Media Journal of General Computer Science, 2025, Vol 2, Issue 1,1-11
Published Online January 2025 in MASE (http://www.mase.or.id)
DOI: 10.62205/mjgcs.v2i1.26

Optimizing Predictions For Thyroid Disease Sufferers Using Correlation Matrix And Random Forest With Hyperparameter Tuning

Angelina
Universitas Dinamika Bangsa, Jambi, Indonesia
E-mail: angelinanaa1515@gmail.com
*Corresponding Author

Nadhea Filosofia and Riyan Arga Wijaya
Universitas Dinamika Bangsa, Jambi, Indonesia
E-mail: {nadheana25, riyanargaw}@gmail.com

Received: 28 June, 2024; Accepted: 22 July, 2024; Published: 30 January, 2025

Abstract: *Thyroid disease is one of the most common endocrine disorders, affecting the body's hormone function and balance. Symptoms can include changes in weight, fatigue, and temperature regulation issues. Although the causes are varied, thyroid disease can generally be treated with medications or medical interventions. The objective of this study is to present and optimize a predictive model for thyroid disease patients by measuring the comparison between correlation analysis of traits and the variables used, as well as evaluating the performance of the Random Forest method in optimizing predictions. One machine learning method that can be used to optimize the prediction of thyroid disease patients is Random Forest. The features used include age, gender, smoking history, radiotherapy history, and pathology characteristics, which are utilized to optimize predictions using this Random Forest algorithm. This study employs hyperparameter tuning, with the best parameters being (n_estimators)* 100 *and (max_depth)* 30*, which are then used to predict the occurrence of thyroid disease with an accuracy of* 95%*.*

Keywords: Thyroid, Random Forest, Hyperparameter Tuning, Optimizing, Prediction

I. Introduction

Kelenjar tiroid adalah massa atau benjolan yang dapat berisi cairan, yang berkembang di dalam kelenjar tiroid. Tidak semua kelainan pada kelenjar tiroid dapat terlihat atau terdeteksi secara klinis. Oleh karena itu, diperlukan pemeriksaan diagnostik tambahan untuk mengidentifikasi dan mengevaluasi kelainan pada kelenjar ini. Salah satu metode yang digunakan adalah pemindaian tiroid (*thyroid scan*), yang membantu dalam mendiagnosis dan menilai kondisi kelenjar tiroid secara lebih mendalam [1]. Data statistik mengenai nodul tiroid di Indonesia masih terbatas. Namun, berdasarkan informasi dari Kementerian Kesehatan melalui data Sistem Informasi Rumah Sakit (SIRS) pada tahun 2015, Provinsi Sumatera Selatan mencatat jumlah kasus gangguan tiroid tertinggi di Indonesia dengan mencapai 1.400 kasus. Di Provinsi Jawa Barat, terdapat sekitar 1.100 kasus gangguan tiroid, menjadikannya daerah dengan jumlah kasus tertinggi ketiga setelah Sumatera Selatan dan Jawa Tengah [2]. Gangguan tiroid dapat dikategorikan menjadi dua jenis utama yaitu hipertiroidisme dan hipotiroidisme. Hipertiroidisme adalah kondisi di mana kelenjar tiroid berfungsi terlalu aktif, menghasilkan hormon tiroid dalam jumlah berlebihan. Sebaliknya, hipotiroidisme terjadi ketika kelenjar tiroid tidak cukup aktif, mengakibatkan produksi hormon tiroid yang lebih rendah dari yang dibutuhkan tubuh [3].

Dalam era digital, machine learning telah memainkan peran penting dalam bidang medis untuk membantu mendiagnosis berbagai kondisi kesehatan, termasuk gangguan kelenjar tiroid. Salah satu metode yang dapat digunakan untuk optimalisasi prediksi pasien dengan gangguan kelenjar tiroid adalah metode Random Forest. Metode Random Forest merupakan pengembangan dari metode Classification and Regression Tree (CART) dengan menerapkan metode bootstrap aggregating dan random feature selection. Kelebihan metode ini antara lain dapat menghasilkan akurasi yang lebih tinggi, dapat mengatasi data dalam jumlah yang besar secara efisien, dan tidak terdapat pemangkasan variabel seperti pada algoritma pohon klasifikasi tunggal [4].

Metode pendukung lainnya adalah *Correlation Matrix* dan *Hyperparameter Tuning*, *Correlation Matrix* adalah alat statistik penting yang digunakan untuk mengevaluasi dan menggambarkan hubungan antara dua atau lebih variabel dalam sebuah dataset. Setiap elemen dalam matriks ini menunjukkan seberapa kuat hubungan antara sepasang variabel, serta apakah hubungan itu bersifat positif atau negatif [5].

Dalam pengembangan model *machine learning*, *tuning hyperparameter* adalah langkah yang sangat penting. Memilih kombinasi *hyperparameter* yang paling tepat dapat secara drastis meningkatkan performa model. Berbagai teknik seperti *Grid Search*, *Random Search*, *Bayesian Optimization*, dan *Hyperband* memberikan cara yang berbeda-beda untuk menentukan hyperparameter terbaik. Setiap metode ini memiliki kelebihan dan kekurangannya sendiri, memungkinkan kita untuk menyesuaikan pendekatan yang digunakan dengan kebutuhan spesifik dan batasan sumber daya yang tersedia [6]. Penelitian sebelumnya memperkirakan kambuhnya kelenjar tiroid pada besar terjadi antara usia 50 – 60 tahun, dan wanita lebih mungkin terkena tiroid dibanding pria [7].

