Analisis Seleksi Fitur Menggunakan Metode ANOVA F-test dan Algoritma Random Forest Untuk Deteksi Diabetes

Galih Hendro Martono, Ria Rismayati, Iptijanul Karor

Universitas Bumigora, Mataram, Indonesia

Correspondence : *e-mail*: iptijanul@gmail.com

Abstrak

Peningkatan level kadar glukosa darah yang melampaui batas normal merupakan ciri-ciri utama dari gangguan metabolisme yang dikenal sebagai diabetes mellitus, atau yang secara umum disebut penyakit kencing manis. Hal ini biasanya terjadi karena gangguan produksi atau fungsi insulin, baik secara absolut maupun relatif. Diperkirakan pada tahun 2030 diabetes akan menjadi penyebab kematian terbesar ke-7 di dunia hal ini didasari laporan dari *World Health Organization* (WHO). Ironisnya, sekitar 70% penderita diabetes tidak menyadari bahwa mereka mengidap penyakit ini, dan sekitar 25% telah mengalami komplikasi serius sebelum diagnosis ditegakkan. Oleh karena itu, deteksi dini serta manajemen risiko yang efektif sangat krusial untuk mencegah dampak kesehatan yang lebih berat. Pentingnya pemilihan fitur dalam meningkatkan akurasi prediksi diabetes adalah fokus penelitian ini. Metode seleksi fitur berbasis *ANOVA F-test* yang digabungkan dengan algoritma *Random Forest* dalam penyusunan model prediksi diabetes digunakan pada penelitian ini . Dataset yang digunakan terdiri dari 70.000 data dengan 33 atribut, yang kemudian diseleksi hingga diperoleh 13 fitur paling relevan berdasarkan nilai *P-value* < 0,05. Hasil evaluasi menunjukkan bahwa penerapan seleksi fitur secara signifikan meningkatkan performa model. Akurasi prediksi mencapai 73% saat menggunakan 5 fitur, meningkat menjadi 86% dengan 10 fitur, dan mencapai 90% ketika menggunakan 13 fitur. Temuan ini menggaris bawahi pentingnya proses seleksi fitur dalam pengembangan model prediktif penyakit diabetes, serta memberikan kontribusi penting dalam mendukung upaya deteksi dini dan pengelolaan risiko secara lebih optimal.

Kata kunci: Diabetes, klasifikasi, Seleksi Fitur, *ANOVA F-test*, *Random Forest*

Abstract

Increased blood glucose levels that exceed normal limits are the main characteristics of a metabolic disorder known as diabetes mellitus, or what is commonly called diabetes. This usually occurs due to impaired insulin production or function, both absolutely and relatively. It is estimated that by 2030 diabetes will be the 7th leading cause of death in the world, this is based on a report from the World Health Organization (WHO). Ironically, around 70% of people with diabetes are unaware that they have this disease, and around 25% have experienced serious complications before the diagnosis is made. Therefore, early detection and effective risk management are crucial to prevent more severe health impacts. The importance of feature selection in improving the accuracy of diabetes prediction is the focus of this study. The ANOVA F-test-based feature selection method combined with the Random Forest algorithm in compiling the diabetes prediction model was used in this study. The dataset used consisted of 70,000 data with 33 attributes, which were then selected to obtain the 13 most relevant features based on a P-value <0.05. The evaluation results show that the application of feature selection significantly improves model performance. Prediction accuracy reaches 73% when using 5 features, increases to 86% with 10 features, and reaches 90% when using 13 features. These findings underline the importance of the feature selection process in developing predictive models for diabetes, as well as providing important contributions in supporting early detection efforts and more optimal risk management.

Keywords*: Diabetes, Classification, Feature Selection, Anova F-test, Random Forest*

1. Pendahuluan

Salah satu tanda utama gangguan metabolisme yang disebut diabetes mellitus, atau penyakit kencing manis, adalah peningkatan kadar glukosa dalam darah di atas ambang batas normal. Seringkali

kondisi ini disebabkan oleh ketidakseimbangan dalam produksi maupun fungsi hormon insulin, baik secara keseluruhan (absolut) maupun sebagian (relatif). Diperkirakan pada tahun 2030 diabetes akan menjadi penyebab kematian terbesar ke-7 di dunia hal ini didasari laporan dari World Health Organization (WHO). Lebih dari itu, sekitar 70% penderita diabetes tidak menyadari bahwa mereka mengidap penyakit ini [1] yang berarti banyak kasus tidak terdiagnosis hingga komplikasi serius muncul. Oleh karena itu, upaya deteksi dini serta pengelolaan risiko secara tepat menjadi aspek krusial dalam pencegahan dampak kesehatan jangka panjang yang ditimbulkan oleh penyakit ini.

