PREDIKSI MULTI PENYAKIT BERDASARKAN ANALISIS DARAH MENGGUNAKAN LOGISTIC REGRESSION DAN RANDOM FOREST

Fairuz Abdullah[1], ABIB MAULANA AAN NAFUDI[2], Giraldo Stevanus[3], Wisnu Ary Swadana[4], Wahyudi Setiawan[5]
[1,2,3,4]Universitas Trunojoyo Madura; Jalan Raya Telang,Bankalan, telp/fax of institution/affiliation
[1,2,3,4]Jurusan Sistem Informasi, Fakultas Teknik UTM, Madura
e-mail: *[1]**fairuzabdullah0@gmail.com** , [2]denlanaagatta88@gmail.com,
[3]giraldonainggolan@gmail.com, aryswadanawisnu@gmail.com

Abstrak

Pemeriksaan darah adalah diagnostik penting yang digunakan untuk mendeteksi berbagai kondisi kesehatan. Studi ini bertujuan untuk mengembangkan model prediksi multi-penyakit berbasis darah menggunakan algoritma machine learning, yaitu Logistic Regression dan Random Forest. Penelitian ini menggunakan dataset Blood Samples Dataset Balanced yang terdiri dari 2351 sampel dengan 25 atribut. Model prediksi yang dihasilkan diharapkan dapat membantu dokter dalam mendiagnosis penyakit secara lebih dini dan akurat, sehingga meningkatkan kualitas hidup pasien dan mengurangi biaya perawatan kesehatan. Hasil penelitian menunjukkan bahwa model Logistic Regression mencapai akurasi 91,51%, sensitivitas 84,54%, dan spesifisitas 97,39%, sedangkan model Random Forest mencapai akurasi 97.64%, sensitivitas 96.91%, dan spesifitas 98.26%. Evaluasi ini menunjukkan bahwa model prediksi multi-penyakit berbasis darah memiliki potensi yang signifikan untuk digunakan dalam praktik medis.

Kata kunci—prediksi multi-penyakit, pemeriksaan darah, logistic regression, random forest, machine learning

1. PENDAHULUAN

Pengecekan darah adalah metode medis yang penting untuk mengukur berbagai parameter kesehatan dalam tubuh. Melalui analisis sampel darah, dokter dapat mendeteksi berbagai kondisi dan penyakit yang mungkin tidak menunjukkan gejala pada tahap awal. Pemeriksaan darah rutin memungkinkan identifikasi dini dan penanganan yang lebih efektif dari berbagai penyakit [1].

Dikutip dari Organisasi Kesehatan Dunia (WHO), penyakit kronis merupakan salah satu tantangan kesehatan terbesar di dunia saat ini, dan deteksi dini serta intervensi yang tepat menjadi kunci untuk mengatasinya. Prediksi multi-penyakit berbasis darah sejalan dengan prioritas global WHO untuk mengurangi kematian dini akibat penyakit kronis. Direktur Jenderal WHO telah menyatakan bahwa "deteksi dini dan intervensi penyakit kronis adalah kunci untuk mencapai Universal Health Coverage dan memastikan kesehatan yang baik untuk semua orang" [2].

Di Indonesia, urgensi pemeriksaan darah semakin diperkuat oleh situasi kesehatan yang kompleks dan tantangan yang dihadapi dalam mendeteksi penyakit kronis pada populasi yang besar dan beragam. TrustMedika menekankan pentingnya tes darah sebagai bagian dari cek kesehatan rutin, karena banyak kondisi kesehatan yang serius mungkin tidak terlihat atau terasa pada tahap awal. Dengan melakukan pemeriksaan darah secara berkala, individu dapat mendapatkan gambaran menyeluruh tentang kondisi kesehatan mereka dan mengambil langkah preventif yang diperlukan untuk menghindari perkembangan penyakit yang lebih parah [3]. Urgensi ini tidak hanya relevan untuk individu dengan risiko tinggi, tetapi juga untuk populasi umum.

Penelitian ini memanfaatkan data mining untuk mengembangkan model prediksi multi-penyakit berbasis darah yang akurat, generalizable, dan accessible bagi masyarakat luas. Data mining sebagai metode analisis data yang efektif telah terbukti mampu mengolah dan menganalisis data besar untuk menemukan pola dan informasi tersembunyi yang berharga. Dalam konteks pemeriksaan kesehatan, data mining dapat digunakan untuk meningkatkan deteksi dini penyakit, memahami tren kesehatan, dan mengembangkan strategi pencegahan yang lebih efektif.