Oleh karena itu pada penelitian ini dalam konteks kesehatan, khususnya penyakit tiroid, penelitian ini berfokus pada pendekatan optimalisasi prediksi menggunakan *Random Forest*, beserta pengujian dengan beberapa metode yang dapat memperkuat penelitian ini.

2. Method

Tahapan dari proses optimalisasi prediksi penyakit tiroid digambarkan dalam *flowchart* yaitu mengumpulkan data, *preprocessing*, penerapan metode *Random Forest - Correlation Matrix - Hyperparameter Tuning*, pengujian data, dan kesimpulan.

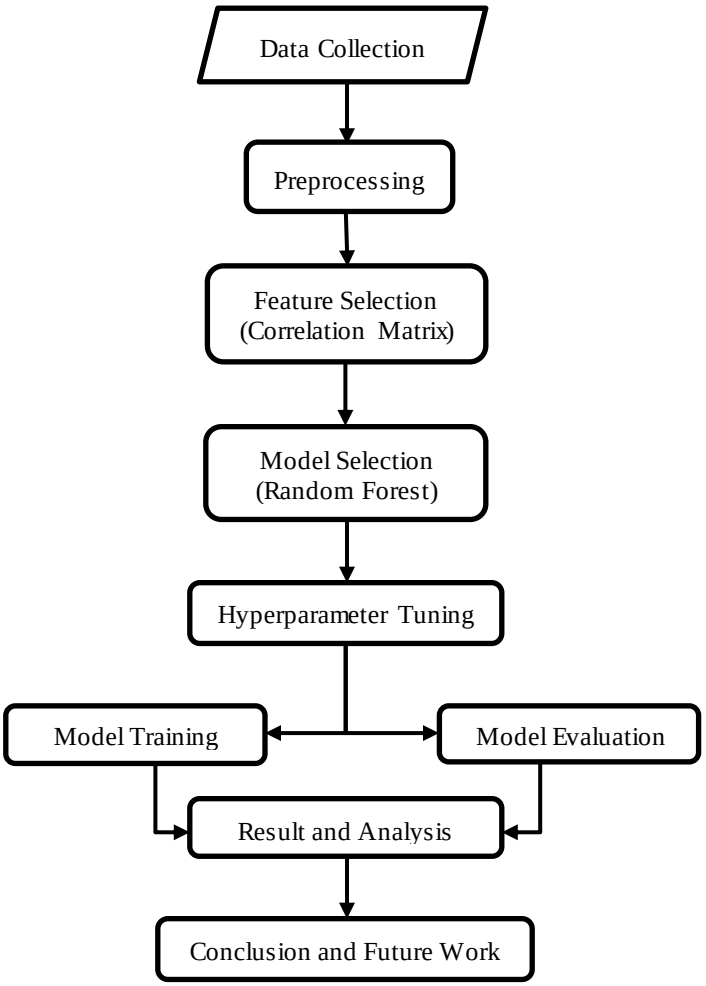


Figure 1. Flowchart Proses Penelitian

2.1 Data Preprocessing

Dataset yang digunakan dalam penelitian ini diambil dari repositori *Kaggle*. Dataset penyakit tiroid yang digunakan terdiri dari 383 sampel, dengan masing-masing sampel memiliki 17 fitur. Dataset ini mencakup berbagai catatan untuk berbagai kondisi tiroid dan termasuk kelas target yang mencakup kondisi kesehatan dan klasifikasi diagnosis. Penting untuk mengevaluasi fitur-fitur ini untuk memilih jumlah fitur yang optimal dalam klasifikasi penyakit tiroid. Fitur-fitur dalam dataset ini meliputi berbagai tipe data, yaitu *int64* untuk data kontinu dan *object* untuk data kategori. Berikut adalah rincian atribut-atribut dalam dataset:

Tabel 1. Atribut Data

Attribute	Description	Data Type
Age	Usia pasien	int64
Gender	Jenis kelamin pasien	object
Smoking	Status merokok pasien	object
Hx Smoking	Riwayat merokok pasien	object
Hx Radiotherapy	Riwayat terapi radiasi	object
Thyroid Function	Fungsi kelenjar tiroid	object
Physical Examination	Temuan dari pemeriksaan fisik	object
Adenopathy	Pembesaran kelenjar getah bening	object
Pathology	Jenis kanker tiroid berdasarkan patologi	object
Focality	Lokasi dan penyebaran nodul atau tumor tiroid	object
Risk	Penilaian risiko keseluruhan terkait gangguan tiroid	object
T	Tumor (berbasis informasi tentang tumor)	object
N	Status kelenjar getah bening (Nodal involvement)	object
M	Adanya metastasis kanker tiroid ke organ lain	object
Stage	Stadium kanker tiroid	object
Response	Respons terhadap pengobatan	object
Recurred	Apakah kanker tiroid telah kambuh setelah pengobatan	object

Berikut adalah *boxplot* yang kami buat untuk menganalisis rata-rata umur (*age*) pada berbagai tahap Tumor dan *Cancer*, Dengan visualisasi ini, kita bisa melihat bagaimana umur tersebar di setiap kategori serta mengidentifikasi keberadaan outlier yang signifikan dalam setiap kelompok.

