SENASTIKA Universitas Malikussaleh

PENERAPAN NAIVE BAYES, REGRESI LOGISTIK, RANDOM FOREST, SVM, DAN KNN UNTUK PREDIKSI DIABETES

Khairul Huda[1], Munirul Ula[2]

[12]Program Studi Teknik Informatika, Universitas Malikussaleh, Lhokseumawe, Aceh
Email: [1]khairulhuda.210170119@mhs.unimal.ac.id , [2]munirulula@unimal.ac.id

Abstrak

Diabetes merupakan penyakit kronis yang prevalensinya terus meningkat. Oleh karena itu, prediksi dan deteksi dini diabetes sangat penting untuk pencegahan dan pengendalian penyakit ini. Penelitian ini bertujuan untuk membandingkan akurasi beberapa algoritma machine learning populer untuk memprediksi diabetes berdasarkan fitur-fitur diagnosis pasien. Data 768 pasien yang terdiri dari 9 fitur diagnosis diabetes digunakan dalam penelitian. Data dibagi menjadi data latih dan data uji dengan perbandingan 80:20. Beberapa algoritma machine learning yang diaplikasikan meliputi Regresi Logistik (LR), Random Forest (RF), Support Vector Machine (SVM), dan K-Nearest Neighbor (KNN). Akurasi model dievaluasi menggunakan matriks konfusi serta metrik presisi, recall, dan f1-score. Hasil penelitian menunjukkan bahwa Random Forest memberikan akurasi paling tinggi yaitu 81% diikuti SVM (80%), KNN (80%), *naive bayes* (76%) dan LR (78%). Namun, untuk kelas minoritas (pasien diabetes) LR dan RF menghasilkan recall lebih tinggi daripada SVM dan KNN. Dengan demikian kombinasi beberapa model dapat meningkatkan performa klasifikasi diabetes secara keseluruhan.

Kata Kunci: Prediksi Diabetes, Regresi Logistik, Random Forest, Support Vector Machine, K-Nearest Neighbor dan *Naïve Bayes*.

1. PENDAHULUAN

1.1 Latar Belakang

Diabetes mellitus (DM) telah berkembang menjadi permasalahan kesehatan global dengan prevalensi yang menunjukkan peningkatan signifikan setiap tahun. Data dari International Diabetes Federation (IDF) menunjukkan bahwa jumlah penderita diabetes di dunia mencapai 463 juta orang pada tahun 2019 dan diproyeksikan akan meningkat menjadi 700 juta pada tahun 2045 jika tidak ada intervensi yang efektif. Menghadapi tantangan ini, berbagai studi telah dilakukan untuk mengembangkan model prediksi risiko diabetes yang akurat, dengan tujuan memungkinkan deteksi dini dan implementasi strategi pencegahan yang tepat waktu. Secara historis, regresi logistik telah menjadi metode yang paling sering digunakan untuk memprediksi diabetes berdasarkan data diagnostik pasien. Namun, perkembangan terkini dalam bidang machine learning telah memunculkan algoritma yang lebih canggih seperti random forest, support vector machine (SVM), K-nearest neighbor (KNN), dan Naive Bayes, yang menunjukkan tingkat akurasi yang lebih tinggi dalam berbagai studi komparatif.

Beberapa penelitian telah dilakukan untuk membandingkan efektivitas berbagai metode klasifikasi dalam konteks prediksi diabetes. Salah satu studi membandingkan performa naive Bayes classifier, decision tree, dan KNN, dengan hasil menunjukkan bahwa KNN mencapai tingkat akurasi tertinggi sebesar 81%. Studi lain yang lebih komprehensif membandingkan akurasi random forest, KNN, dan deep neural network untuk klasifikasi diabetes, dengan random forest mengungguli metode lainnya dengan akurasi mencapai 97,38%. Penelitian ini bertujuan untuk melakukan evaluasi komparatif terhadap performa algoritma regresi logistik, random forest, SVM, KNN, dan Naive Bayes dalam memprediksi diabetes mellitus tipe 2, menggunakan dataset diagnostik yang identik. Tujuan utama dari studi ini adalah untuk menguji hipotesis bahwa dengan menggunakan dataset dan fitur seleksi yang sama, algoritma machine learning modern seperti random forest, SVM, KNN, dan Naive Bayes akan menunjukkan performa superior dibandingkan dengan regresi logistik yang telah lama menjadi standar dalam prediksi diabetes.

Hasil dari penelitian ini diharapkan dapat memberikan landasan ilmiah yang kuat untuk pemilihan algoritma optimal dalam konteks prediksi diabetes pada era kontemporer, serta berkontribusi pada pengembangan sistem deteksi dini yang lebih akurat dan efisien. Implikasi dari peningkatan akurasi prediksi ini dapat sangat signifikan dalam upaya global untuk mengatasi epidemi diabetes, memungkinkan intervensi yang lebih tepat sasaran dan penghematan sumber daya kesehatan yang substansial.

1.2 Tujuan Penelitian

Berdasarkan latar belakang sebelumnya, tujuan dari penelitian ini adalah :

1. Melakukan evaluasi komparatif terhadap kinerja lima algoritma pembelajaran mesin - regresi logistik, random forest, support vector machine (SVM), K-nearest neighbor (KNN), dan Naive Bayes - dalam memprediksi diabetes melitus tipe 2, menggunakan kumpulan data diagnostik yang identik.
2. Menguji hipotesis bahwa algoritma pembelajaran mesin modern (random forest, SVM, KNN, dan Naive Bayes) memiliki tingkat akurasi yang lebih tinggi dibandingkan dengan metode regresi logistik konvensional dalam konteks prediksi diabetes, dengan menggunakan kumpulan data dan pemilihan fitur yang sama.
3. Mengidentifikasi algoritma prediksi diabetes yang paling akurat dan efisien di antara kelima metode yang diuji, untuk memberikan justifikasi ilmiah dalam pemilihan algoritma optimal untuk deteksi dini dan penatalaksanaan diabetes, serta berkontribusi pada pengembangan sistem prediksi yang lebih efektif dalam upaya menanggulangi epidemi diabetes global.