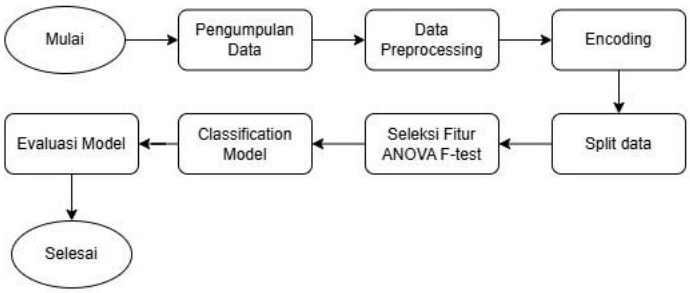
Dalam menghadapi tantangan tersebut, pemanfaatan teknik data mining untuk membangun sistem prediksi penyakit diabetes menjadi semakin relevan. Penelitian ini secara khusus menyoroti pentingnya proses seleksi fitur sebagai upaya untuk meningkatkan akurasi model prediktif. Seleksi fitur, yang merupakan salah satu metode reduksi dimensi, berfungsi untuk menyederhanakan kompleksitas data dengan cara memilih atribut-atribut yang paling relevan dan informatif [2]. Dengan menyaring fitur-fitur penting, model tidak hanya menjadi lebih efisien, tetapi juga berpotensi menghasilkan performa prediksi yang lebih akurat.

Penelitian ini mengusulkan pengembangan model prediksi diabetes berbasis algoritma Random Forest, yang dikenal handal dalam menangani data kompleks dan menghasilkan klasifikasi yang baik. Selain itu, metode ANOVA F-test digunakan dalam proses seleksi fitur untuk menemukan faktor-faktor yang secara statistik paling memengaruhi hasil diagnosis diabetes. Metode ini diharapkan dapat membantu dalam pembangunan sistem pendukung keputusan medis berbasis data.

Penelitian terdahulu terkait klasifikasi penyakit diabetes telah mengeksplorasi berbagai algoritma data mining untuk meningkatkan akurasi prediksi dini terkena diabetes. Penelitian yang dilakukan oleh [3] mengkaji penerapan pendekatan explainable artificial intelligence (XAI) dalam memprediksi penyakit diabetes menggunakan algoritma pembelajaran mesin. Studi ini secara khusus membandingkan dua metode interpretasi model, yaitu LIME dan SHAP, dalam menjelaskan mekanisme kerja yang dilakukan oleh model pembelajaran mesin yang digunakan. Dalam penelitian ini, model utama Logistic Regression dan Random Forest berhasil mencapai tingkat akurasi sebesar 86% pada data pengujian. Selanjutnya, studi oleh [4] mengeksplorasi penggunaan algoritma Random Forest untuk meramalkan perkembangan Diabetes Mellitus Tipe 1. Pendekatan ini berfokus pada kemampuan model dalam mendeteksi kemungkinan munculnya penyakit berdasarkan pola data yang tersedia. Sementara itu, penelitian oleh [5] menyoroti pemanfaatan berbagai teknik pembelajaran mesin untuk memprediksi komplikasi yang timbul akibat diabetes, khususnya Diabetic Retinopathy (DR) dan Diabetic Nephropathy (DN). Penelitian ini menggunakan sejumlah model seperti Random Forest, XGBoost, LightGBM, CatBoost, dan Neural Networks, serta menggabungkan beberapa model melalui pendekatan ensemble seperti Voting dan Stacking. Hasil analisis menunjukkan bahwa metode ensemble, terutama Voting dan Stacking, dapat memberikan kinerja yang sangat baik dengan nilai AUC mencapai 1.0 pada dataset yang telah melalui oversampling. Penelitian [6] melakukan penelitian yang membandingkan metode Support Vector Machine (SVM) dan Modified Balanced Random Forest (MBRF) untuk mengidentifikasi pasien yang menderita diabetes, dengan hasil menunjukkan bahwa MBRF mendapatkan akurasi maksimum sebesar 97,8%, sedangkan SVM hanya mencapai 87,94%. Penelitian ini menekankan pentingnya pemilihan algoritma yang tepat dalam pengembangan model prediksi diabetes. Selanjutnya, [7] membandingkan Random Forest dan SVM untuk mendeteksi risiko awal diabetes melitus. Hasil penelitian menunjukkan bahwa Random Forest memiliki akurasi maksimum 98,08%, lebih tinggi dari 91,03% dari SVM. Hasil ini menunjukkan bahwa Random Forest mungkin merupakan pilihan yang lebih baik untuk mengklasifikasikan risiko diabetes.[8] Dalam jurnal ini juga ditemukan bahwa algoritma Random Forest dapat dengan akurasi akhir sebesar 98% mengidentifikasi kemungkinan diabetes mellitus pada ibu hamil. Penelitian ini memberikan wawasan tambahan tentang efektivitas algoritma klasifikasi dalam mendeteksi diabetes dan menekankan pentingnya pemilihan algoritma yang tepat.