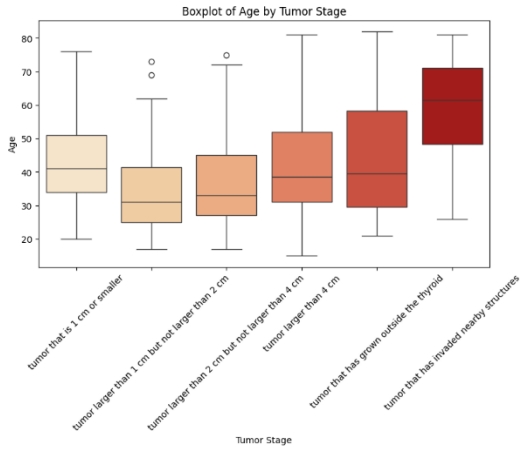


Figure 3. Boxplot Umur pada tahap Tumor

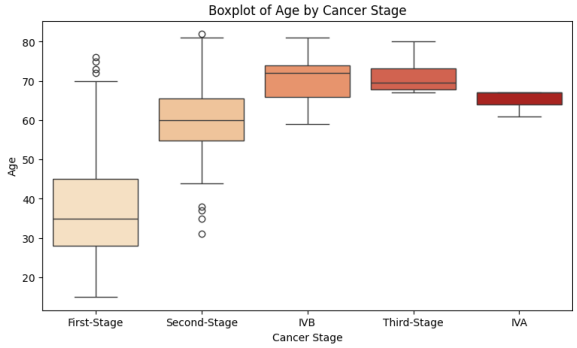


Figure 4. Boxplot Umur pada tahap Cancer

2.1.1 Correlation Matrix

Dalam mengklasifikasikan data menggunakan metode *Random Forest*, ada beberapa faktor penting yang perlu diperhatikan untuk memastikan model bekerja dengan efektif dan efisien. Dua di antaranya yang sangat penting adalah jumlah pohon (*number of trees*) dan kedalaman maksimum pohon (*maximum depth of trees*). Dan *Correlation Matrix* kami terapkan untuk mengidentifikasi hubungan dari tiap variabel. Berikut adalah *Correlation Heatmap* dari seluruh variabel / fitur yang tertera pada dataset.

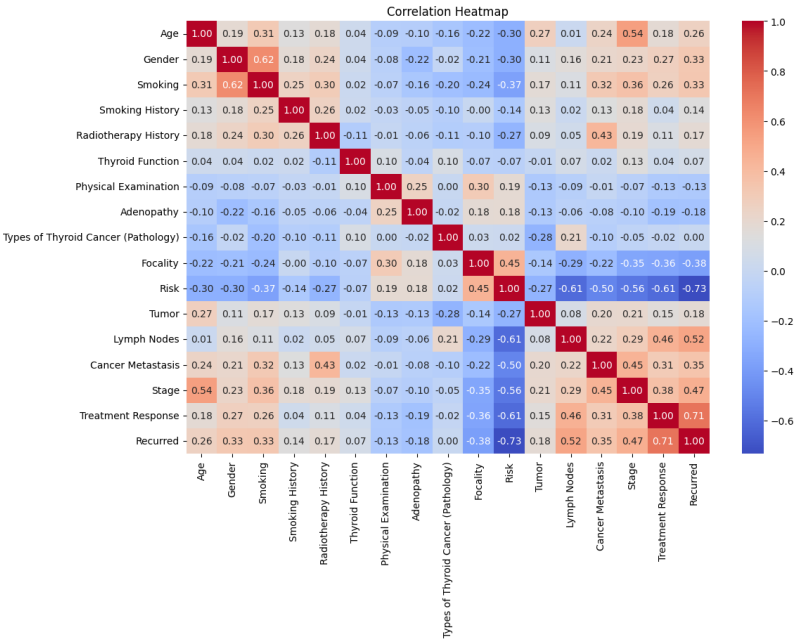


Figure 5. Correlation Heatmap Seluruh Fitur

Dan berikut ini adalah beberapa fitur yang dipilih karena sangat berkorelasi terhadap penelitian yang dilakukan , dapat dilihat pada figure 6. Beberapa diantaranya yaitu usia, jenis kelamin, riwayat merokok, riwayat *radiotherapy* dan *pathology*.

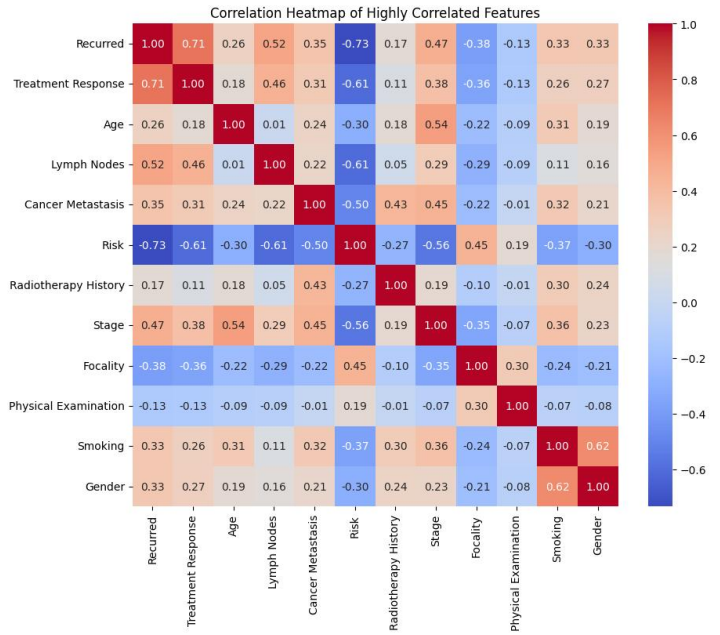


Figure 6. Correlation Heatmap Fitur yang sangat berkorelasi

Dan terakhir yang kami lakukan dalam penerapan *Correlation Matrix* adalah memilih Fitur yang memiliki Korelasi lebih dari 3%, Berikut adalah fitur yang berkorelasi lebih dari 3%.