2. STUDI LITERATUR

2.1 Diabetes

Diabetes adalah penyakit metabolisme kronis yang ditandai dengan peningkatan kadar gula darah (atau gula darah) yang, dari waktu ke waktu, dapat merusak organ dalam lainnya, seperti jantung, pembuluh darah, mata, ginjal, dan saraf. Ada dua jenis diabetes, diabetes tipe 1 atau juga dikenal sebagai diabetes remaja adalah penyakit kronis di mana pankreas menghasilkan sedikit atau tidak ada insulin. Biasanya menyerang orang dewasa, diabetes tipe 2 adalah ketika tubuh menjadi resisten terhadap insulin atau tidak dapat memproduksi cukup insulin. Sekitar 422 juta orang di dunia menderita diabetes(World Health Organization, 2022). Banyak penderita diabetes yang tidak menyadari bahwa dirinya mengidap diabetes karena gejala tersebut seringkali dianggap normal. Misalnya gejala diabetes adalah haus, sering buang air kecil, berat badan turun drastis, massa otot kurang, dll [1].

2.2 Regresi Logistik

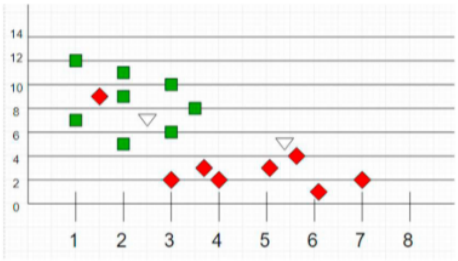
Regresi logistik adalah semacam pembelajaran terawasi yang memperkirakan hubungan antara ketergantungan biner
variabel dan setidaknya satu variabel independen dengan mengevaluasi probabilitas dengan bantuan fungsi sigmoid. Di dalam bertentangan dengan namanya, regresi logistik tidak digunakan untuk masalah regresi, melainkan merupakan jenis pembelajaran mesin masalah klasifikasi dimana variabel terikatnya dikotomis (0/1, -1/1, benar/salah) dan variabel bebas dapat tingkat binominal, ordinal, interval atau rasio [2].

$$y = \frac{1}{1 + e^{-x}} \quad (1)$$

Dimana y adalah keluaran yang merupakan hasil penjumlahan tertimbang dari variabel masukan x. Jika keluarannya lebih dari 0,5, keluarannya adalah 1, jika tidak, keluarannya adalah 0

2.3 *K-Nearest Neighbor Classifier*

Metode *K-Nearest Neighbor* (KNN) dapat digunakan untuk menyelesaikan masalah yang berkaitan dengan regresi dan klasifikasi, meskipun secara umum metode ini digunakan untuk menyelesaikan masalah klasifikasi dalam bisnis. Keuntungan utamanya adalah kesederhanaan terjemahan dan waktu komputasi yang rendah. Pada gambar 1, titik (2.5, 7) dan (5.5, 4.5) akan dialokasikan ke salah satu cluster. KNN menggunakan fungsi jarak Euclidean untuk mencari jarak antara titik data yang ada dan titik data baru. Jadi, (2.5, 7) akan menjadi milik cluster hijau, sedangkan (5.5, 4.5) akan menjadi milik cluster merah [2].

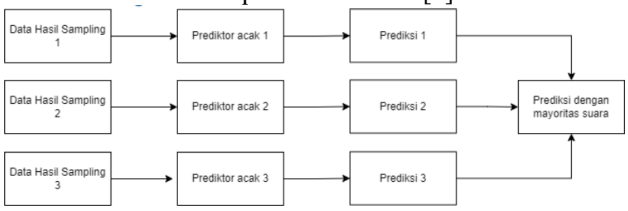


Gambar 1. Contoh KNN

2.4 Random Forest

Algorime Random Forest Classifier adalah pengembangan dari model Algoritme Decision Tree, dimana setiap pohon fikiran dilatih dengan sampel individu. Model Random Forest Classifier yang menghasilkan banyak tree dan dengan cara yang sama. Seiring bertambahnya data, maka tree ikut berkembang. Random Forest Classifier

merupakan algoritme dengan membuat pohon klasifikasi dan regresi yang nodenya dipisahkan oleh algoritme yang dioptimasi sebagai fungsi untuk meminimalkan squared-error loss [3].



Gambar 2. Alur perhitungan Random Forest Classifer

Random Forest dibangun dengan menggunakan pemilihan atribut secara acak. Metode CART (*Clasification and Regresion Tree*) digunakan untuk membuat pohon keputusan, sehingga pohon keputusan tersebut tumbuh mencapai ukuran maksimum dan tidak dipangkas sehingga dihasilkan kumpulan pohon yang kemudian disebut forest. Waktu komputasi yang dibutuhkan oleh *Algoritme Random Forest Classifier* bekerja mengklasifikasi adalah :

$$T\sqrt{MNlog(N)} \tag{2}$$

Nilai T merupakan banyaknya pohon, M adalah banyaknya peubah yang digunakan pada sub sample, N adalah banyaknya pengujian [3].

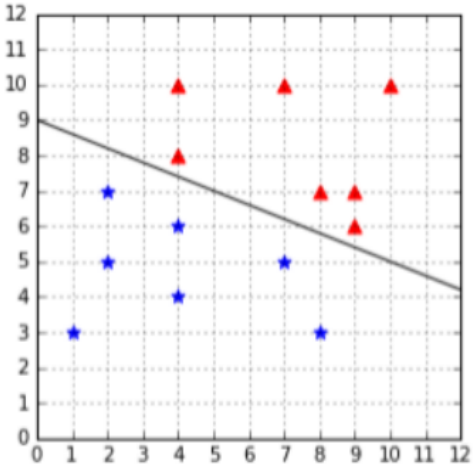
2.5 Support Vector Machine

Support Vector Machine merupakan sistem pembelajaran yang menggunakan hipotesis berupa fungsi-fungsi linear dalam sebuah fitur yang berdimensi tinggi dan dilatih dengan menggunakan algoritma pembelajaran yang didasarkan dengan teori optimasi. Support Vector Machine diperkenalkan pertama kali oleh Vapnik pada tahun 1992 sebagai rangkaian harmonis konsep konsep unggulan dalam bidang pattern recognition. Tingkat akurasi pada model yang akan dihasilkan oleh proses peralihan pada svm sangat bergantung pada fungsi kernel dan parameter yang digunakan. Berdasarkan dengan karakteristiknya metode SVM dibagi menjadi dua yaitu linear dan non linear, SVM linear merupakan data yang dipisahkan secara linear yaitu memsiahkan dua kelas pada hyperplane dengan soft margin.