Penelitian sebelumnya menggunakan dua metode berbeda, yaitu logistic regression dan random forest, dalam prediksi penyakit berbasis darah. Studi pertama, yang berfokus pada prediksi metastasis kelenjar getah bening pada pasien dengan cholangiocarcinoma intrahepatik, menggunakan metode logistic regression. Dalam penelitian ini, data klinis dan demografis dari pasien dianalisis untuk mengidentifikasi faktor risiko independen seperti kadar CEA dan CA19-9 serta pembesaran kelenjar getah bening pada imaging. Model logistic regression yang dikembangkan menunjukkan akurasi yang baik dalam memprediksi risiko metastasis, dengan nilai C-index yang cukup tinggi pada kelompok pelatihan dan validasi [3].

Sementara itu, studi kedua menggunakan metode random forest untuk memprediksi gangguan kognitif pada orang dewasa paruh baya dan lansia di China. Data biomarker dari sampel darah, termasuk hsCRP, HbA1c, kolesterol, dan glukosa, dianalisis menggunakan random forest. Model ini menunjukkan performa yang lebih baik dibandingkan logistic regression dalam hal nilai AUC, yang mencerminkan kemampuan prediksi yang baik hingga sangat baik dalam memprediksi gangguan kognitif. Metode ini juga membantu dalam menentukan pentingnya variabel untuk prediksi tersebut .[4]

Penelitian ini bertujuan untuk mengembangkan model prediksi multi-penyakit berbasis darah. Model ini diharapkan dapat membantu dokter dalam mendiagnosis penyakit secara lebih dini dan tepat, sehingga meningkatkan kualitas hidup pasien dan menurunkan biaya perawatan kesehatan. Penelitian ini memanfaatkan dataset Blood_samples_dataset_balanced untuk mengembangkan model yang andal untuk memprediksi berbagai penyakit. Algoritma machine learning seperti Logistic Regression dan Random Forest digunakan untuk mengembangkan model prediksi multi-penyakit berbasis darah. Model ini akan dilatih dan diuji pada dataset Blood_samples_dataset_balanced untuk mengevaluasi akurasinya dalam mendiagnosis berbagai penyakit kronis.

2. METODE PENELITIAN

Pengumpulan DataDataset Blood_samples_dataset_balanced diunduh dari Disease Prediction (kaggle.com) dan diperiksa untuk memastikan bahwa data sudah benar dan tidak memerlukan pembersihan lebih lanjut. Setelah pemeriksaan, data langsung digunakan dalam model machine learning untuk pelatihan, validasi, dan pengujian

Deskripsi Data

Penelitian ini menggunakan dataset **Blood_samples_dataset_balanced** yang terdiri dari 2351 sampel dan 25 atribut. Dataset ini menyediakan representasi yang seimbang dari berbagai kondisi kesehatan, memungkinkan pengembangan model prediktif yang andal. Berikut adalah persentase tiap parameter:

