurasi. Metode ini bekerja dengan membangun pohon keputusan secara acak dari kumpulan data. Teknik ini dapat menghasilkan tingkat akurasi yang tinggi dan menghindari masalah overfitting, tetapi juga memerlukan waktu pelatihan yang lebih lama. [6].

Berdasarkan penelitian terdahulu melakukan penelitian untuk penyakit tiroid dengan menerapkan metode Decision Tree dan Support Vector Machine, didapatkan hasil nilai akurasi sebesar 99.89% [7]. Penelitian tambahan tentang pengujian dengan menggunakan dataset penyakit tiroid sebanyak 500 data, algoritma Naïve Bayes memiliki akurasi sebesar 63% pada data testing dan 64% pada data training, sedangkan algoritma K-Nearest Neighbors (KNN) memiliki akurasi sebesar 63% [8].

Berdasarkan latar belakang yang telah disebutkan, penelitian ini bertujuan untuk mengetahui bagaimana penerapan metode random forest dalam proses klasifikasi diagnosa penyakit tiroid dengan hasil performa terbaik serta membuat model klasifikasi yang sederhana dan memiliki performa baik.

1.1. Tinjauan Literatur

Data Mining

Data mining merupakan teknik yang untuk mengekstrak informasi dan pengetahuan penting dari kumpulan data yang sangat besar. Informasi dan pengetahuan ini dapat digunakan untuk berbagai tujuan, seperti riset ilmiah, analisis bisnis, dan pengambilan keputusan [9]. Sederhananya, data mining adalah penemuan pola pengolahan data yang bermanfaat [10]. Data mining merupakan proses mengekstrak informasi dari data dalam jumlah besar. Informasi tersebut dapat berupa fakta, kesimpulan, pola, atau anomali. Data mining memiliki lima fungsi utama, yaitu [11]:

- Klasifikasi adalah proses memprediksi kelas data baru berdasarkan data yang telah dilatih sebelumnya.
- Clustering adalah proses membagi data ke dalam kelompok-kelompok berdasarkan atributnya.
- Association adalah proses mengidentifikasi hubungan antara dua atau lebih variabel.
- Sequencing adalah proses mengidentifikasi urutan kejadian atau peristiwa.
- Forecasting adalah proses memprediksi nilai pada masa yang akan datang berdasarkan data masa lalu.

Klasfikasi

Klasifikasi adalah proses pembelajaran mesin yang menggunakan fitur-fitur dari suatu objek untuk memprediksi kelas objek tersebut. Metode ini bekerja dengan membangun model yang dapat mengidentifikasi ciri-ciri yang membedakan satu kelas dengan kelas lainnya. Selanjutnya, model ini dapat digunakan untuk memprediksi kelas objek baru yang label kelasnya belum diketahui [10]. Klasifikasi adalah proses memprediksi kelas data baru dengan menggunakan data yang telah dilatih sebelumnya. Data yang telah dilatih disebut data training, sedangkan data baru disebut data testing. Kemungkinan kelas data baru ditentukan berdasarkan label atau atribut tujuan dari data training [12].

Random Forest

Random Forest merupakan kumpulan pohon keputusan yang digunakan untuk mengklasifikasi data ke suatu kelas. Ini adalah salah satu teknik yang dapat meningkatkan akurasi hasil dalam menciptakan atribut untuk setiap node yang dilakukan secara acak. Pohon keputusan dibangun dengan menentukan node akar, lalu cabang-cabang pohon akan bercabang lagi hingga mencapai node daun. Hasil akhir dari pohon keputusan adalah kelas yang diprediksi untuk data baru [13]. Metode Random Forest digunakan untuk melakukan klasifikasi dan regresi. Terdapat tiga poin kunci: (1) pembangunan pohon prediksi dengan melakukan bootstrap sampling; (2) Setiap pohon keputusan melaksanakan memanfaatkan variabel prediktor secara acak; (3) hasil dari setiap pohon keputusan digabungkan oleh Random Forest menggunakan mayoritas suara untuk klasifikasi atau rata-rata untuk regresi. Metode ini menggabungkan hasil dari berbagai pohon keputusan untuk meningkatkan akurasi dan kinerja model [14].