Sedangkan non linear yaitu merupakan fungsi dari kernel trick terhadap ruang yang berdimensi tinggi [4].

Klasifikasi ini dipilih selama pelatihan sebagai hyperplane unik yang memisahkan instance positif yang diketahui dari instance negative, dalam kalsifikasi SVM memiliki keungulan penting dalam pendekatan teorinya yang dibenarkan atas masalah overfitting yang memungkinkannya bekerja dengan baik [4].

Support Vector Machine menggunakan 2 titik (vector) yang selanjutnya dua titik tersebut akan membentuk garis pembatas (sisi pembatas jika 3 dimensi atau lebih) garis pembatas yang dibentuk dari dau buah vector ini disebut hyperplane [4].



Gambar 3. Hyperplane memisahkan dua kelas

Dua titik yang menjadi patokan hyperplane disebut dengan support vector. Dapat dilihat bahwa memiliki dua kelompok data yang disebut klasifikasi, kemudian tugas SVM adalah membagi dua kelompok ini sebaik mungkin atau menentukan hyperplane terbaik, pembagian dimana garis batasnya dapat memisahkan dua kelompok dengan jarak terjauh antara titik terluar di masing masing kelompok dengan garis pembatas itu sendiri. Permasalahan non linear dapat diatasi dengan memodifikasi trick kernel ke dalam SVM yang akan menjadi pemisah kelas atau

hyperplane menjadi dua kelas didalam ruang vector dalam penelitian ini kernel yang akan digunakan adalah kernel linear [4]. Seperti yang dapat dilihat persamaannya pada Tabel 1 di bawah ini.

Tabel 1. Rumus Kernel

Jenis Kernel	Model
Linear	$K(x.x') = x.x'$
Polynomial	$K(x.x') = (x.x' + c)'$
RBF Gaussian	$K(x.x') = \exp(-\gamma\|\|x - x'\|\|^2)$
Siqmoid	$K(x.x') = \tanh(\alpha x.x' + \beta)$

2.6 Naive Bayes

Pengklasifikasi Bayesian adalah pengklasifikasi statistik, dan dioperasikan berdasarkan teorema Bayes, mengklasifikasikan data ke dalam kategori yang telah ditentukan menggunakan probabilitas bersyarat. Probabilitas bersyarat dapat dipahami sebagai peluang terjadinya suatu peristiwa jika peristiwa lain telah terjadi. Aturan Bayesian adalah pendekatan yang digunakan untuk memperkirakan kemungkinan suatu atribut diberikan kumpulan data sebagai masukan. Istilah "naif" pada nama algoritme mengacu pada asumsi bahwa setiap nilai atribut bersifat independen [5].

Naive Bayes (NB) dianggap sebagai algoritma deskriptif dan prediktif. Probabilitasnya bersifat deskriptif dan kemudian digunakan untuk memprediksi kategori data yang tidak terlatih. Cara ini mempunyai beberapa kelebihan, sebagai berikut. Pertama-tama, mudah digunakan. Kedua, jumlah data pelatihan yang dibutuhkan NB untuk klasifikasi belum tentu besar. Selain itu, meskipun pengklasifikasi NB dirancang secara naif dan asumsinya tampaknya terlalu sederhana, pengklasifikasi NB berfungsi dengan baik dalam sejumlah situasi dunia nyata yang rumit [5].

3. DATA DAN METODOLOGI

3.1 Eksplorasi dan Pra-Pemrosesan Dataset

Meskipun sekarang sudah ada dataset diabetes yang lebih besar dan lebih kompleks, dataset Diabetes Pima Indian tetap menjadi tolok ukur untuk penelitian klasifikasi diabetes. Dengan adanya variabel hasil biner, dataset ini secara alami cocok untuk pembelajaran terawasi dan, khususnya, regresi logistik. Namun, berbagai algoritma Machine Learning telah digunakan untuk menghasilkan model klasifikasi berdasarkan dataset ini agar tidak terbatas pada satu jenis model.

Dalam penelitian ini, fokus kami adalah menganalisis Dataset Pima Indian dengan algoritma canggih untuk bekerja dengan Internet of Medical Things (IoMT) secara efektif. Dataset ini diunduh dari Kaggle (https://www.kaggle.com/uciml/pima-indians-diabetes) dan tersedia melalui Lisensi CC0: Domain Publik serta telah di-anonimkan dengan baik dan tidak mengandung fitur identifikasi subjek pasien. Seperti yang terlihat di Tabel 1, dataset mencatat delapan karakteristik penyebab dan klasifikasinya yang sesuai. Dataset ini memiliki 9 kolom dan 768 baris (500 non-diabetik dan 268 diabetik). Variabel hasil biner klasifikasi mengambil nilai (0 atau 1), di mana 0 menunjukkan hasil tes negatif untuk diabetes, dan 1 mengimplikasikan hasil tes positif.

Dataset ini tidak memiliki nilai null atau nilai yang hilang. Namun, menurut pengetahuan domain. ada nilai yang tidak konsisten untuk atribut: konsentrasi glukosa (Gluc), tekanan darah (BP), ketebalan lipatan kulit (Skin), insulin, dan BMI, di mana nilai nol tidak berada dalam rentang normal dan oleh karena itu tidak akurat.

Terdapat perbedaan yang nyata dalam kinerja dan efisiensi model klasifikasi prediksi tergantung pada metodologi pra-pemrosesan. Oleh karena itu, dalam putaran pertama eksperimen, dilakukan pemrosesan minimal. Namun, dalam putaran kedua, algoritma seleksi fitur diterapkan.

Karena tidak ada nilai yang hilang atau null, hanya satu teknik pra-pemrosesan data yang diterapkan dalam putaran pertama. Ini adalah menggantikan nilai median pada fitur yang memiliki nilai nol yang tidak valid.