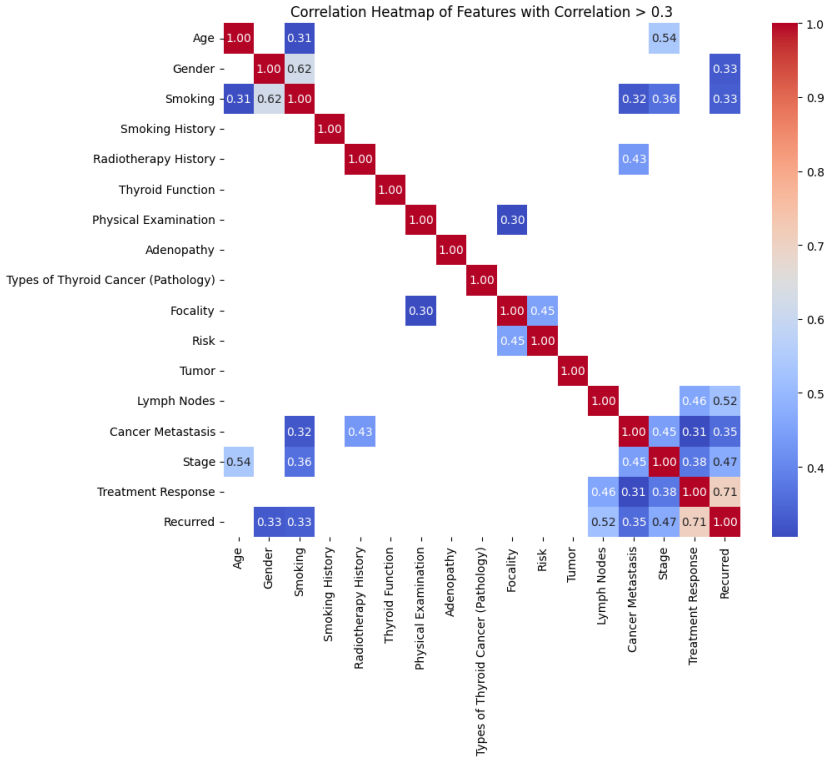


Figure 7. Correlation Heatmap Fitur dengan Korelasi > 0.3

Tabel 2. Atribut Data yang dipilih

Gender	Age	Smoking	Hx Smoking	Hx Radiotherapy	Thyroid Function	Pathology
F	27	No	No	No	Euthyroid	Micropapillary
F	34	No	Yes	No	Euthyroid	Micropapillary
F	30	No	No	No	Euthyroid	Micropapillary
F	41	No	Yes	No	Clinical Hyperthyroidism	Micropapillary
M	52	Yes	No	No	Eutyhroid	Micropapillary
F	70	No	No	No	Euthyroid	Micropapillary
M	43	No	No	No	Subclinical Hyperthyroidism	Micropapillary

2.1.2 Data Splitting

Untuk mengevaluasi kinerja model, data dibagi menjadi data pelatihan (80%) dan data pengujian (20%) menggunakan *train_test_split* dari *scikit-learn*. Parameter *test_size=0.2* menetapkan proporsi data pengujian, sementara *random_state=42* memastikan hasil yang konsisten dan dapat direproduksi. Pembagian ini memastikan model dilatih dengan data yang cukup dan dievaluasi secara adil pada data yang tidak terlihat selama pelatihan.

2.2 Method Prediction

Metode *Random Forest* adalah metode klasifikasi grup paling diterima karena memiliki fitur-fitur yang sangat baik seperti Pengukuran Pentingnya Variabel, Kesalahan *Out-of-bag*, Proksimitas, dll [10]. Saat mengklasifikasi penderita penyakit tiroid, *Random Forest* dapat digunakan untuk memprediksi persentase penyakit tiroid muncul kembali pada penderita. *Random forest* adalah suatu algoritma yang digunakan untuk klasifikasi data dalam jumlah yang besar. Penggunaan *tree* yang semakin banyak akan mempengaruhi akurasi yang akan didapatkan menjadi lebih

baik. Penentuan klasifikasi dengan *random forest* diambil berdasarkan hasil voting dari *tree* yang terbentuk. Langkah-langkah dalam melakukan perhitungan dari klasifikasi dengan metode *Random Forest* dapat dijelaskan sebagai berikut.