Metode Random Forest memiliki dua parameter utama, yaitu jumlah pohon keputusan (tree) dan jumlah variabel independen yang digunakan untuk proses pencabangan. Untuk menentukan nilai parameter mtry, disarankan menggunakan persamaan berikut [15]:

$$mtry_1 = \frac{1}{2}|\sqrt{p}|$$

$$mtry_2 = |\sqrt{p}|$$

$$mtry_3 = 2 \times |\sqrt{p}|$$

mtry = jumlah variabel independen yang signifikan untuk setiap split

p = jumlah variabel independent

Confusion Matrix

Confusion matrix adalah tabel yang digunakan untuk mengukur akurasi dan kinerja algoritma klasifikasi.

Tabel ini menunjukkan berapa banyak data testing yang diprediksi dengan benar dan salah oleh algoritma tersebut. Confusion matrix dapat digunakan untuk menilai kinerja algoritma klasifikasi dalam menyelesaikan masalah klasifikasi [13]. Confusion matrix merupakan sebuah metode yang menghitung akurasi, recall, dan precision melalui tabel matriks [15].

Tabel 1. Confusion Matrix

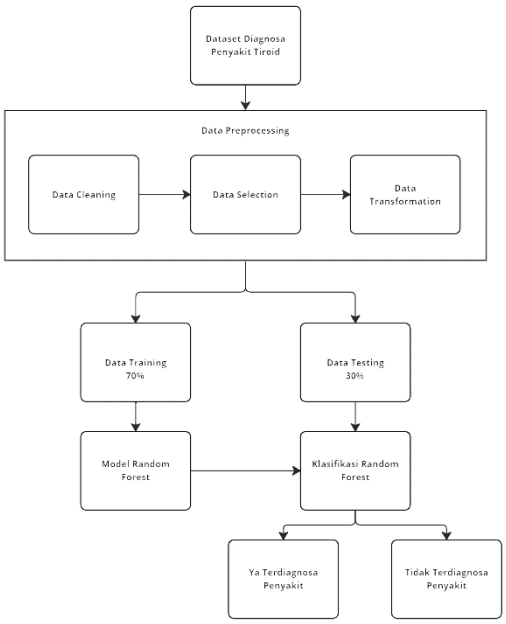
		True Class	
		Postitve	Negative
Predicted Class	Positive	True positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

2. Metode Penelitian



Gambar 1. Alur Penelitian

2.1. Pengumpulan Data

Pada penelitian ini data yang digunakan didapatkan dari Kaggle. Data ini menangkap berbagai atribut terkait kondisi tiroid untuk diagnosis medis. Ini mencakup informasi demografis seperti usia dan jenis kelamain. Fitur riwayat kesehatan mencakup asupan tiroksin, obat antitiroid, dan operasi sebelumnya. Status kesehatan pasien terkini mengenai penyakit, kehamilan, adanya penyakit gondok, tumor, atau kondisi hipofisis dicatat. Selain itu, ia mencatat apakah pasien mencurigai hipotiroidisme atau hipertiroidisme. Hasil laboratorium seperti kadar TSH, T3, TT4, T4U, FTI disertakan jika diukur. Klasifikasi biner menunjukkan ada tidaknya hipertiroidisme. Sumber rujukan juga ditunjukkan Dataset berisi atribut-atribut.