Algoritma seperti logistic regression, model Naive Bayes, dan logistic regression tidak terlalu sensitif terhadap data yang tidak ternormalisasi, jadi tidak ada penskalaan yang dilakukan untuk menjaga pengujian tetap serupa untuk ketiga model pembelajaran mesin tersebut. Dengan hanya delapan fitur, mungkin terlihat kontra-intuitif untuk mengurangi fitur lebih lanjut, tetapi ini dapat mengurangi beberapa noise dalam klasifikasi dan menyorot pengelompokan halus yang disintesis dengan menggabungkan kelas yang ada.

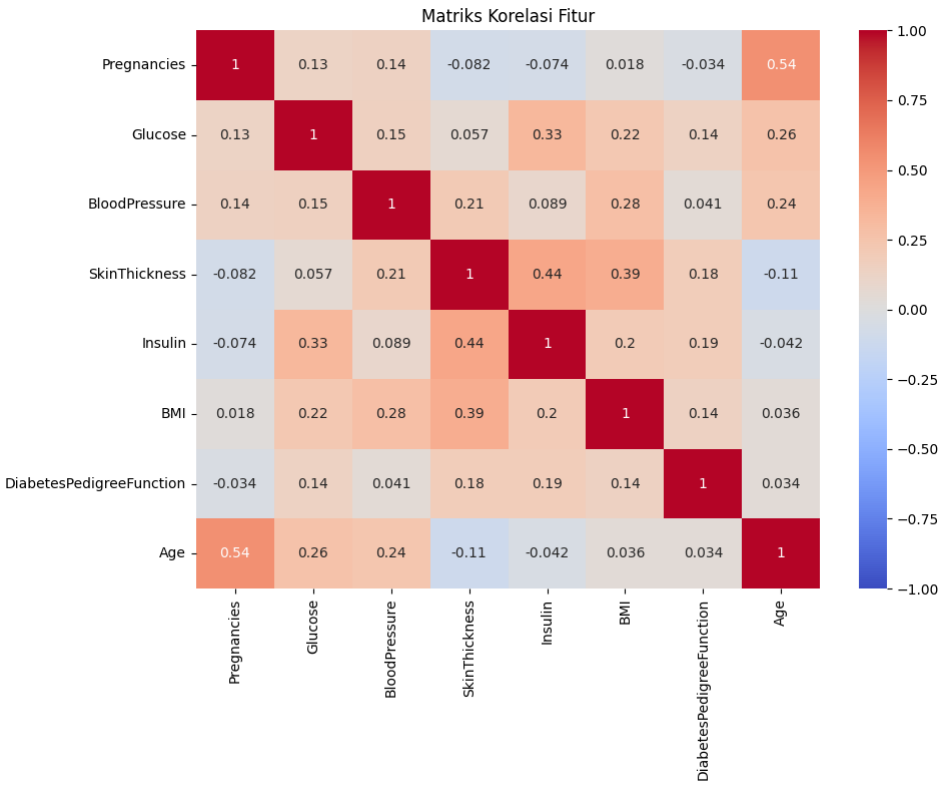
Tabel 2. Gambaran Umum Dataset Diabetes Pima Indian

Feature	Description	Data type	Range
Preg	Number of times pregnant	Numeric	[0, 17]
Gluc	Plasma glucose concentration at 2 Hours in an oral glucose tolerance test (GTIT)	Numeric	[0, 199]
BP	Diastolic Blood Pressure (mm Hg)	Numeric	[0, 122]
Skin	Triceps skin fold thickness (mm)	Numeric	[0, 99]
Insulin	2-Hour Serum insulin (µh/ml)	Numeric	[0, 846]
BMI	Body mass index [weight in kg/(Height in m)]	Numeric	[0, 67.1]
DPF	Diabetes pedigree function	Numeric	[0.078, 2.42]
Age	Age (years)	Numeric	[21, 81]
Outcome	Binary value indicating non-diabetic /diabetic	Factor	[0,1]

Tabel 3. Ringkasan Statistik Dataset Diabetes Pima Indian

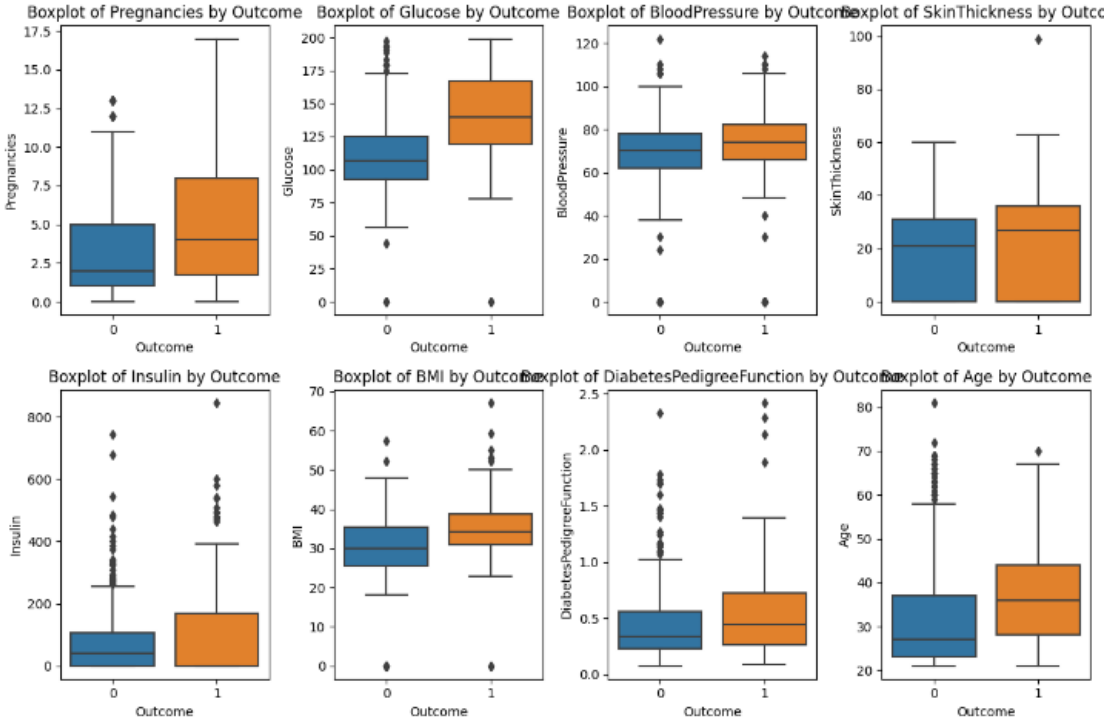
Features	Preg	Gluc	BP	Skin	Insulin	BMI	DPF	Age
Min.	0.000	0.0	0.00	0.00	0.0	0.00	0.0780	21.00
1st Qu.	1.000	99.0	62.00	0.00	0.0	27.30	0.2437	24.00
Median	3.000	117.0	72.00	23.00	30.5	32.00	0.3725	29.00
Mean	3.845	120.9	69.11	20.54	79.8	31.99	0.4719	33.24
3rd Qu.	6.000	140.2	80.00	32.00	127.2	36.60	0.6262	41.00
Max	17.000	199.0	122.00	99.00	846.0	67.10	2.4200	81.00