Banyak penelitian telah dilakukan untuk mengembangkan model prediksi diabetes, namun dalam penelitian sebelumnya, belum ada penelitian yang mendasari klasifikasi risiko diabetes menggunakan algoritma *Random Forest* dan metode *ANOVA F-test* untuk seleksi fitur. Oleh karena itu, tujuan penelitian ini adalah untuk menganalisis seberapa penting proses seleksi fitur dalam meningkatkan akurasi prediksi penyakit diabetes. Metode ANOVA F-test digunakan untuk menentukan fitur-fitur yang memiliki pengaruh terbesar pada prediksi diabetes, dengan cara menganalisis nilai *P-value* dari setiap fitur. Fitur yang nilai P-valuenya kurang dari 0,05 akan dipertahankan untuk digunakan dalam model prediksi. Dengan demikian, tujuan penelitian ini adalah untuk meningkatkan akurasi prediksi diabetes dengan mengoptimalkan fitur algoritma Random Forest.

2. Metode Penelitian



Gambar 1. Alur Penelitian

Karena data yang dikumpulkan dalam penelitian ini berbentuk numerik dan akan melalui proses analisis, jenis penelitian ini dikategorikan sebagai penelitian kuantitatif. Penelitian kuantitatif adalah salah satu ciri penelitian yang menonjolkan penggunaan angka dalam teknik pengumpulan datanya di dalam konteks penelitian. Gambar 1 menunjukkan tahapan penelitian ini. Tahapan ini sangat penting untuk memastikan bahwa penelitian memiliki struktur dan tujuan yang jelas.

2.1 Pengumpulan Data

Dataset merupakan komponen esensial dalam pengembangan sistem berbasis *machine learning* yang berfungsi sebagai tempat penyimpanan berbagai jenis data yang diperlukan dalam proses pelatihan dan pengujian model. Secara umum, dataset dapat memuat beragam format data, seperti citra digital, teks, audio, video, hingga data numerik. Keakuratan dan efektivitas model yang dibangun sangat bergantung pada keberadaan dataset yang representatif dan berkualitas [9]. Dataset sekunder yang dikumpulkan melalui Kaggle terdiri dari 70.000 data dengan 34 variabel atau atribut, yang terdiri dari 33 variabel independen dan satu variabel dependen. Penelitian ini menggunakan dataset ini. Untuk mengklasifikasikan jenis penyakit pasien, target yang berperan sebagai variabel dependen ini digunakan. Pada dataset yang digunakan terdapat 13 jenis penyakit diabetes. Rincian kategori jenis penyakit yang digunakan ditunjukkan temuan penelitian ini pada Tabel 1.

Tael 1 klasifikasi jenis penyakit

Jenis Penyakit	Deskripsi
Cystic Fibrosis-Related Diabetes (CFRD)	Diabetes yang terjadi pada penderita fibrosis kistik, karena gangguan fungsi pankreas.
Gestational Diabetes	Diabetes yang muncul selama kehamilan.
LADA (Latent Autoimmune Diabetes in Adults)	Diabetes genetik yang terjadi biasanya di usia muda.
MODY (Maturity-Onset Diabetes of the Young)	Pengaruh lingkungan, seperti polusi atau gaya hidup.
Neonatal Diabetes Mellitus (NDM)	Diabetes langka yang terjadi pada bayi baru lahir atau anak-anak yang belum berusia enam bulan.
Prediabetic	Jumlah gula darah lebih tinggi dari batas normal, tetapi belum cukup untuk mendiagnosis diabetes.
Secondary Diabetes	Diabetes yang diakibatkan oleh penyakit atau kondisi lain, seperti pankreatitis atau gangguan hormonal.
Steroid-Induced Diabetes	Diabetes yang terjadi akibat penggunaan steroid jangka panjang.
Type 1 Diabetes	Diabetes autoimun adalah kondisi di mana tubuh tidak memproduksi insulin sama sekali.
Type 2 Diabetes	Diabetes yang sering dijumpai, yaitu kondisi di mana tubuh tidak memanfaatkan insulin dengan efektif.
Type 3c Diabetes (Pancreatogenic Diabetes)	Diabetes yang disebabkan oleh kerusakan pankreas, misalnya akibat operasi atau cedera.
Wolcott-Rallison Syndrome	Diabetes genetik langka yang biasanya terjadi pada bayi dan disertai masalah kesehatan lain, seperti tulang atau hati.
Wolfram Syndrome	Gangguan genetik langka yang melibatkan diabetes dan masalah saraf, penglihatan, atau pendengaran.