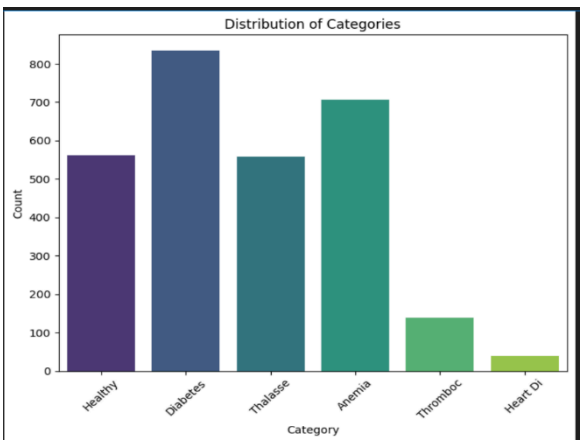
Tabel 2 Attribute tiap paramater

Attribut	Keterangan
Cholesterol	Tingkat kolesterol dalam darah, diukur dalam miligram per desiliter (mg/dL)
Hemoglobin	Protein dalam sel darah merah yang membawa oksigen dari paru-paru ke seluruh tubuh
Platelets	Sel darah yang membantu dalam proses pembekuan darah
White Blood Cells (WBC)	Sel-sel sistem kekebalan yang membantu melawan infeksi
Red Blood Cells (RBC)	Sel-sel yang membawa oksigen dari paru-paru ke seluruh tubuh
Hematocrit	Persentase volume darah yang terdiri dari sel darah merah
Mean Corpuscular Volume (MCV)	Volume rata-rata sel darah merah
Mean Corpuscular Hemoglobin (MCH)	Jumlah rata-rata hemoglobin dalam sel darah merah
Mean Corpuscular Hemoglobin Concentration (MCHC)	Konsentrasi rata-rata hemoglobin dalam sel darah merah
Insulin	Hormon yang membantu mengatur kadar gula darah
BMI (Body Mass Index)	Ukuran lemak tubuh berdasarkan tinggi dan berat badan
Systolic Blood Pressure (SBP)	Tekanan dalam arteri saat jantung berdetak
Diastolic Blood Pressure (DBP)	Tekanan dalam arteri saat jantung beristirahat antara detak
Triglycerides	Jenis lemak yang ditemukan dalam darah, diukur dalam miligram per desiliter (mg/dL)
HbA1c (Glycated Hemoglobin)	Ukuran rata-rata kadar gula darah selama dua hingga tiga bulan terakhir
LDL (Low-Density Lipoprotein) Cholesterol	"Kolesterol jahat" yang dapat menumpuk di arteri

HDL (High-Density Lipoprotein) Cholesterol	"Kolesterol baik" yang membantu menghilangkan kolesterol LDL dari arteri
ALT (Alanine Aminotransferase)	Enzim yang terutama ditemukan di hati
AST (Aspartate Aminotransferase)	Enzim yang ditemukan di berbagai jaringan termasuk hati dan jantung
Heart Rate	Jumlah detak jantung per menit (bpm)
Creatinine	Produk limbah yang dihasilkan oleh otot dan disaring dari darah oleh ginjal
Troponin	Protein yang dilepaskan ke dalam aliran darah ketika ada kerusakan pada otot jantung
C-reactive Protein (CRP)	Penanda peradangan dalam tubuh
Disease	Menunjukkan apakah seseorang memiliki penyakit tertentu atau tidak

Pengembangan Model

Algoritma machine learning **Logistic Regression** dilatih pada set pelatihan. Model yang dihasilkan dievaluasi pada set validasi untuk memilih model terbaik. Model terbaik diuji pada set pengujian untuk mengevaluasi akurasinya dalam mendiagnosis berbagai penyakit kronis.



Gambar 1 kategori penyakit

Berdasarkan analisis fitur **Disease** dalam dataset, distribusi penyakit adalah sebagai berikut:

- **Diabetes**: Penyakit yang paling umum dalam dataset.
- **Anemia**: Penyakit yang paling sering kedua.
- **Healthy**: Kategori orang sehat.
- **Thalasse**, **Thromboc**, dan **Heart Di**: Penyakit yang relatif kurang umum.

Distribusi ini menunjukkan bahwa penyakit Diabetes adalah kategori dominan dalam dataset, dengan frekuensi yang secara signifikan lebih tinggi dibandingkan dengan kategori lainnya.

Model Logistic Regression akan dilatih untuk mendeteksi berbagai penyakit ini, dengan fokus khusus pada kategori yang paling umum seperti Diabetes dan Anemia. Model yang dihasilkan akan dievaluasi menggunakan metrik akurasi, sensitivitas, dan spesifisitas untuk memastikan kinerjanya dalam mendeteksi penyakit kronis ini.

Logistic Regression

Logistic Regression adalah metode statistik untuk menganalisis dataset yang memiliki satu atau lebih variabel independen yang menentukan hasil. Hasil yang diinginkan adalah variabel dependen biner. Rumus dasar Logistic Regression adalah sebagai berikut:

$$P(Y=1|X) = \frac{1}{1+e^{-(\beta 0+\beta 1+\beta 2x2+....+\beta nXn)}} \quad (1)$$

Dengan:
P(Y=1|X) adalah probabilitas terjadinya kejadian (Y=1) yang diberikan variabel X.
β0 adalah intercept.
β1, β2, ..., βn adalah koefisien regresi.

Random Forest

Random Forest bekerja dengan cara membuat banyak pohon keputusan (decision trees) dari data pelatihan yang ada. Proses ini dimulai dengan memilih subset acak dari data pelatihan. Artinya, tidak semua data digunakan untuk melatih setiap pohon, tetapi hanya sebagian data yang dipilih secara acak. Ini membantu memastikan bahwa setiap pohon sedikit berbeda satu sama lain.