Pembagian yang digunakan pohon untuk mempartisi sebuah node menjadi dua turunannya dipilih dengan mempertimbangkan setiap kemungkinan pemisahan pada setiap variabel prediktor dan memilih yang "terbaik" sesuai dengan beberapa kriteria. Dalam regresi, jika nilai respons pada node adalah y1, ..., yn, kriteria pemisahan yang khas adalah rata-rata kuadrat pada node yang dinyatakan sebagai:

$$Q = \frac{1}{n}\sum_{i=1}^{n} = 1(y_i - \bar{y})^2$$

Dimana

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} = 1 y_i$$

Merupakan nilai prediksi pada simpul (rata-rata nilai respons). Dalam konteks klasifikasi di mana kelas K dilambangkan 1,..., K, kriteria pemisahan yang umum adalah indeks Gini yang dinyatakan:

$$Q = \sum_{k \neq k'}^{K} \neq \hat{P}k\hat{P}k'$$

Dimana $[\hat{P}k]$ merupakan pengamatan kelas 'k' dalam simpul yang dinyatakan:

$$\hat{P}k = \frac{1}{n}\sum_{i=1}^{n} = 1 y_i = k$$

Di dalam *Correlation Matrix*, korelasi *Pearson* adalah ukuran yang sering digunakan untuk menilai hubungan linier antara dua variabel. Nilai korelasi ini berkisar dari -1 hingga 1. Nilai +1 menunjukkan adanya hubungan positif sempurna, di mana kedua variabel bergerak dalam arah yang sama. Sebaliknya, nilai -1 menunjukkan hubungan negatif sempurna, di mana satu variabel meningkat sementara variabel lainnya menurun. Nilai 0 menunjukkan tidak adanya hubungan linier antara variabel-variabel tersebut.

Nilai korelasi ini dapat diinterpretasikan dengan panduan berikut:

- Korelasi antara 0,7 hingga 1,0 atau -0,7 hingga -1,0 dianggap kuat.
- Korelasi antara 0,4 hingga 0,7 atau -0,4 hingga -0,7 adalah moderat.
- Korelasi antara 0,1 hingga 0,4 atau -0,1 hingga -0,4 adalah lemah.
- Korelasi antara 0 dan 0,1 atau -0,1 dan 0 dianggap sangat lemah atau hampir tidak ada.

Sebagai contoh, jika kita memiliki tiga variabel X, Y, dan Z, maka matriks korelasi untuk variabel-variabel ini bisa terlihat sebagai berikut:

	X	Y	Z
X	1.0	0.8	-0.2
Y	0.8	1.0	0.5
Z	-0.2	0.5	1.0

Matriks ini menunjukkan bahwa X dan Y memiliki hubungan positif kuat dengan nilai 0,8, sementara hubungan antara X dan Z lemah dan negatif dengan nilai -0,2. Hubungan antara Y dan Z adalah positif moderat dengan nilai 0,5.

Correlation Matrix dapat dihitung dengan rumus:

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2} \cdot \sqrt{\sum(Y_i - \bar{Y})^2}}$$

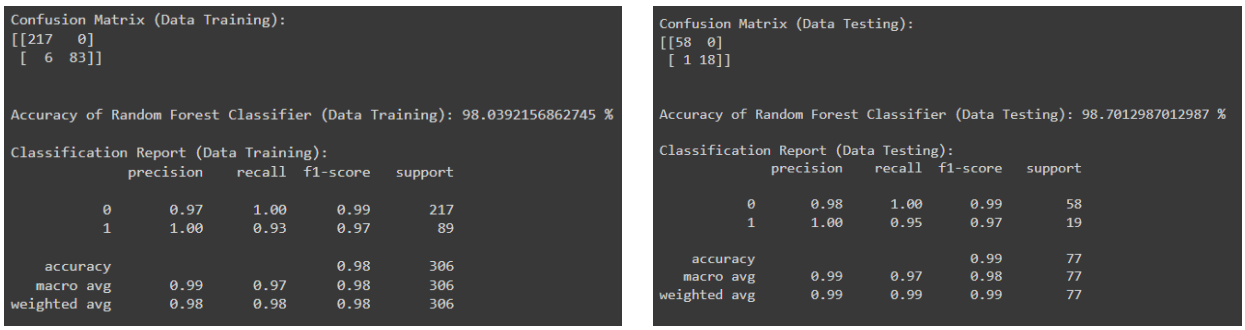
Dimana X_i dan Y_i adalah nilai-nilai individu dari variabel X dan Y, dan $\bar{X}$ dan $\bar{Y}$ adalah rata-rata dari variabel-variabel tersebut.

Menentukan *hyperparameter* yang tepat bisa sangat meningkatkan performa model. Jika *hyperparameter* tidak disetel dengan benar, model bisa mengalami *overfitting*, di mana model terlalu sesuai dengan data pelatihan dan berkinerja buruk pada data baru, atau *underfitting*, di mana model gagal menangkap pola dari data pelatihan. Oleh karena itu, *tuning hyperparameter* adalah langkah krusial untuk mencapai keseimbangan antara bias dan varians, yang penting untuk membuat model dapat bekerja dengan baik pada data yang belum pernah dilihat sebelumnya. Beberapa metode populer yang dapat digunakan pada *hyperparameter tuning* adalah *Grid Search*, *Random Search*, *Bayesian Optimization*, *Hyperband* dan *Cross Validation*.

2.3 Evaluation

Tahap training ini bertujuan untuk memanfaatkan data pelatihan dalam mengajarkan model bagaimana membuat prediksi atau melakukan klasifikasi berdasarkan pola-pola yang terdapat dalam data tersebut. Selama proses ini, model dipasangkan dengan data pelatihan, dan parameter-parameter internalnya disesuaikan agar dapat menyesuaikan diri dengan pola-pola yang ada. Tujuan utamanya adalah agar model dapat menangkap dan memahami hubungan antar variabel dalam data. Setelah pelatihan selesai, diharapkan model akan mampu memberikan prediksi yang akurat ketika dihadapkan pada data baru yang belum pernah dilihat sebelumnya.