2.2. Preprocessing

Proses ini dilakukan untuk memeriksa dan memperbaiki kesalahan pada dataset. Teknik yang gunakan pada penelitian ini yaitu *cleaning*, *selection*, dan *transformation*. *Cleaning* diperlukan untuk mengatasi data yang hilang menggunakan hasil rata-rata data. Lalu proses *selection* dilakukan untuk menunjukan kontribusi relative setiap fitur dalam membuat keputusan. Semakin tinggi nilai penting suatu fitur, semakin besar pengaruh fitur tersebut dalam model. Selain itu, proses *transformation* digunakan untuk merubah data dengan tipe categorical menjadi numerik yaitu 0 dan 1.

2.3. Split Data

Setelah melewati proses preprocessing, kemudian dilakukan split data untuk data training dan data testing. Pembagian data pada penelitian ini menggunakan 70/30. 70% data training dan 30% data testing.

2.4. Klasifikasi Random Forest

Pada tahap ini dilakukan pembuatan model random forest untuk klasifikasi diagnose penyakit tiroid. Random forest ini memiliki beberapa parameter yang diujikan untuk mengoptimalkan model. Beberapa parameter tersebut adalah *n_estimator* (jumlah pohon keputusan), *max_depth* (kedalaman maksimum pohon keputusan), *min_sample_leaf* (jumlah minimum sampel daun), dan *random state* (nilai inisialisasi).

2.5. Evaluasi

Setelah itu dilakukan proses evaluasi terhadap performa model random forest. Evaluasi ini melibatkan penggunaan matriks seperti confusion matrix untuk mengukur metrik-metrik penting seperti akurasi, presisi, dan recall. Evaluasi ini membantu dalam memahami seberapa baik model tersebut dalam mengklasifikasikan data.

3. Hasil dan Pembahasan

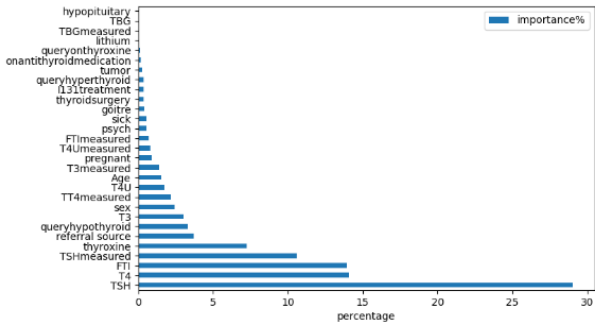
3.1. Preprocessing

Pada penelitian ini, sebelum data masuk ke tahap pemodelan, dilakukan terlebih dahulu proses pra-pemrosesan data. Pada tahapan *cleaning* data dilakukan proses pengisian melengkapi data yang hilang menggunakan rata-rata data. Berikut hasil data yang sudah di *cleaning* dalam dalam gambar 2.

```
Age                          0
sex                          0
thyroxine                    0
queryonthyroxine             0
onantithyroidmedication      0
sick                         0
pregnant                     0
thyroidsurgery               0
I131treatment                0
queryhypothyroid             0
queryhyperthyroid            0
lithium                      0
goitre                       0
tumor                        0
hypopituitary                0
psych                        0
TSHmeasured                  0
TSH                          0
T3measured                   0
T3                           0
TT4measured                  0
T4                           0
T4Umeasured                  0
T4U                          0
FTImeasured                  0
...
TBGmeasured                  0
TBG                          0
referral source              0
binaryClass                  0
```

Gambar 2. Cleaning

Selanjutnya, pada tahapan selection data dilakukan untuk menunjukkan kontribusi relatif setiap fitur dalam pembuatan keputusan. Semakin tinggi nilai kepentingan suatu fitur, semakin besar pengaruh fitur tersebut dalam model. Dengan memahami nilai kepentingan masing-masing fitur dapat diidentifikasikan fitur mana yang paling signifikan, sehingga dapat meningkatkan akurasi dan memiliki dampak lebih besar terhadap hasil akhir dari model. Hasil data yang sudah di *selection* ditampilkan dalam gambar 3.