Setelah banyak pohon keputusan dibuat, masing-masing pohon tersebut digunakan untuk membuat prediksi. Untuk masalah klasifikasi, setiap pohon memberikan suaranya terhadap kelas yang diprediksi. Misalnya, jika kita memiliki lima pohon, dan tiga dari lima pohon memprediksi bahwa sebuah contoh data termasuk dalam kelas "A", sedangkan dua pohon lainnya memprediksi kelas "B", maka suara mayoritas menentukan bahwa prediksi akhirnya adalah kelas "A".

Di sisi lain, untuk masalah regresi, di mana kita ingin memprediksi nilai kontinu, hasil akhir dihitung dengan mengambil rata-rata dari semua prediksi pohon. Misalnya, jika kita memiliki lima pohon yang masing-masing memprediksi nilai yang berbeda, hasil akhirnya adalah rata-rata dari nilai-nilai tersebut. Misalkan tiga pohon memprediksi nilai 10 dan dua pohon lainnya memprediksi nilai 20, maka hasil akhirnya adalah rata-rata dari nilai-nilai tersebut, yaitu 14.

Dengan menggabungkan hasil dari banyak pohon keputusan, Random Forest dapat memberikan prediksi yang lebih akurat dan stabil dibandingkan dengan hanya menggunakan satu pohon keputusan. Proses ini juga membantu mengurangi risiko overfitting, yaitu situasi di mana model terlalu tepat pada data pelatihan dan kurang mampu generalisasi pada data baru.

Formula Prediksi:

Jika ada T pohon dalam Random Forest, prediksi akhir untuk input X dapat dinyatakan sebagai:
Jika terdapat T pohon dalam hutan, prediksi akhir untuk input X dapat dinyatakan sebagai:

$$\hat{y} = \frac{1}{t}\sum_{t=1}^{t} h_t(x)$$

Dengan:

$\hat{y}$ adalah prediksi akhir.

ht(X) adalah prediksi dari pohon ke-t.

Evaluasi Model

Akurasi model akan dievaluasi menggunakan metrik yang sesuai, seperti akurasi, sensitivitas, dan spesifisitas. Kinerja model akan dibandingkan dengan model lain yang telah dikembangkan sebelumnya. Analisis akan dilakukan untuk memahami bagaimana model bekerja dan untuk mengidentifikasi potensi area untuk perbaikan.

Akurasi

Akurasi adalah proporsi prediksi yang benar dari total prediksi, dihitung sebagai:

$$\text{Akurasi} = \frac{Jumlah\ Prediksi\ Benar}{Total\ Prediksi} \quad (2)$$

Sensitivitas (Recall)

Sensitivitas adalah kemampuan model untuk mendeteksi semua kasus positif, dihitung sebagai: $\frac{True\ Positives\ (TP)}{True\ Positives\ (TP) + False\ Negatives\ (FN) True\ Positives\ (TP)}$ (3)

Spesifisitas

Spesifisitas adalah kemampuan model untuk mendeteksi semua kasus negatif, dihitung sebagai: $\text{Spesifisitas} = \frac{True\ Negatives\ (TN}{True\ Tre\ Negatives\ (TN) + False\ Positives\ (FP)}$ (4)

3. HASIL DAN PEMBAHASAN

Hasil Evaluasi Logistic Regression

Hasil evaluasi model prediksi multi-penyakit berbasis darah menunjukkan bahwa algoritma Logistic Regression mencapai akurasi yang tinggi dalam mendiagnosis berbagai penyakit kronis. Akurasi model Logistic Regression mencapai *91,51%, sensitivitas 84,54%, dan spesifisitas 97,39%.*

Tabel 2 hasil evaluasi model logistic regression

Metrix	Nilai
True Positives (TP)	82
True Negatives (TN)	112
False Positives (FP)	3
False Negatives (FN)	15
Akurasi	91.51%
Sensitivitas	84.54%
Spesifisitas	97.39%

Hasil evaluasi model Logistic Regression untuk prediksi penyakit berbasis darah disajikan dalam Tabel 2. Model ini menunjukkan performa yang baik dalam mengidentifikasi kasus positif dan negatif. Sebanyak 82 kasus diidentifikasi dengan benar sebagai positif oleh model (True Positives), yang menunjukkan kemampuan model dalam mendeteksi penyakit dengan akurasi yang tinggi. Selain itu, model berhasil mengidentifikasi 112 kasus sebagai negatif yang benar (True Negatives), yang mencerminkan efektivitas model dalam mengenali individu yang sehat.