2.2 *Data Preprocessing*

Data preprocessing merupakan serangkaian langkah terstruktur yang dilakukan pada data sebelum

dilakukan analisis atau diaplikasikan dalam pemodelan. Tujuannya adalah untuk merapikan, menyesuaikan, dan menyiapkan data agar dapat diolah dengan lebih efisien oleh algoritma analisis atau *machine learning* [10]. Tahapan *preprocessing* data melibatkan serangkaian tindakan, termasuk penghapusan duplikasi data dan penanganan nilai kosong. Setelah dilakukan kedua tindakan ini, diketahui bahwa dataset tidak mengandung data duplikat maupun *missing value*. Sebagai hasilnya, jumlah data yang dianalisis dalam penelitian ini tetap utuh, yaitu sejumlah 70.000 data.

2.3 *Encoding*

Apabila suatu dataset mengandung variabel dengan tipe kategorikal, maka diperlukan proses encoding untuk mengubah variabel tersebut ke dalam bentuk numerik. Transformasi ini menjadi langkah krusial karena sebagian besar algoritma machine learning hanya dapat memproses data dalam format angka, Oleh karena itu, mengubah variabel kategori menjadi angka memungkinkan algoritma untuk belajar pola dengan lebih efektif [11]. Dataset yang digunakan dalam penelitian ini masih mengandung variabel dengan tipe data *string*. Oleh karena itu, pada tahapan ini, variabel-variabel yang memiliki nilai *string* perlu diubah menjadi nilai numerik agar dapat diproses lebih lanjut. Dari total 34 variabel dalam dataset, teridentifikasi sebanyak 21 variabel yang awalnya berupa *string*.

Proses encoding yang dilakukan merupakan bagian dari tahapan dari pra-pemrosesan data. Setiap variabel kategorikal yang awalnya berupa teks dikonversi ke dalam bentuk angka agar dapat dipahami oleh sistem pembelajaran mesin. Contohnya, dalam variabel *Genetic Markers*, nilai “Negative” dikodekan menjadi 0 dan “Positive” menjadi 1, dengan asumsi bahwa nilai positif mengindikasikan risiko yang lebih tinggi. Pendekatan serupa diterapkan pada variabel *Autoantibodies*, *Family History*, *Environmental Factors*, dan *Early Onset Symptoms*, di mana kategori “Yes” atau “Present” dikonversi menjadi 1, sedangkan “No” atau “Absent” menjadi 0.

Beberapa variabel seperti *Physical Activity* dan *Socioeconomic Factors* memiliki lebih dari dua kategori, sehingga dilakukan penyandian bertingkat. Sebagai contoh, “Low” dikodekan menjadi 1, “Moderate” menjadi 2, dan seterusnya, sesuai dengan tingkat intensitas atau tingkat risiko yang diwakilinya. Khusus untuk variabel *Urine Test*, terdapat empat kategori berbeda yang masing-masing diberikan kode angka unik mulai dari 0 hingga 3. Hal ini memungkinkan model membedakan setiap kondisi dengan jelas tanpa kehilangan informasi penting.

Pada dataset yang digunakan, variabel "Target" atau kategori jenis penyakit juga memiliki nilai dalam format *string*. Tabel 2 menyajikan hasil dari proses konversi nilai *string* pada variabel jenis penyakit diabetes menjadi bentuk numerik.