Pada langkah testing ini, model diuji menggunakan data yang belum pernah dilihat sebelumnya selama fase pelatihan untuk menilai kemampuannya dalam melakukan generalisasi terhadap data baru. Model diuji pada kumpulan data independen yang tidak terlibat dalam proses pelatihan. Prediksi yang dihasilkan oleh model dibandingkan dengan nilai sebenarnya dalam data uji untuk mengevaluasi keakuratan dan kinerjanya. Tujuan utama dari tahap pengujian adalah untuk menilai sejauh mana model dapat membuat prediksi yang tepat dengan data yang tidak digunakan saat model dilatih.



```
Confusion Matrix (Data Training):
[[217   0]
 [  6  83]]


Accuracy of Random Forest Classifier (Data Training): 98.0392156862745 %

Classification Report (Data Training):
              precision    recall  f1-score   support

           0       0.97      1.00      0.99       217
           1       1.00      0.93      0.97        89

    accuracy                           0.98       306
   macro avg       0.99      0.97      0.98       306
weighted avg       0.98      0.98      0.98       306
```

```
Confusion Matrix (Data Testing):
[[58  0]
 [ 1 18]]


Accuracy of Random Forest Classifier (Data Testing): 98.7012987012987 %

Classification Report (Data Testing):
              precision    recall  f1-score   support

           0       0.98      1.00      0.99        58
           1       1.00      0.95      0.97        19

    accuracy                           0.99        77
   macro avg       0.99      0.97      0.98        77
weighted avg       0.99      0.99      0.99        77
```

Figure 8. Confussion Matrix Data Training dan Testing

2.4 Cross Validation

Cross-validation adalah teknik untuk mengevaluasi kinerja model dengan membagi data pelatihan menjadi beberapa *subset*, melatih model pada subset yang berbeda, dan mengukur kinerjanya pada subset yang tersisa. Dalam penelitian ini, digunakan metode *10-fold cross-validation* untuk menilai model *Random Forest*.

Dengan menggunakan *cross_val_score* dari *scikit-learn*, nilai *cross-validation* untuk model diukur dengan parameter *cv=10* yang menunjukkan bahwa data dibagi menjadi 10 lipatan. Skor akurasi diperoleh untuk setiap lipatan, yang memberikan indikasi sejauh mana model dapat generalisasi pada data yang tidak terlihat. Rata-rata skor akurasi sebesar 0.941 menunjukkan bahwa model memiliki performa yang konsisten dan akurat pada data pelatihan.

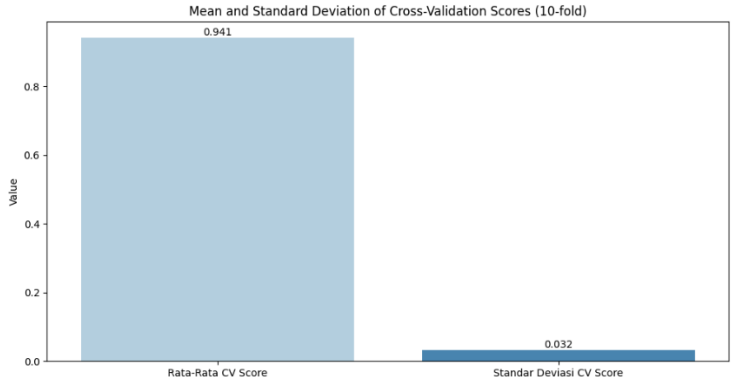


Figure 9. CV Score

2.5 Hyperparameter Tuning

Hyperparameter tuning dalam *Random Forest* melibatkan penyesuaian parameter berikut:

- *n_estimators*: Jumlah pohon dalam hutan. Nilai lebih tinggi biasanya meningkatkan akurasi tetapi juga waktu komputasi.
- *max_depth*: Kedalaman maksimum setiap pohon. Mengatur kedalaman pohon mempengaruhi kompleksitas model dan risiko *overfitting*.
- *min_samples_split*: Jumlah minimum sampel untuk membagi simpul. Mengontrol pemisahan simpul dan mencegah *overfitting*.
- *min_samples_leaf*: Jumlah minimum sampel di setiap daun. Mencegah daun simpul terlalu kecil, yang membantu mengurangi *overfitting*.
- Penggunaan *param_grid* dengan kombinasi berbagai nilai dari parameter ini memungkinkan penyesuaian model untuk mencapai performa optimal dalam klasifikas

Disini kami melakukan *hyperparameter tuning* untuk mencari kombinasi terbaik *hyperparameter* untuk model yang akan digunakan, dimana kami meletakkan jumlah pohon [100, 200, 300] dan kedalaman maksimum pohon [*None*, 10, 20, 30] pada setiap pohon. *Hyperparameter tuning* dilakukan dengan *GridSearchCV* yang merupakan bagian dari modul *scikit learn* yang bertujuan untuk melakukan validasi untuk satu atau lebih *hyperparameter* dan mengukur tingkat akurasinya secara langsung.