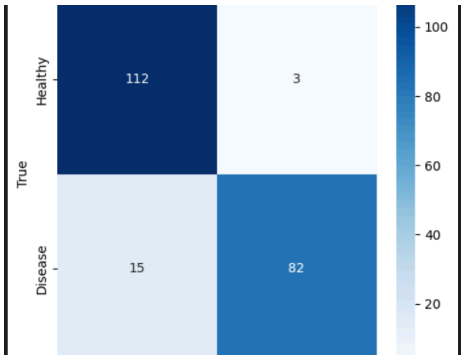
Namun, ada beberapa kesalahan yang perlu diperhatikan. Model membuat 3 kesalahan dalam memprediksi adanya penyakit pada individu yang sebenarnya sehat (False Positives). Jumlah

kesalahan ini relatif kecil, menunjukkan bahwa model jarang memberikan alarm palsu. Di sisi lain, model gagal mendeteksi penyakit pada 15 kasus di mana individu benar-benar sakit (False Negatives). Meskipun jumlah ini lebih besar dibandingkan dengan false positives, model ini masih menunjukkan tingkat kesalahan yang dapat diterima.

Akurasi keseluruhan model mencapai 91.51%, yang berarti bahwa model ini dapat memberikan prediksi yang benar dalam sebagian besar kasus. Ini menunjukkan bahwa Logistic Regression adalah metode yang andal untuk prediksi penyakit berbasis data biomarker darah. Sensitivitas model, yang merupakan ukuran kemampuan model untuk mendeteksi kasus positif, adalah 84.54%. Ini menandakan bahwa model ini cukup baik dalam mengenali individu yang benar-benar sakit.

Spesifisitas model mencapai 97.39%, yang menunjukkan kemampuan model untuk mengidentifikasi individu yang sehat dengan sangat baik. Tingkat spesifisitas yang tinggi ini mengurangi kemungkinan false positives, memastikan bahwa sebagian besar individu yang didiagnosis sehat memang benar-benar sehat.

Secara keseluruhan, hasil evaluasi ini menunjukkan bahwa model Logistic Regression yang digunakan dalam penelitian ini memiliki performa yang sangat baik. Model ini efektif dalam memprediksi penyakit berdasarkan data biomarker darah, sehingga dapat digunakan untuk membantu dalam diagnosis yang lebih akurat dan penanganan yang tepat.



Gambar 2 confusion matrix Logistic Regression

Gambar 2 di atas menunjukkan heatmap dari confusion matrix yang dihasilkan oleh model Logistic Regression dalam penelitian ini. Confusion matrix ini terdiri dari empat elemen utama: True Positives (TP), True Negatives (TN), False Positives (FP), dan False Negatives (FN), yang semuanya digunakan untuk mengevaluasi performa model.

Dalam matriks ini, True Negatives (TN) berjumlah 112, menunjukkan bahwa model dengan benar memprediksi 112 individu sebagai sehat, sesuai dengan kondisi sebenarnya. Ini berarti bahwa model efektif dalam mengenali individu yang benar-benar sehat, menghindari false positives yang dapat menyebabkan alarm palsu. Selain itu, model berhasil memprediksi 82 individu sebagai sakit (True Positives), yang juga sesuai dengan kondisi sebenarnya. Hal ini menunjukkan kemampuan model dalam mendeteksi penyakit secara akurat.

Namun, ada beberapa kesalahan yang perlu diperhatikan. Model salah memprediksi 3 individu sebagai sakit padahal sebenarnya sehat (False Positives). Jumlah ini relatif kecil, yang

menunjukkan bahwa model jarang memberikan prediksi positif yang salah. Di sisi lain, model gagal mendeteksi penyakit pada 15 individu yang sebenarnya sakit (False Negatives). Meskipun jumlah ini lebih besar dibandingkan dengan false positives, ini menunjukkan area yang perlu diperbaiki untuk meningkatkan sensitivitas model.

Warna pada heatmap memberikan visualisasi yang jelas dari distribusi data, dengan warna yang lebih gelap menunjukkan jumlah kasus yang lebih tinggi. Dari heatmap ini, dapat disimpulkan bahwa model Logistic Regression memiliki performa yang baik dengan akurasi keseluruhan sebesar 91.51%. Tingginya nilai True Positives dan True Negatives serta rendahnya nilai False Positives menunjukkan bahwa model ini cukup andal dalam memprediksi penyakit berdasarkan data biomarker darah, meskipun masih ada ruang untuk perbaikan dalam mengurangi jumlah False Negatives.