Matriks scatterplot berguna untuk mengidentifikasi hubungan berpasangan dari fitur-fitur secara awal. Jika titik-titik tersebar, itu berarti tidak ada hubungan yang jelas, sementara jika titik-titik disusun secara kasar dalam garis lurus, itu berarti mereka berhubungan secara linear. Saat merujuk pada matriks scatterplot dalam Gambar 4, fitur-fitur yang paling erat berkorelasi/proporsional termasuk [Kehamilan dan Usia], [Ketebalan Kulit dan IMT], dan [Glukosa dan Insulin] karena gambar scatterplot mereka semua menunjukkan korelasi positif.



Gambar 4. Matriks Korelasi fitur

Seperti yang terlihat dalam Gambar 4, terdapat pencilan (outliers) dalam fitur DPF, Usia, Insulin, Glukosa, IMT, dan Tekanan Darah, yang mungkin disebabkan oleh faktor-faktor lain yang mendasarinya. Akan lebih baik untuk menstandarisasi data guna menghindari dampak buruk dari pencilan-pencilan tersebut. Dataset ini tidak terlalu besar, sehingga lebih baik menghindari penghapusan baris-baris secara tidak perlu.



Gambar 5. Diagram kotak dan garis menggambarkan distribusi fitur untuk setiap kelas hasil.

3.2 Metode

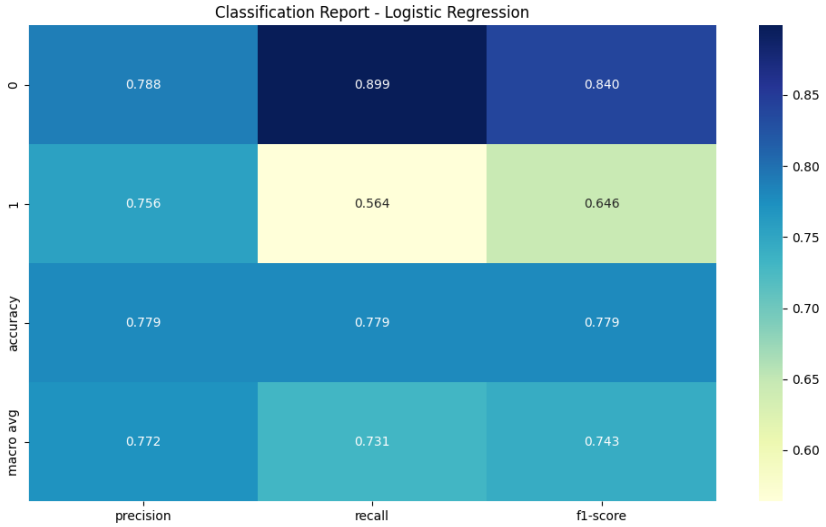
Terkait dengan klasifikasi dan prediksi diabetes serta penyakit tidak menular lainnya, model pembelajaran mesin (ML) dan pembelajaran mendalam (DL) telah menjadi area penelitian penting selama bertahun-tahun. Banyak alat dan model telah diajukan untuk membantu menyelesaikan masalah prediksi diagnosis, termasuk Convolutional Neural Networks (CNN), Artificial Neural Networks (ANN), dan model pembelajaran mesin gabungan atau hibrida. Metode evaluasi umum untuk kinerja model termasuk akurasi, presisi, sensitivitas, spesifisitas, F-measure (F-score), dan Mean Square Error (MSE), serta perbandingan kinerja pada data yang telah di-preproses dan tidak di-preproses.

Explainable AI (XAI) adalah konsep dalam kecerdasan buatan di mana keputusan yang diambil oleh model pembelajaran mesin dapat dipahami oleh pengguna. Konsep interpretabilitas merujuk pada kemampuan untuk mengamati sebab dan akibat dalam model tersebut. Interpretabilitas model dapat bersifat intrinsik, seperti pada pohon keputusan. Namun, interpretabilitas juga dapat diperkenalkan ke model (post-hoc) dengan menerapkan fungsi pada model yang telah di-pretrain untuk menghasilkan penjelasan. Terkait penyakit tidak menular, belum banyak penelitian yang dilakukan dalam pemeriksaan model pembelajaran mesin yang dapat dijelaskan. Perlu dicatat bahwa model yang dapat diinterpretasikan tidak selalu dapat dijelaskan sejauh di mana pikiran manusia sepenuhnya memahami langkah-langkah yang terjadi untuk mencapai keputusan yang diambil oleh model pembelajaran mesin.

Plot ringkasan SHAP menunjukkan pentingnya fitur yang diurutkan secara menurun yang ditandai oleh sumbu y serta efek pada bagaimana nilai fitur terkait dengan prediksi yang ditandai oleh sumbu x, yang pada gilirannya dapat digunakan untuk menginterpretasikan korelasi antara fitur dan hasil. Kita dapat menggunakan interpretasi post-hoc untuk mengidentifikasi apakah model yang telah kita latih telah dengan akurat menangkap detail proses pengambilan keputusan dunia nyata dari dataset. Selain itu, dapat digunakan untuk menyoroti bias dan kesalahan dalam model pembelajaran mesin. Dalam makalah ini, tujuannya adalah menggunakan metode kecerdasan buatan yang dapat diinterpretasikan untuk membuat model kita jelas dan dapat dimengerti oleh pengguna akhir dari dua aspek. Aspek pertama adalah menilai fitur mana yang penting, dan detailnya.