Tabel 3. Tabel hasil kombinasi Hyperparameter

Parameter	n_estimators	max_depth	min_samples_split	min_samples_leaf	Akurasi
84	100	30	5	1	0.957536
3	100	0	5	1	0.957536
57	100	20	5	1	0.957536
0	100	0	2	1	0.957483
27	100	10	2	1	0.957483
81	100	30	2	1	0.957483
54	100	20	2	1	0.957483
5	300	0	5	1	0.954257
86	300	30	5	1	0.954257
8	300	0	10	1	0.954257
58	200	20	5	1	0.954257
59	300	20	5	1	0.954257
62	300	20	10	1	0.954257
35	300	10	10	1	0.954257
68	300	20	5	1	0.954257
32	300	10	5	2	0.954257
31	200	10	5	1	0.954257
30	100	10	5	1	0.954257
1	200	0	2	1	0.954257
4	200	0	5	1	0.954257

Berdasarkan tabel 2, bisa diasumsikan parameter terbaik berada pada parameter 84. Akurasi yang didapatkan mungkin terlihat sama dengan 2 dibawahnya, tetapi pada sisi lain, parameter 84 lebih unggul. Keunggulan pada max_depth menjadikan parameter 84 menjadi parameter terbaik dalam *Hyperparameter Tuning*.

2.6 Learning Curve

Learning Curve adalah alat visual yang digunakan untuk menilai bagaimana performa model berubah seiring dengan jumlah data pelatihan yang digunakan. Dengan mem-*plot learning curve*, kita dapat memantau apakah model mengalami *overfitting* atau *underfitting*, dan bagaimana kinerja model meningkat dengan penambahan data pelatihan. Berikut ini adalah *Learning Curve*-nya.

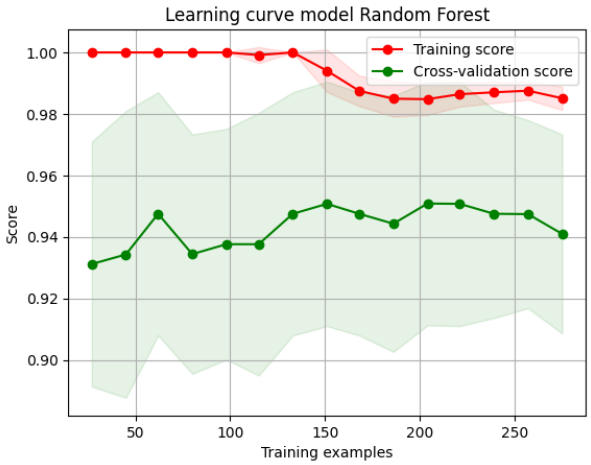


Figure 10. Learning Curve

Berikut adalah ROC curve yang merupakan pengukur kinerja model klasifikasi di berbagai *threshold*. Berdasarkan model diatas dengan parameter jumlah pohon 100 dan maksimal kedalaman pohon 30 didapatkan kurva ROC dengan nilai AUC sebesar 0.98 yang menunjukkan bahwa model memiliki kinerja yang sangat baik dalam memisahkan kelas positif dan kelas negatif. Nilai AUC mendekati 1 menunjukkan bahwa model memiliki kemampuan yang sangat baik dalam membedakan antara sampel positif dan negatif. Hal ini menandakan kinerja model dianggap sudah lebih dari cukup.

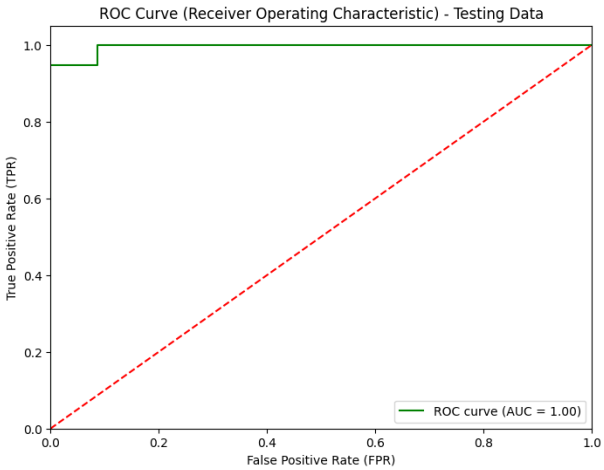


Figure 11. ROC curv

3. Results and Discussion

Model Random Forest mencapai akurasi 98.70% pada data pengujian, menunjukkan performa yang sangat baik. Namun, dalam 10-fold cross-validation, akurasi rata-rata sedikit lebih rendah di 94.09%, yang menunjukkan variasi dalam kinerja model pada subset data. Peningkatan akurasi menjadi 95.75% setelah hyperparameter tuning menegaskan bahwa penyesuaian parameter seperti jumlah pohon dan kedalaman maksimal pohon signifikan untuk meningkatkan performa model. Hasil ini menunjukkan bahwa model

Random Forest sangat efektif untuk klasifikasi penyakit tiroid, tetapi pengujian lebih lanjut dengan dataset yang lebih besar diperlukan untuk memastikan keandalannya dan kemampuannya untuk generalisasi.