Gambar 3. Selection

Pada tahapan *transformation* data, dilakukan perubahan data kategorikal menjadi numerik dengan mengonversi nilai “false” dan “true” menjadi “0” dan “1”. Hasil data yang sudah di transformation dapat dilihat dalam Gambar 4.

```
Age                          int64
sex                          int32
thyroxine                    int32
queryonthyroxine             int32
onantithyroidmedication      int32
sick                         int32
pregnant                     int32
thyroidsurgery               int32
I131treatment                int32
queryhypothyroid             int32
queryhyperthyroid            int32
lithium                      int32
goitre                       int32
tumor                        int32
hypopituitary                int32
psych                        int32
TSHmeasured                  int32
TSH                        float64
T3measured                   int32
T3                         float64
TT4measured                  int32
T4                         float64
T4Umeasured                  int32
T4U                        float64
FTImeasured                  int32
...
TBG                          int32
referral source              int32
binaryClass                  int32
```

Gambar 4. Tranformation

3.2. Split Data

Setelah melalui tahapan-tahapan tersebut, dataset kemudian dibagi menjadi data pelatihan dan data pengujian. Dalam penelitian ini, pembagian dataset dilakukan dengan dua cara. Pembagian pertama adalah 80/20, di mana 80% digunakan untuk data pelatihan dan 20% untuk data pengujian. Pembagian kedua adalah 70/30, dengan 70% untuk data pelatihan dan 30% untuk data pengujian. Kedua pembagian ini bertujuan untuk menguji kinerja model dalam melakukan klasifikasi. Berikut adalah pembagian data yang dilakukan dalam tabel 2.

Tabel 2. Split Data

Data Training	Data Testing
70%	30%
80%	20%

3.3. Klasifikasi Random Forest

Dalam penelitian ini, metode random forest digunakan untuk melakukan klasifikasi penyakit tiroid. Pada random forest, terdapat beberapa parameter yang diuji untuk meningkatkan kinerja dan mengoptimalkan model. Berikut adalah pengujian dengan pengaturan parameter random forest yang telah dilakukan untuk klasifikasi penyakit tiroid pada tabel 3 dan tabel 4.

Tabel 3. Pengujian Pertama

Split Data 80/20

Pengujian	1	2	3
n_estimator	50	100	150
max_depth	5	10	15
min_samples_leaf	5	10	15
random_state	42	42	42
Accuracy	95%	98%	99%

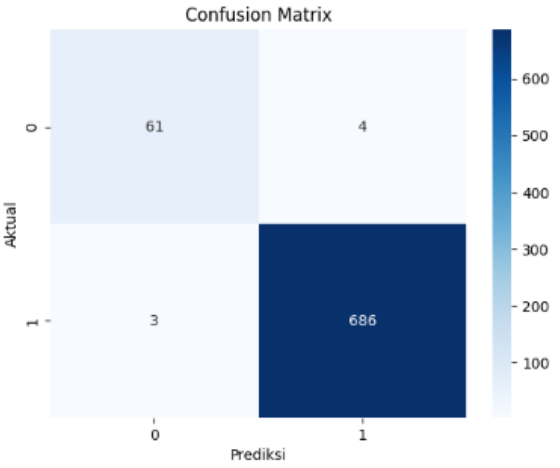
. Tabel 4. Pengujian Kedua

Split Data 70/30			
Pengujian	1	2	3
n_estimator	50	100	150
max_depth	5	10	15
min_samples_leaf	5	10	15
random_state	42	42	42
Accuracy	97%	98%	98%

Pada pengujian pertama dan kedua, diterapkan berbagai parameter yaitu (*n_estimators*) jumlah pohon keputusan diuji dengan nilai (50, 100, 150), (*max_depth*) kedalaman maksimum pohon keputusan diuji dengan nilai (5, 10, 15), (*min_samples_leaf*) jumlah minimum sampel diuji dengan nilai (5, 10, 15), serta *random_state* dengan nilai 42. Setiap kombinasi parameter ini diuji untuk mengevaluasi pengaruhnya terhadap akurasi model. Hasil pengujian menunjukkan bahwa variasi parameter menghasilkan tingkat akurasi yang beragam. Hal ini menggarisbawahi pentingnya pemilihan parameter yang tepat untuk meningkatkan kinerja model. Dengan mengoptimalkan nilai *n_estimators*, *max_depth*, dan *min_samples_leaf*, model dapat mencapai tingkat akurasi yang lebih tinggi, menunjukkan betapa signifikan dampak tuning parameter terhadap hasil akhir.