4. HASIL DAN PEMBAHASAN

4.1 Logistic Regression



Gambar 6. Classification Report Logistic Regression

- Precision

Precision untuk kelas 0 adalah sebesar 0.79, artinya model dapat memprediksi 79% data pada kelas 0 dengan benar, Precision untuk kelas 1 adalah sebesar 0.76, artinya model dapat memprediksi 76% data pada kelas 1 dengan benar. Rata-rata precision adalah sebesar 0.77, artinya model dapat memprediksi rata-rata 77% data dengan benar

- Recall

Recall untuk kelas 0 adalah sebesar 0.90, artinya model dapat menghasilkan 90% prediksi yang tepat pada kelas 0. Recall untuk kelas 1 adalah sebesar 0.56, artinya model dapat menghasilkan 56% prediksi yang tepat pada kelas 1. Rata-rata recall adalah sebesar 0.73, artinya model dapat menghasilkan rata-rata 73% prediksi yang tepat

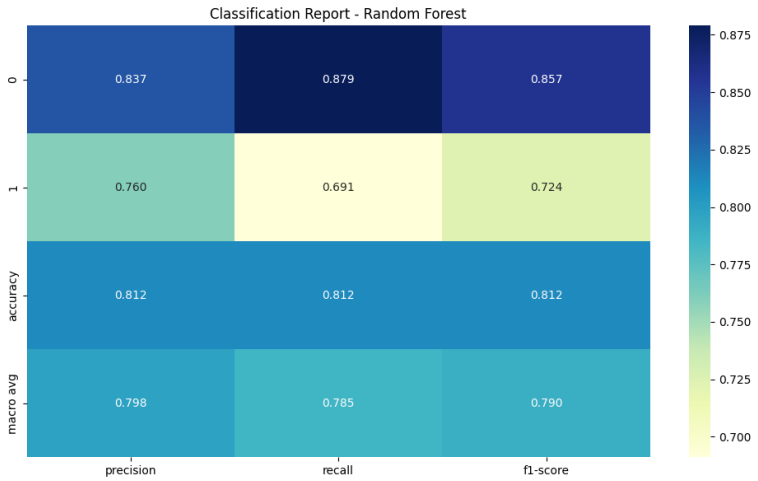
- F1-Score

F1-score untuk kelas 0 adalah sebesar 0.84, artinya kemampuan model untuk dapat memprediksi data dengan benar dan menghasilkan prediksi yang tepat pada kelas 0 adalah sebesar 84%. F1-score untuk kelas 1 adalah sebesar 0.65, artinya kemampuan model untuk dapat memprediksi data dengan benar dan menghasilkan prediksi yang tepat pada kelas 1 adalah sebesar 65%. Rata-rata f1-score adalah sebesar 0.74, artinya rata-rata kemampuan model untuk dapat memprediksi data dengan benar dan menghasilkan prediksi yang tepat adalah sebesar 74%

- Accuracy

Model memiliki akurasi sebesar 0.78, artinya kemampuan model untuk dapat menghasilkan prediksi dengan benar dan tepat dari keseluhuran data yang ada adalah sebesar 78%.

4.2 Random Forest



Gambar 7. Classification Report Random Forest

- Precision

Precision untuk kelas 0 adalah sebesar 0.84, artinya model dapat memprediksi 84% data pada kelas 0 dengan benar. Precision untuk kelas 1 adalah sebesar 0.81, artinya model dapat memprediksi 81% data pada kelas 1 dengan benar. Rata-rata precision adalah sebesar 0.82, artinya model dapat memprediksi rata-rata 82% data dengan benar

- Recall

Recall untuk kelas 0 adalah sebesar 0.91, artinya model dapat menghasilkan 91% prediksi yang tepat pada kelas 0. Recall untuk kelas 1 adalah sebesar 0.69, artinya model dapat menghasilkan 69% prediksi yang tepat pada kelas 1. Rata-rata recall adalah sebesar 0.80, artinya model dapat menghasilkan rata-rata 80% prediksi yang tepat

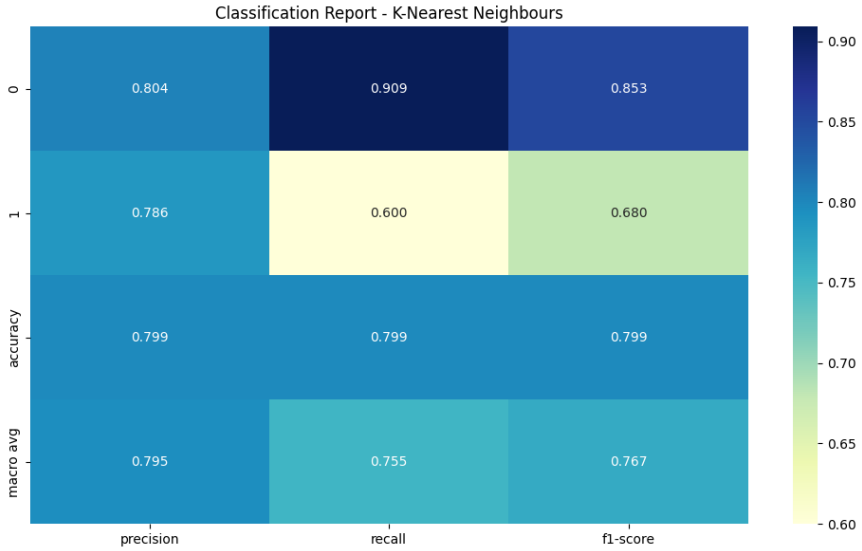
- F1-Score

F1-score untuk kelas 0 adalah sebesar 0.87, artinya kemampuan model untuk dapat memprediksi data dengan benar dan menghasilkan prediksi yang tepat pada kelas 0 adalah sebesar 87%. F1-score untuk kelas 1 adalah sebesar 0.75, artinya kemampuan model untuk dapat memprediksi data dengan benar dan menghasilkan prediksi yang tepat pada kelas 1 adalah sebesar 75%. Rata-rata f1-score adalah sebesar 0.81, artinya rata-rata kemampuan model untuk dapat memprediksi data dengan benar dan menghasilkan prediksi yang tepat adalah sebesar 81%