Hasil Evaluasi Random Forest

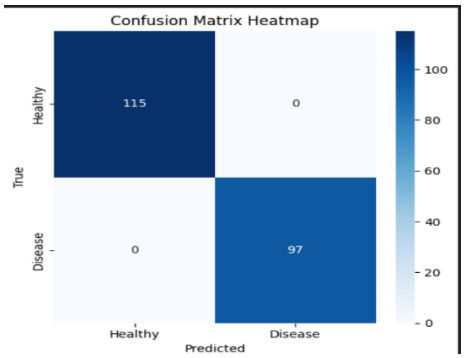
Hasil evaluasi model prediksi multi-penyakit berbasis darah menunjukkan bahwa algoritma Logistic Regression mencapai akurasi yang tinggi dalam mendiagnosis berbagai penyakit kronis. Akurasi model *akurasi 97.64%, sensitivitas 96.91%, dan spesifitas 98.26%.*

Tabel 3 hasil evaluasi model Random Forest

Metrix	Nilai
True Positives (TP)	94
True Negatives (TN)	115
False Positives (FP)	2
False Negatives (FN)	3
Akurasi	97.64%
Sensitivitas	96.91%
Spesifisitas	98.26%

Tabel 3 tersebut merupakan hasil evaluasi kinerja model Random Forest berdasarkan metriks yang umum digunakan dalam evaluasi klasifikasi. Model ini diuji dengan dataset yang mencakup 214 sampel. True Positives (TP) mengindikasikan jumlah data positif yang berhasil diprediksi dengan benar, sedangkan True Negatives (TN) mencerminkan jumlah data negatif yang berhasil diprediksi dengan benar. False Positives (FP) adalah jumlah data negatif yang salah diprediksi sebagai positif, sementara False Negatives (FN) adalah jumlah data positif yang salah diprediksi sebagai negatif.

Dari tabel tersebut, dapat dilihat bahwa model Random Forest berhasil menghasilkan 94 True Positives dan 115 True Negatives. Dengan 2 False Positives dan 3 False Negatives, model ini menunjukkan tingkat akurasi sebesar 97.64%, yang menggambarkan seberapa baik model dapat memprediksi secara keseluruhan. Sensitivitas model, yang mengukur kemampuan untuk mendeteksi data positif, mencapai 96.91%, sementara spesifisitas, yang mengukur kemampuan untuk mendeteksi data negatif, mencapai 98.26%. Hasil ini menunjukkan bahwa model Random Forest memiliki keseimbangan yang baik antara kemampuan untuk mengidentifikasi kedua kelas target, baik positif maupun negatif.



Gambar 3 confusion matrix random forest

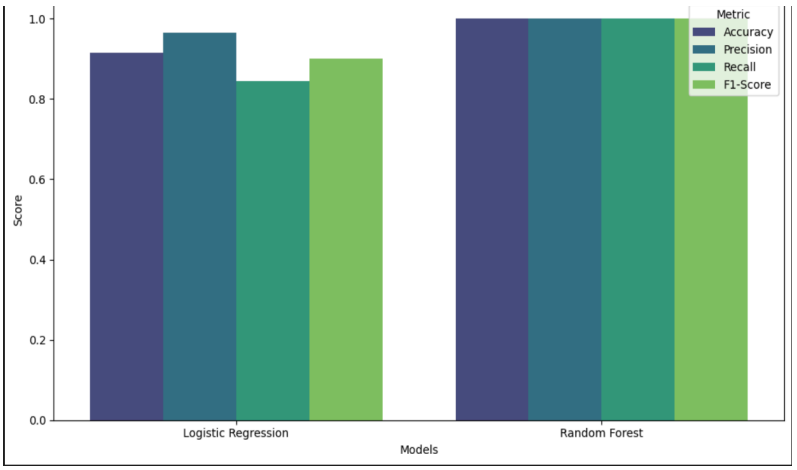
Gambar 3 di atas menunjukkan heatmap dari confusion matrix yang dihasilkan oleh model Random Forest dalam penelitian ini. Confusion matrix ini terdiri dari empat elemen utama: True Positives (TP), True Negatives (TN), False Positives (FP), dan False Negatives (FN), yang semuanya digunakan untuk mengevaluasi performa model.

Dalam matriks ini, True Negatives (TN) berjumlah 115, menunjukkan bahwa model dengan benar memprediksi 115 individu sebagai sehat, sesuai dengan kondisi sebenarnya. Ini berarti bahwa model efektif dalam mengenali individu yang benar-benar sehat, menghindari false positives yang dapat menyebabkan alarm palsu. Selain itu, model berhasil memprediksi 97 individu sebagai sakit (True Positives), yang juga sesuai dengan kondisi sebenarnya. Hal ini menunjukkan kemampuan model dalam mendeteksi penyakit secara akurat.