Tabel 4. Hasil Akurasi tiap model

Model	Akurasi
Random Forest	0.9870
Cross Validation	0.9409
Hyperparameter Tuning	0.9575

4. Conclusion

Penelitian ini berhasil mengembangkan model prediksi gangguan kelenjar tiroid menggunakan metode *Random Forest* dengan akurasi tinggi. Proses pengembangan melibatkan seleksi fitur untuk mengidentifikasi variabel yang relevan, penggunaan *correlation matrix* untuk mengeliminasi fitur tidak signifikan, dan pemilihan *Random Forest* untuk menangani data berdimensi tinggi serta menghindari *overfitting*. *Hyperparameter tuning*, termasuk penyesuaian jumlah pohon (*n_estimators* = 100) dan kedalaman pohon (*max_depth* = 30), meningkatkan akurasi model menjadi 95%. Hasil ini menekankan potensi *besar Random Forest* dalam klasifikasi gangguan tiroid dan pentingnya teknik seleksi fitur serta *tuning* parameter dalam mengoptimalkan performa model. Penelitian lebih lanjut dengan dataset yang lebih besar dan variatif dapat memperkuat keandalan model dan memperluas aplikasinya dalam deteksi penyakit tiroid.

REFERENCES

[1] R. S. Tantika and A. Kudus, "Penggunaan Metode Support Vector Machine Klasifikasi Multiclass pada Data Pasien Penyakit Tiroid," *Bandung Conf. Ser. Stat.*, vol. 2, no. 2, pp. 159–166, 2022, doi: 10.29313/bcss.v2i2.3590.

[2] Yuyun Saputri and Meta Maulida Damayanti, "Karakteristik Pasien dengan Nodul Tiroid di Rumah Sakit X Bandung," *J. Ris. Kedokt.*, vol. 1, no. 2, pp. 71–79, 2021, doi: 10.29313/jrk.v1i2.438.

[3] 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.

[4] A. Ramadhan, B. Susetyo, and Indahwati, "Penerapan Metode Klasifikasi Random Forest Dalam Mengidentifikasi Faktor Penting Penilaian Mutu Pendidikan," *J. Pendidik. dan Kebud.*, vol. 4, no. 2, pp. 169–182, 2019, doi: 10.24832/jpnk.v4i2.1327.

[5] E. Nasti, T. H. Setiawan, H. Warianto, A. Andi, and G. Gerry, "Faktor-Faktor Yang Mempengaruhi Tingkat Kecerdasan Emosional Anak Terhadap Pelajaran Matematika Dengan Menggunakan Analisis Faktor," *J. Lebesgue J. Ilm. Pendidik. Mat. Mat. dan Stat.*, vol. 3, no. 1, pp. 44–59, 2022, doi: 10.46306/lb.v3i1.72.

[6] A. Handayani, A. Jamal, and A. A. Septiandri, "Evaluasi Tiga Jenis Algoritme Berbasis Pembelajaran Mesin untuk Klasifikasi Jenis Tumor Payudara," *JNTETI*, vol. 6, no. 4, pp. 394–403, 2017.

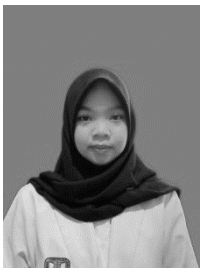
[7] M. Alnaggar, M. Handosa, T. Medhat, and M. Z. Rashad, "Thyroid Disease Multi-class Classification based on Optimized Gradient Boosting Model," *Egypt. J. Artif. Intell.*, vol. 2, no. 1, pp. 1–14, 2023, doi: 10.21608/ejai.2023.205554.1008.

[8] "Thyroid Disease Dataset," *kaggle.com*, no. https://www.kaggle.com/datasets/jainaru/thyroid-disease-data, [Online]. Available: https://www.kaggle.com/datasets/jainaru/thyroid-disease-data

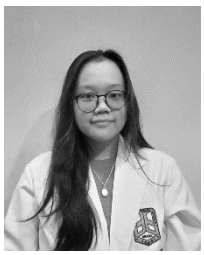
[9] A. E. Budiman and A. Widjaja, "Analisis Pengaruh Teks Preprocessing Terhadap Deteksi Plagiarisme Pada Dokumen Tugas Akhir," *J. Tek. Inform. dan Sist. Inf.*, vol. 6, no. 3, pp. 475–488, 2020, doi: 10.28932/jutisi.v6i3.2892.

[10] D. P. Sinambela, H. Naparin, M. Zulfadhilah, and N. Hidayah, "Implementasi Algoritma Decision Tree dan Random Forest dalam Prediksi Perdarahan Pascasalin," *J. Inf. dan Teknol.*, vol. 5, no. 3, pp. 58–64, 2023, doi: 10.60083/jidt.v5i3.393.

Riyan Arga Wijaya merupakan seorang yang lahir di Kota Jambi dengan tanggal lahir 11 Maret 2003. Hingga saat ini penulis masih berkuliah di Universitas Dinamika Bangsa dengan jurusan teknik informatika. Dengan ketekunan dan motivasi tinggi untuk terus belajar dan berusaha kami berhasil menyelesaikan paper ini.



Nadhea Filosofia merupakan seorang yang lahir di Jambi dengan tanggal lahir 25 Januari 2003. Hingga saat ini penulis masih berkuliah di Universitas Dinamika Bangsa dengan jurusan teknik informatika. Dengan ketekunan dan motivasi tinggi untuk terus belajar dan berusaha kami berhasil menyelesaikan paper ini.



Angelina merupakan seorang yang lahir di Jambi dengan tanggal lahir 08 Maret 2003. Hingga kini penulis masih berkuliah di Universitas Dinamika Bangsa dengan jurusan teknik informatika. Dengan ketekunan dan motivasi tinggi untuk terus belajar dan berusaha kami berhasil menyelesaikan paper ini dan berharap paper ini mampu memberikan kontribusi positif bagi dunia Pendidikan.