Tabel 2 konversi jenis penyakit menjadik numerik

Nama Variabel	Nilai *String*	Konversi
Target	*Cystic Fibrosis-Related Diabetes (CFRD)*	0
	Gestational Diabetes	1
	LADA (Latent Autoimmune Diabetes in Adults)	2
	MODY (Maturity-Onset Diabetes of the Young)	3
	Neonatal Diabetes Mellitus (NDM)	4
	Prediabetic	5
	Secondary Diabetes	6
	Steroid-Induced Diabetes	7
	Type 1 Diabetes	8
	Type 2 Diabetes	9
	Type 3c Diabetes (Pancreatogenic Diabetes)	10
	Wolcott-Rallison Syndrome	11
	Wolfram Syndrome	12

2.4 *Split Data*

Akurasi suatu model sangat dipengaruhi oleh proporsi pembagian data yang dilakukan menjadi dua jenis, yaitu data latih dan data uji. [12]. Langkah selanjutnya dalam penelitian ini adalah

mengidentifikasi dan memisahkan variabel bebas x dan variabel terikat y. Variabel x terdiri atas kolom-kolom mulai dari "*Genetic Markers*" hingga "*Early Onset Symptoms*", untuk saat ini variabel y adalah kolom "Target" yang berisi kategori penyakit diabetes. Kemudian, dataset terbagi menjadi dua bagian yaitu data pelatihan dan data pengujian, dengan menggunakan fungsi *train_test_split* dari pustaka *sklearn.model_selection* di *Python*. Proporsi pembagian yang diterapakan yaitu 80% untuk data latih dan 20% untuk data uji.

2.5 Seleksi Fitur *ANOVA F-test*

Tahapan selanjutnya ialah seleksi fitur tujuannya adalah meningkatkan kinerja model, mengurangi kompleksitas komputasi, dan mencegah overfitting. Signifikansi statistik dari masing-masing elemen yang memprediksi variabel target dalam penelitian ini dinilai melalui penggunaan metode *ANOVA F-test*. Fitur dengan nilai *F-statistic* yang tinggi dan *P-value* < 0,05 dipilih untuk pelatihan model karena memiliki pengaruh yang signifikan [13]. Untuk nilai F pada setiap fitur dihitung berdasarkan rasio variansi antar grup terhadap variansi dalam grup. Untuk nilai F yang paling tinggi dihitung menggunakan metode *SelectKBest* untuk menciptakan subset fitur yang paling signifikan terhadap variabel target [14]. Persamaan *ANOVA F-Test* dapat dilihat pada Persamaan 1 sampai dengan persamaan 3.

$$F = \frac{\text{variance between groups}}{\text{variance within groups}} \tag{1}$$

$$\text{Variance between groups} = \frac{\sum n_i(\overline{Y}_i - \overline{Y})^2}{k-1} \tag{2}$$

$$\text{Variance within groups} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij} - \overline{Y}_i\right)^2}{n-k} \tag{3}$$

n_i menyatakan jumlah sampel pada grup ke-i, sedangkan $\overline{Y}_i$ adalah rata-rata nilai sampel pada grup ke-i. Nilai $\overline{Y}$ menunjukkan rata-rata total dari seluruh sampel yang dianalisis. Adapun k merepresentasikan jumlah total grup yang dibandingkan dalam analisis, dan n merupakan jumlah total dalam *dataset*.

2.6 *Classification Model*

Langkah selanjutnya yakni membangun model klasifikasi untuk memprediksi penyakit diabetes setelah tahap seleksi fitur. Pada penelitian ini, *Random Forest* merupakan algoritma yang diterapkan untuk klasifikasi. *Random Forest* merupakan sekumpulan pohon keputusan yang dibangun menggunakan sampel yang diambil secara acak, namun dengan ketentuan yang berbeda dalam memecah simpul. Model ini beroperasi dengan memanfaatkan subset fitur pada setiap pohon dan berusaha mencari ambang batas yang paling optimal dalam pembagian data. Oleh karena itu, sejumlah pohon akan dilatih dengan metode yang kurang kuat dan setiap pohon akan memberikan prediksi yang berbeda [15]. *Random Forest* memiliki sejumlah kelebihan, seperti kemampuannya untuk meningkatkan akurasi dalam situasi di mana data tidak lengkap, kemampuan untuk mengurangi terjadinya kesalahan, dan kemampuan untuk menyimpan data secara efisien [16]. *Random Forest* merupakan algoritma yang tepat untuk digunakan pada masalah klasifikasi dalam *machine learning* dan pemrosesan data [17]. *Random Forest* merupakan metode pembelajaran ensemble yang bertujuan untuk meningkatkan akurasi prediksi dengan membangkitkan atribut secara acak pada setiap node pohon keputusan. Model ini terdiri dari kumpulan banyak pohon keputusan yang bekerja secara kolektif untuk mengelompokkan data ke dalam kelas tertentu. Setiap pohon dalam *Random Forest* dimulai dari sebuah simpul akar dan berakhir pada sejumlah simpul daun yang menghasilkan keputusan akhir [18]. *Random Forest* pada proses pembentukan pohon keputusan mirip dengan algoritma Classification and Regression Tree (CART), perbedaan utama yakni dalam algoritma *Random Forest* tidak ada proses pemangkasan pohon yang dilakukan. Untuk menentukan fitur yang paling baik pada setiap simpul internal pohon, *Random Forest* menggunakan ukuran impuritas berupa Indeks Gini. Nilai Gini dihitung menggunakan Persamaan 4.