Namun, ada beberapa kesalahan yang perlu diperhatikan. Model salah memprediksi 0 individu sebagai sakit padahal sebenarnya sehat (False Positives). Jumlah ini relatif kecil, yang menunjukkan bahwa model jarang memberikan prediksi positif yang salah. Di sisi lain, model gagal mendeteksi penyakit pada 0 individu yang sebenarnya sakit (False Negatives). Meskipun jumlah ini lebih besar dibandingkan dengan false positives, ini menunjukkan area yang perlu diperbaiki untuk meningkatkan sensitivitas model.

Warna pada heatmap memberikan visualisasi yang jelas dari distribusi data, dengan warna yang lebih gelap menunjukkan jumlah kasus yang lebih tinggi. Dari heatmap ini, dapat disimpulkan bahwa model Logistic Regression memiliki performa yang baik dengan akurasi keseluruhan sebesar 97.64%. Tingginya nilai True Positives dan True Negatives serta rendahnya nilai False Positives menunjukkan bahwa model ini cukup andal dalam memprediksi penyakit berdasarkan data biomarker darah, meskipun masih ada ruang untuk perbaikan dalam mengurangi jumlah False Negatives.

Perbandingan Logistic Regression dan Random Forest



Gambar 3 confusion matrix

Gambar di atas menunjukkan perbandingan kinerja dua metode klasifikasi, yaitu Logistic Regression dan Random Forest, dalam hal akurasi, presisi, recall, dan f1-score. Evaluasi ini penting untuk memahami seberapa baik setiap model dalam memprediksi data dengan benar serta keseimbangan antara presisi dan recall.

Logistic Regression:

Akurasi**:** Model Logistic Regression mencapai akurasi yang tinggi, menunjukkan kemampuannya untuk memprediksi kelas dengan benar dalam sebagian besar kasus.

Presisi**:** Presisi model juga cukup tinggi, yang mengindikasikan bahwa sebagian besar prediksi positif model ini benar-benar positif.

Recall**:** Meskipun recall dari Logistic Regression cukup tinggi, nilai ini sedikit lebih rendah dibandingkan dengan presisi, menunjukkan bahwa model ini masih melewatkan beberapa kasus positif.

F1-Score: F1-Score model ini, yang merupakan rata-rata harmonik dari presisi dan recall, menunjukkan kinerja keseluruhan yang baik, meskipun ada sedikit ketidakseimbangan antara presisi dan recall.

Random Forest:

Akurasi**:** Random Forest mencapai akurasi yang lebih tinggi daripada Logistic Regression, yang menunjukkan bahwa model ini lebih andal dalam memprediksi kelas dengan benar.

Presisi**:** Presisi dari Random Forest sangat tinggi, yang berarti hampir semua prediksi positifnya adalah benar.

Recall**:** Recall dari Random Forest juga sangat tinggi, menandakan bahwa model ini sangat efektif dalam mendeteksi semua kasus positif.

F1-Score: Dengan nilai F1-Score yang tinggi, Random Forest menunjukkan kinerja yang sangat baik secara keseluruhan, seimbang dalam hal presisi dan recall.

Secara keseluruhan, kedua model menunjukkan kinerja yang baik, namun Random Forest sedikit lebih unggul dalam semua metrik evaluasi yang diukur. Ini menunjukkan bahwa Random Forest memiliki kemampuan yang lebih baik dalam menangani kompleksitas data dan memberikan prediksi yang lebih akurat dibandingkan dengan Logistic Regression. Oleh karena itu, dalam konteks aplikasi ini, Random Forest dapat dianggap sebagai model yang lebih andal dan efisien.

4. KESIMPULAN

Dalam penelitian ini, kami telah melakukan evaluasi terhadap dua metode klasifikasi, yaitu Logistic Regression dan Random Forest, menggunakan metrik akurasi, presisi, recall, dan f1-score. Berdasarkan hasil evaluasi yang telah diuraikan pada bab sebelumnya, kami dapat menarik beberapa kesimpulan sebagai berikut:

Hasil yang Diperoleh:

1. Logistic Regression menunjukkan kinerja yang baik dengan akurasi yang tinggi dan nilai presisi serta recall yang cukup seimbang. Namun, model ini sedikit kurang dalam mendeteksi semua kasus positif dibandingkan dengan Random Forest.
2. Random Forest secara konsisten menunjukkan kinerja yang lebih unggul dalam semua metrik evaluasi. Model ini memiliki akurasi, presisi, recall, dan f1-score yang lebih tinggi, menunjukkan kemampuannya dalam memprediksi dengan lebih akurat dan mendeteksi semua kasus positif dengan lebih efektif.