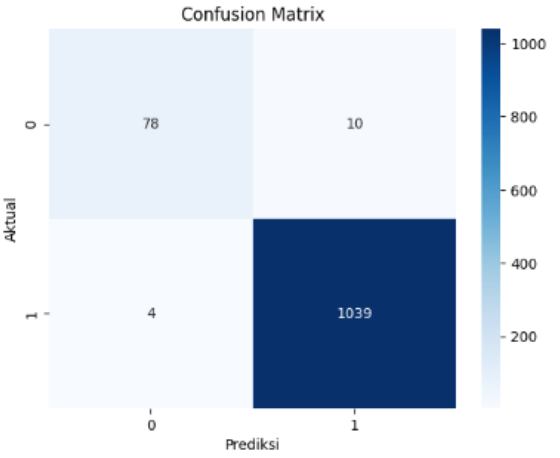
3.4. Evaluasi

Pada penelitian ini, dilakukan 6 pengujian menggunakan metode klasifikasi random forest. Pengujian ke-3 dengan pembagian data 80/20 menghasilkan akurasi terbaik. Berikut adalah confusion matrix untuk pengujian ke-3 yang ditampilkan dalam gambar.



Gambar 3. Confusion matrix pengujian ke-3

Untuk pembagian data 70/30, akurasi terbaik dicapai pada pengujian ke-6 dengan akurasi sebesar 98%. Pengujian kedua ini menggunakan parameter yang sama dengan pengujian pertama. Hasil dari pengujian ini lebih rendah dibandingkan pengujian pertama karena jumlah data pelatihan yang berkurang dan data pengujian yang bertambah. Berikut adalah confusion matrix untuk pengujian ke-6 yang ditampilkan dalam gambar 3.



Gambar 3. Confusion Matrix pengujian ke-6

4. Kesimpulan

Berdasarkan analisis yang dilakukan, dapat disimpulkan bahwa:

1. Penelitian ini mencakup 6 pengujian, dan hasil terbaik diperoleh pada pengujian pertama ke-3 dengan pembagian data 80/20, yang mencapai akurasi 99%. Ini lebih tinggi dibandingkan dengan pembagian data 70/30, yang menghasilkan akurasi 98%.
2. Perubahan nilai pada parameter *max_depth* dan *min_samples_leaf* dapat mempengaruhi akurasi hasil pada model random forest.

Daftar Rujukan

[1] D. Sartika and Y. Yupianti, "Klasifikasi Penyakit Tiroid Menggunakan Algoritma C4.5 (Studi Kasus : Rumah Sakit Umum Daerah (RSUD) Hasanuddin Damrah Manna)," *Rekayasa*, vol. 13, no. 1, pp. 71–76, 2020, doi: 10.21107/rekayasa.v13i1.5912.

[2] D. C. Yadav and S. Pal, "To generate an ensemble model for women thyroid prediction using data mining techniques," *Asian Pacific Journal of Cancer Prevention*, vol. 20, no. 4, pp. 1275–1281, 2019, doi: 10.31557/APJCP.2019.20.4.1275.

[3] S. Agustiani, A. Mustopa, A. Saryoko, W. Gata, and S. K. Wildah, "Penerapan Algoritma J48 Untuk Deteksi Penyakit Tiroid," *Paradigma - Jurnal Komputer dan Informatika*, vol. 22, no. 2, pp. 153–160, 2020, doi: 10.31294/p.v22i2.8174.