$$Gini(S_i) = 1 - \sum_{i=0}^{c-1} p_i^2 \tag{4}$$

P_i menyatakan frekuensi relatif dari kelas C_i dalam suatu subset data tertentu, yaitu proporsi jumlah sampel dari kelas C_i terhadap tota sampel dalam subset tersebut. Di mana C_i

adalah kelas ke-i dan c adalah jumlah total kelas dalam *dataset.* Nilai P_i ini digunakan untuk menghitung seberapa homogen atau seberapa bersih sebuah *dataset* setelah pemisahan dilakukan berdasarkan fitur tertentu. Kualitas pemisahan untuk fitur ke-k kemudian dihitung berdasarkan distribusi P_i pada masing-masing subset S_i menggunakan persamaan 5.

$$Gini_{split} = \sum_{i=0}^{k-1} \left(\frac{n_i}{n}\right) Gini(S_i) \tag{5}$$

n_i merujuk pada total jumlah sampel di dalam subset S_i setelah dilakukan pemisahan, sedangkan n merupakan total keseluruhan sampel yang terdapat pada node sebelum dilakukan pemisahan .s

Random Forest membangun banyak pohon keputusan $\{\boldsymbol{h}(\boldsymbol{x}, \boldsymbol{\Theta_k}), \boldsymbol{k} = \boldsymbol{1}, \ldots\}$ dimana $\boldsymbol{\Theta_k}$ adalah vektor parameter acak yang bersifat independent identically distributed (i.i.d). Setiap pohon menghasilkan prediksi, dan hasil klasifikasi akhir ditentukan berdasarkan mayoritas suara (majority vote) dari semua pohon. kekuatan masing-masing pohon klasifikasi dan korelasi antar pohon. Fungsi margin dari *Random Forest* didefinisikan pada Persamaan 6.
3.

$$mr(X,Y) = P_\Theta(h(X,\Theta) = Y) - \max_{j \neq Y} P_\Theta\,(h(X,\Theta) = j) \tag{6}$$

Kekuatan dari himpunan pengklasifikasi $\{h(X,\Theta)\}$ dihitung dengan Persamaan 7.

$$PE \leq \frac{\bar{\rho}(1-s^2)}{s^2} \tag{7}$$

Di mana $\bar{\rho}$ adalah rata-rata korelasi antar prediksi pohon, dihitung dengan Persamaan 8.

$$\bar{\rho} = \frac{E_{\Theta,\Theta'}\big(\rho(\Theta,\Theta')sd(\Theta)sd(\Theta')\big)}{E_{\Theta,\Theta'}\big(sd(\Theta)sd(\Theta')\big)} \tag{8}$$

2.7 Evaluasi Model

Untuk menilai dampak seleksi fitur terhadap kinerja model prediksi, beberapa model *Random Forest* dilatih dan dievaluasi menggunakan subset fitur yang berbeda, mulai dari subset fitur yang lebih kecil hingga subset fitur dengan kualitas terbaik yang dipilih oleh *ANOVA F-test*. Akurasi, presisi, *recall*, serta *F1-score* ialah metrik evaluasi yang sesuai digunakan untuk mengevaluasi kemampuan model. Akurasi menggambarkan bagian total dari prediksi yang benar di antara semua data yang telah diklasifikasikan oleh model, sementara presisi menunjukkan seberapa akurat model dalam menghasilkan prediksi untuk kelas positif. *Recall* mengukur sejauh mana model mampu dengan benar mengidentifikasi semua contoh dari kelas positif, sedangkan *F1-score* adalah rata-rata harmonis dari presisi dan *recall* yang menunjukkan keseimbangan antara kedua metrik ini dalam menilai kemapuan model [19]. Untuk mengukur akurasi, presisi, Recall, dan F1-Score. Persamaan 9 hingga 12 adalah rumus:

$$\text{Akurasi} = \frac{TP+TN}{TP+TN+FP+FN} \tag{9}$$

$$\text{Presisi} = \frac{TP}{TP+FP} \tag{10}$$

$$\textit{Recall} = \frac{TP}{TP+FN} \tag{11}$$

$$\textit{F1-Score} = 2 \times \frac{\text{Presisi} \times \textit{Recall}}{\text{Presisi} + \textit{Recall}} \tag{12}$$