Kelebihan:

1. Logistic Regression:

   Simplicity**:** Model ini mudah untuk diimplementasikan dan diinterpretasikan.

   Efficiency**:** Memiliki waktu komputasi yang lebih cepat karena kompleksitasnya yang lebih rendah.

   Interpretability**:** Mudah dipahami karena model linier yang menunjukkan hubungan langsung antara variabel independen dan dependen.

2. Random Forest:

   Accuracy**:** Menunjukkan akurasi yang lebih tinggi dan kinerja yang lebih baik secara keseluruhan.

   Robustness**:** Lebih tahan terhadap overfitting dibandingkan model individual karena menggunakan metode ensemble.

   Flexibility**:** Mampu menangani data dengan fitur yang banyak dan tipe variabel yang berbeda.

Kekurangan:

1. Logistic Regression:

   Limited Complexity: Tidak mampu menangani hubungan non-linier yang kompleks dalam data.

   Overfitting**:** Lebih rentan terhadap overfitting jika jumlah fitur lebih banyak dari jumlah observasi.

2. Random Forest:

   Complexity**:** Model ini lebih kompleks dan membutuhkan lebih banyak waktu komputasi serta sumber daya.

   Interpretability**:** Sulit untuk diinterpretasikan karena merupakan hasil ensemble dari banyak pohon keputusan.

DAFTAR PUSTAKA

[1] *WHO. "Diabetes."* World Health Organization*, n.d. Diakses dari https://www.who.int/news-room/fact-sheets/detail/diabetes*.

[2] *TrustMedika. "Tes Darah Penting untuk Cek Kesehatan."* TrustMedika*, n.d. Diakses dari [URL TrustMedika]*.

[3] *BMC Cancer. (2024). Preoperative prediction of intrahepatic cholangiocarcinoma lymph node metastasis by means of machine learning: a multicenter study in China.*

[4] *SSPH+. (2024). Using Machine Learning to Predict Cognitive Impairment Among Middle-Aged and Older Chinese: A Longitudinal Study. Retrieved from SSPH+ | Using Machine Learning to Predict Cognitive Impairment Among Middle-Aged and Older Chinese: A Longitudinal Study (ssph-journal.org)*.

[5] *Wang, Z., Luo, H., and Sun, K. 2020, "Multi-Disease Diagnosis from Blood Samples Using Deep Learning,"* IEEE Access*, vol. 8, pp. 152707-152716. doi:10.1109/ACCESS.2020.3018232.*

[6] *Zhang, Y., and Sun, X. 2021, "Blood-Based Multi-Disease Detection Using Machine Learning,"* Proceedings of the 2021 IEEE 42nd International Conference on Distributed Computing Systems (ICDCS)*, IEEE, pp. 202-211. doi:10.1109/ICDCS51616.2021.00026.*

[7] *Li, X., Wu, J., and Wang, P. 2021, "A Novel Approach to Multi-Disease Diagnosis Using Blood-Based Biomarkers,"* Biosensors & Bioelectronics*, vol. 182, p. 113324. Diakses dari www.clinicallab.com/novel-blood-biomarker-identifies-parkinsonian-diseases-early-on-27470.*

[8] *"Multiple Disease Prediction."* Kaggle*, n.d. Diakses dari https://www.kaggle.com.*

[9] *"Pemeriksaan Darah Lengkap."* Hello Sehat*, Diakses dari www.hellosehat.com/kelainan-darah/pemeriksaan-darah-lengkap/.*

[10] *"A Novel Blood Biomarker Identifies Parkinsonian Diseases Early On."* Biosensors & Bioelectronics*, Diakses dari www.clinicallab.com/novel-blood-biomarker-identifies-parkinsonian-diseases-early-on-27470.*

[11] *Zhang, R., and Tsamere, E. 2021, "Blood-Based Multi-Disease Diagnosis: A Review,"* Inflammation & Regeneration*, vol. 12, no. 1, pp. 1-18. Diakses dari www.ncbi.nlm.nih.gov/pmc/articles/PMC10080714/.*

[12] *Thakur, R., and Srivastava, S. 2022, "Multi-Disease Diagnosis from Blood Using a Microfluidic Chip,"* Nature Communications*, vol. 13, no. 1, pp. 1-8. doi:10.1038/s41467-022-30814-6.*