- Accuracy

Model memiliki akurasi sebesar 0.83, artinya kemampuan model untuk dapat menghasilkan prediksi dengan benar dan tepat dari keseluhuran data yang ada adalah sebesar 83%

4.3 K-Nearest Neighbour



Gambar 8. Classification Report K-Nearest Neighbour

- Precision

Precision untuk kelas 0 adalah sebesar 0.80, artinya model dapat memprediksi 80% data pada kelas 0 dengan benar. Precision untuk kelas 1 adalah sebesar 0.79, artinya model dapat memprediksi 79% data pada kelas 1 dengan benar. Rata-rata precision adalah sebesar 0.79, artinya model dapat memprediksi rata-rata 79% data dengan benar

- Recall

Recall untuk kelas 0 adalah sebesar 0.91, artinya model dapat menghasilkan 91% prediksi yang tepat pada kelas 0. Recall untuk kelas 1 adalah sebesar 0.60, artinya model dapat menghasilkan 60% prediksi yang tepat pada kelas 1. Rata-rata recall adalah sebesar 0.75, artinya model dapat menghasilkan rata-rata 75% prediksi yang tepat

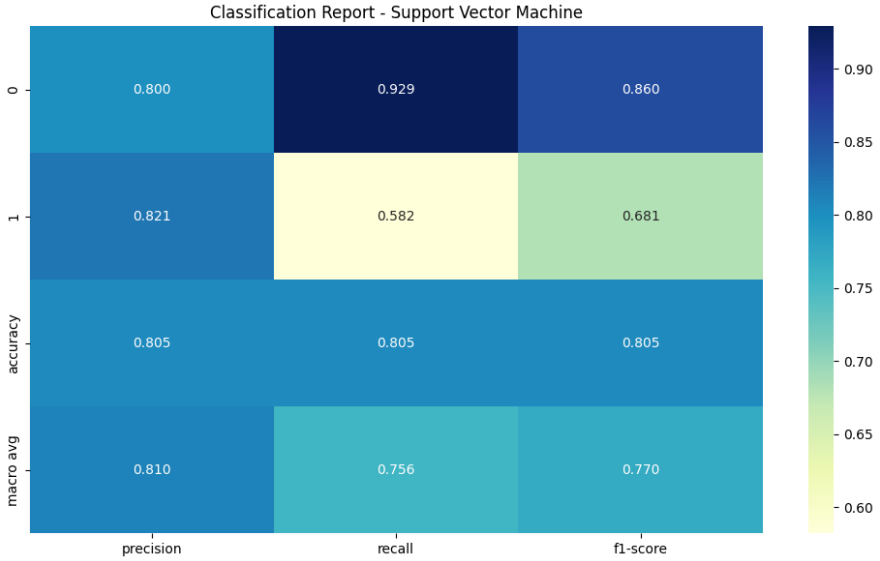
- F1-Score

F1-score untuk kelas 0 adalah sebesar 0.85, artinya kemampuan model untuk dapat memprediksi data dengan benar dan menghasilkan prediksi yang tepat pada kelas 0 adalah sebesar 85%. F1-score untuk kelas 1 adalah sebesar 0.68, artinya kemampuan model untuk dapat memprediksi data dengan benar dan menghasilkan prediksi yang tepat pada kelas 1 adalah sebesar 68%. Rata-rata f1-score adalah sebesar 0.77, artinya rata-rata kemampuan model untuk dapat memprediksi data dengan benar dan menghasilkan prediksi yang tepat adalah sebesar 77%

- Accuracy

Model memiliki akurasi sebesar 0.80, artinya kemampuan model untuk dapat menghasilkan prediksi dengan benar dan tepat dari keseluhuran data yang ada adalah sebesar 80%

4.3 Support Vector Mechine



Gambar 9. Classification Report Support Vector Mechine

- Precision

Precision untuk kelas 0 adalah sebesar 0.80, artinya model dapat memprediksi 80% data pada kelas 0 dengan benar. Precision untuk kelas 1 adalah sebesar 0.82, artinya model dapat memprediksi 82% data pada kelas 1 dengan benar. Rata-rata precision adalah sebesar 0.81, artinya model dapat memprediksi rata-rata 81% data dengan benar

- Recall

Recall untuk kelas 0 adalah sebesar 0.93, artinya model dapat menghasilkan 93% prediksi yang tepat pada kelas 0. Recall untuk kelas 1 adalah sebesar 0.86, artinya model dapat menghasilkan 86% prediksi yang tepat pada kelas 1. Rata-rata recall adalah sebesar 0.76, artinya model dapat menghasilkan rata-rata 76% prediksi yang tepat

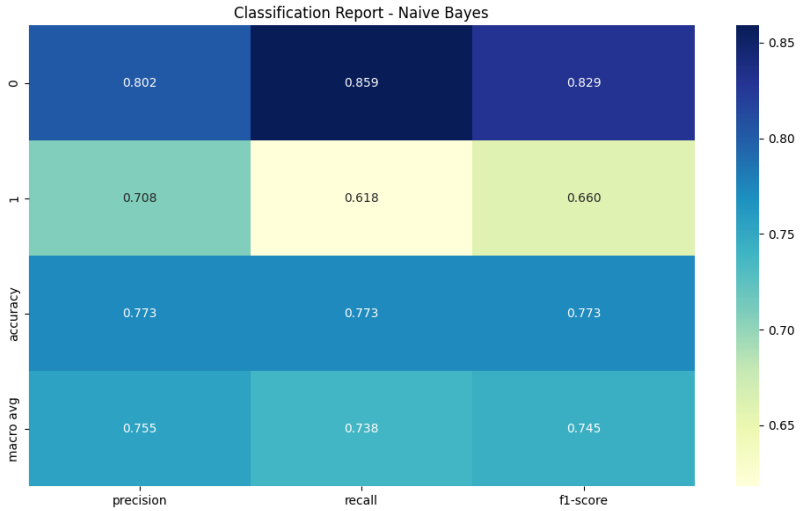
- F1-Score

F1-score untuk kelas 0 adalah sebesar 0.86, artinya kemampuan model untuk dapat memprediksi data dengan benar dan menghasilkan prediksi yang tepat pada kelas 0 adalah sebesar 86%. F1-score untuk kelas 1 adalah sebesar 0.68, artinya kemampuan model untuk dapat memprediksi data dengan benar dan menghasilkan prediksi yang tepat pada kelas 1 adalah sebesar 68%. Rata-rata f1-score adalah sebesar 0.77, artinya rata-rata kemampuan model untuk dapat memprediksi data dengan benar dan menghasilkan prediksi yang tepat adalah sebesar 77%