4. Hasil dan Pembahasan

4.1 Hasil *Encoding*

Dalam langkah berikut, proses *encoding* digunakan untuk mengubah variabel atau atribut yang memiliki nilai kategorikal atau *string* ke dalam nilai numerik. Nilai kategori, atau *string*, dapat diubah menjadi nilai numerik dengan menggunakan fungsi "*LabelEncoder*()" dari library "*sklearn.preprocessing*". Karena algoritma yang digunakan dalam penelitian ini difokuskan untuk memproses data numerik, fungsi ini sangat berguna saat bekerja dengan data kategorikal dalam *data mining*. Dataset disimpan menjadi sebuah dataset baru setelah menjadi numerik. Contoh data dari dataset setelah proses *encoding* ditunjukkan dalam Tabel 3.

Tabel 3 dataset setelah proses *encoding*

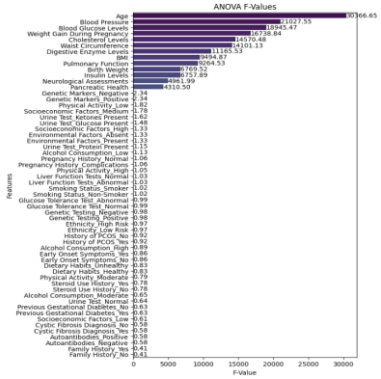
Target	Genetic Markers	Autoantibodies	Family History	Environmental Factors	Insulin Levels	Age	…	Liver Function Tests	Digestive Enzyme Levels	Urine Test	Birth Weight	Early Onset Symptoms
7	1	0	0	1	40	44	…	1	56	1	2629	0
4	1	0	0	1	13	1	…	1	28	0	1881	1
5	1	1	1	1	27	36	…	0	55	1	3622	1
8	0	1	0	1	8	7	…	0	60	1	3542	0
12	0	0	1	1	17	10	…	1	24	3	1770	0
2	1	0	1	1	17	41	…	1	52	1	3835	1
9	0	0	0	0	29	30	…	0	96	1	4426	0
…	…	…	…	…	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…	…	…	…	…
0	1	0	0	0	32	30	…	0	35	1	2592	1
2	1	1	1	0	27	41	…	0	64	1	3593	1

4.2 Hasil Split Data

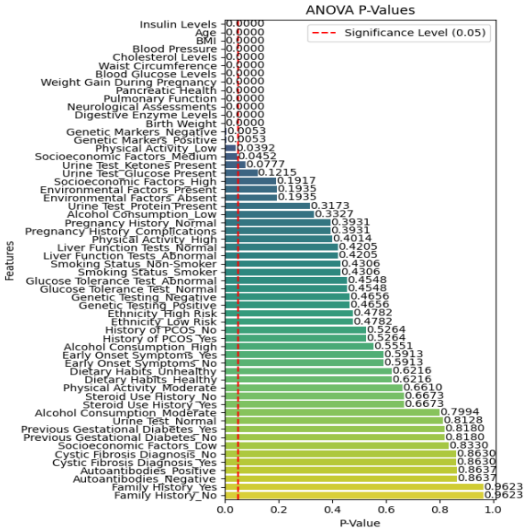
Proses pembagian data merupakan tahap krusial dalam pengembangan model machine learning, guna memisahkan data yang dipakai untuk melatih dan menguji model. Berdasarkan hasil yang ditampilkan pada gambar, data telah dikelompokkan menjadi dua kategori utama, yakni data pelatihan dan data pengujian. Data pelatihan terdiri dari 56.000 sampel dengan masing-masing memiliki 58 atribut atau fitur yang merepresentasikan karakteristik dari setiap entri. Sedangkan data uji terdiri dari 14.000 sampel dengan jumlah fitur yang sama, yakni 58, yang digunakan untuk menilai kinerja model setelah proses pelatihan selesai dilakukan. Selain itu, label atau target yang menyertai data latih berjumlah 56.000, Sementara itu, label untuk data uji berjumlah 14.000. Proporsi pembagian ini menunjukkan bahwa sekitar 80% data diterapkan sebagai pelatihan model, sementara itu, sisa 20% digunakan untuk menevaluasi akurasi dan kemampuan model dalam melakukan generalisasi terhadap data yang belum pernah ditemukan sebelumnya. Strategi ini sangat penting untuk menjamin bahwa model tidak hanya andal terhadap data pelatihan, tetapi juga mampu memberikan prediksi yang baik pada data baru.