[4] D. Sartika and Yupianti, "Klasifikasi Penyakit Tiroid Menggunakan Algoritma C4 . 5," *Journal of Science and Technology*, vol. 13, no. 1, pp. 71–76, 2020.

[5] Ardiyansyah, P. A. Rahayuningsih, and R. Maulana, "Analisis Perbandingan Algoritma Klasifikasi Data Mining Untuk Dataset Blogger Dengan Rapid Miner," *Jurnal Khatulistiwa Informatika*, vol. VI, no. 1, pp. 20–28, 2018.

[6] E. A. F. Elmuna, *Optimasi metode random forest menggunakan Principal Component Analysis untuk memprediksi harga rumah*. 2023. [Online]. Available: http://etheses.uin-malang.ac.id/50422/

[7] K. Balasree and A. S. Arunachalam, "Thyroid Disease Classification Using Decision Tree and SVM," *Indian J Public Health Res Dev*, no. May, 2020, doi: 10.37506/ijphrd.v11i3.822.

[8] C. Untuk and K. Harga, "Perbandingan Kinerja Algoritma Naive Bayes dan," vol. 7, no. 1, pp. 20–24, 2020.

[9] Oon Wira Yuda, Darmawan Tuti, Lim Sheih Yee, and Susanti, "Penerapan Penerapan Data Mining Untuk Klasifikasi Kelulusan Mahasiswa Tepat Waktu Menggunakan Metode Random Forest," *SATIN - Sains dan Teknologi Informasi*, vol. 8, no. 2, pp. 122–131, 2022, doi: 10.33372/stn.v8i2.885.

[10] R. R. Andarista and A. Jananto, "Penerapan Data Mining Algoritma C4. 5 Untuk Klasifikasi Hasil Pengujian Kendaraan Bermotor," *Jurnal Tekno Kompak*, vol. 16, no. 2, pp. 29–43, 2022, [Online]. Available: https://ejurnal.teknokrat.ac.id/index.php/teknokompak/article/view/1525%0Ahttps://ejurnal.teknokrat.ac.id/index.php/teknokompak/article/download/1525/944

[11] F. M. Hana, "Klasifikasi Penderita Penyakit Diabetes Menggunakan Algoritma Decision Tree C4.5," *Jurnal SISKOM-KB (Sistem Komputer dan Kecerdasan Buatan)*, vol. 4, no. 1, pp. 32–39, 2020, doi: 10.47970/siskom-kb.v4i1.173.

[12] T. S. Lestari and D. A. N. Sirodj, "Klasifikasi Penipuan Transaksi Kartu Kredit Menggunakan Metode Random Forest," *Jurnal Riset Statistika*, vol. 1, no. 2, pp. 160–167, 2022, doi: 10.29313/jrs.v1i2.525.

[13] Suci Amaliah, M. Nusrang, and A. Aswi, "Penerapan Metode Random Forest Untuk Klasifikasi Varian Minuman Kopi di Kedai Kopi Konijiwa Bantaeng," *VARIANSI: Journal of Statistics and Its application on Teaching and Research*, vol. 4, no. 3, pp. 121–127, 2022, doi: 10.35580/variansiunm31.

[14] A. Primajaya and B. N. Sari, "Random Forest Algorithm for Prediction of Precipitation," *Indonesian Journal of Artificial Intelligence and Data Mining*, vol. 1, no. 1, p. 27, 2018, doi: 10.24014/ijaidm.v1i1.4903.

[15] M. R. Givari, M. R. Sulaeman, and Y. Umaidah, "Perbandingan Algoritma SVM, Random Forest Dan XGBoost Untuk Penentuan Persetujuan Pengajuan Kredit," *Nuansa Informatika*, vol. 16, no. 1, pp. 141–149, 2022, doi: 10.25134/nuansa.v16i1.5406.