- Accuracy

Model memiliki akurasi sebesar 0.81, artinya kemampuan model untuk dapat menghasilkan prediksi dengan benar dan tepat dari keseluhuran data yang ada adalah sebesar 81%

4.4 Naive Bayes



Gambar 10. Classification Report Naive Bayes

- Precision

  Precision untuk kelas 0 adalah sebesar 0.80, artinya model dapat memprediksi 80% data pada kelas 0 dengan benar. Precision untuk kelas 1 adalah sebesar 0.71, artinya model dapat memprediksi 71% data pada kelas 1 dengan benar. Rata-rata precision adalah sebesar 0.76, artinya model dapat memprediksi rata-rata 76% data dengan benar
- Recall

  Recall untuk kelas 0 adalah sebesar 0.86, artinya model dapat menghasilkan 86% prediksi yang tepat pada kelas 0. Recall untuk kelas 1 adalah sebesar 0.62, artinya model dapat menghasilkan 62% prediksi yang tepat pada kelas 1. Rata-rata recall adalah sebesar 0.74, artinya model dapat menghasilkan rata-rata 74% prediksi yang tepat
- F1-Score

  F1-score untuk kelas 0 adalah sebesar 0.83, artinya kemampuan model untuk dapat memprediksi data dengan benar dan menghasilkan prediksi yang tepat pada kelas 0 adalah sebesar 83%. F1-score untuk kelas 1 adalah sebesar 0.66, artinya kemampuan model untuk dapat memprediksi data dengan benar dan menghasilkan prediksi yang tepat pada kelas 1 adalah sebesar 66%. Rata-rata f1-score adalah sebesar 0.74, artinya rata-rata kemampuan model untuk dapat memprediksi data dengan benar dan menghasilkan prediksi yang tepat adalah sebesar 74%
- Accuracy

  Model memiliki akurasi sebesar 0.77, artinya kemampuan model untuk dapat menghasilkan prediksi dengan benar dan tepat dari keseluhuran data yang ada adalah sebesar 77%

5. KESIMPULAN DAN SARAN

5.1 Kesimpulan

Dalam evaluasi lima algoritma machine learning (Logistic Regression, Random Forest, K-Nearest Neighbour, Support Vector Machine, dan Naive Bayes) terhadap prediksi diabetes, Random Forest menonjol dengan akurasi tertinggi 83%, sementara K-Nearest Neighbour menunjukkan keseimbangan yang baik antara precision (79%) dan recall (75%). Logistic Regression, Support Vector Machine, dan Naive Bayes juga memberikan kinerja memadai dengan akurasi masing-masing 78%, 81%, dan 77%. Analisis menggunakan metrik Precision, Recall, dan F1-Score menyoroti kelebihan dan kekurangan masing-masing model. Pemilihan model bergantung pada kebutuhan proyek dan trade-off antara akurasi dan keakuratan prediksi. Dengan demikian, Random Forest dan K-Nearest Neighbour dapat dianggap sebagai pilihan potensial berdasarkan evaluasi kinerja mereka dalam konteks prediksi diabetes.

5.2 Saran

Berdasarkan evaluasi kinerja empat algoritma machine learning untuk prediksi diabetes, disarankan untuk mempertimbangkan Random Forest dan K-Nearest Neighbour sebagai pilihan utama. Random Forest menonjol dengan akurasi tertinggi, sementara K-Nearest Neighbour menunjukkan keseimbangan yang baik antara precision dan recall. Namun, keputusan akhir sebaiknya disesuaikan dengan konteks dan tujuan spesifik proyek. Jika prioritas utama adalah mendapatkan prediksi yang sangat akurat, Random Forest dapat menjadi pilihan yang solid. Di sisi lain, jika keseimbangan antara precision dan recall lebih diutamakan, K-Nearest Neighbour bisa menjadi alternatif yang baik. Selain itu, mungkin juga bermanfaat untuk melakukan fine-tuning pada parameter model dan melakukan validasi silang untuk memastikan konsistensi kinerja di berbagai dataset.

REFERENSI

[1] R. Rizki, R. Athallah, I. Cholissodin, dan P. P. Adikara, "Prediksi Potensi Pengidap Penyakit Diabetes berdasarkan Faktor Risiko Menggunakan Algoritme Kernel K-Nearest Neighbor," 2022. [Daring]. Tersedia pada: http://j-ptiik.ub.ac.id

[2] N. P. Tigga dan S. Garg, "Prediction of Type 2 Diabetes using Machine Learning Classification Methods," dalam *Procedia Computer Science*, Elsevier B.V., 2020, hlm. 706–716. doi: 10.1016/j.procs.2020.03.336.

[3] M. Y. Aldean, Paradise, dan N. A. S. Nugraha, "Analisis Sentimen Masyarakat Terhadap Vaksinasi Covid-19 di Twitter Menggunakan Metode Random Forest Classifier (Studi Kasus: Vaksin Sinovac)," 2022.

[4] A. Rahman Isnain, A. Indra Sakti, D. Alita, dan N. Satya Marga, "SENTIMEN ANALISIS PUBLIK TERHADAP KEBIJAKAN LOCKDOWN PEMERINTAH JAKARTA MENGGUNAKAN ALGORITMA SVM," *JDMSI*, vol. 2, no. 1, hlm. 31–37, 2021, [Daring]. Tersedia pada: https://t.co/NfhnfMjtXw

[5] V. Chang, J. Bailey, Q. A. Xu, dan Z. Sun, "Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms," *Neural Comput Appl*, vol. 35, no. 22, hlm. 16157–16173, Agu 2023, doi: 10.1007/s00521-022-07049-z.