4.3 Hasil Seleksi Fitur ANOVA F-test

Setelah melakukan split data, tahap selanjutnya ialah menerapkan seleksi fitur menggunakan metode *ANOVA F-test*. Pertama, nilai *F-statistic* dan *P-value* untuk setiap fitur dalam data pelatihan dihitung. Perhitungan *ANOVA F-test* menghasilkan informasi tentang signifikansi setiap fitur dalam membedakan kategori kelas. Gambar 2 dan Gambar 3 menyajikan visualisasi hasil perhitungan *F-statistic* dan *P-value*, secara berurutan. Sedangkan Tabel 4 menyajikan fitur-fitur yang diseleksi, dikelompokkan berdasarkan jumlah fitur yang dipilih, yaitu 5 fitur, 10 fitur, dan 13 fitur.



Gambar 2. Hasil Visualisai *F-statistic*



Gambar 3. Hasil visualisasi *P-value*

Tabel 4. Hasil Fitur Yang Diseleksi Berdasarkan Jumlah

Hasil Seleksi 5 Fitur Terpilih	Hasil Seleksi 10 Fitur Terpilih	Hasil Seleksi 13 Fitur Terpilih
Age	*Insulin Levels*	*Insulin Levels*
Blood Pressure	*Age*	*Age*
Cholesterol Levels	*BMI*	*BMI*
Blood Glucose Levels	*Blood Pressure*	*Blood Pressure*
Weight Gain During Pregnancy	*Cholesterol Levels*	*Cholesterol Levels*
	Waist Circumference	*Waist Circumference*
	Blood Glucose Levels	*Blood Glucose Levels*
	Weight Gain During Pregnancy	*Weight Gain During Pregnancy*
	Pulmonary Function	*Pancreatic Health*
	Digestive Enzyme Levels	*Pulmonary Function*
		Neurological Assessments
		Digestive Enzyme Levels
		Birth Weight

4.4 Hasil Evaluasi Model

Untuk mengevaluasi kinerja model klasifikasi yang digunakan menggunakan algoritma *Random Forest*, hasil prediksi dibandingkan dengan data uji. Untuk setiap subset fitur yang dipilih (5 fitur, 10 fitur, dan 13 fitur), metrik evaluasi yang digunakan mencakup akurasi, presisi, *recall*, serta F1-Score. Tabel 5 menampilkan hasil tingkat akurasi dari seleksi fitur berdasarkan jumlah fitur.

Tabel 5. Hasil Akurasi Berdasarkan Jumlah Fitur

Jumlah Fitur	Akurasi
5 Fitur	72,40%
10 Fitur	89,23%
13 Fitur	90,43%

Berdasarkan hasil evaluasi performa model terhadap jumlah fitur yang digunakan, dapat disimpulkan bahwa seleksi fitur memberikan dampak signifikan terhadap tingkat akurasi prediksi. Saat model hanya dilatih menggunakan lima fitur teratas, akurasi yang diperoleh tercatat sebesar 72,40%. Meskipun jumlah fiturnya terbatas, performa model masih tergolong cukup baik. Namun, ketika jumlah fitur ditingkatkan menjadi sepuluh, akurasi meningkat tajam hingga mencapai 89,23%, yang menunjukkan bahwa penambahan fitur-fitur yang relevan mampu memberikan informasi yang lebih komprehensif bagi proses klasifikasi. Selanjutnya, pemanfaatan tiga belas fitur menghasilkan akurasi tertinggi sebesar 90,43%. Kenaikan akurasi dari sepuluh ke tiga belas fitur memang tidak terlalu signifikan, namun tetap memberikan

indikasi bahwa fitur tambahan tersebut memiliki kontribusi positif. Secara keseluruhan, penerapan metode ANOVA F-test dalam tahap seleksi fitur terbukti efektif dalam menyaring atribut-atribut yang paling berpengaruh, dan ketika dipadukan dengan algoritma *Random Forest*, mampu menghasilkan model prediksi diabetes yang lebih presisi dan efisien. Penelitian ini sejalan dengan penelitian sebelumnya [20]yang juga menggunakan seleksi fitur dan reduksi dimensi dengan hasil yang lebih baik.

5. Kesimpulan

Berdasarkan hasil penelitian, dapat disimpulkan bahwa penggunaan metode seleksi fitur *ANOVA F-test* secara signifikan mampu meningkatkan akurasi model klasifikasi penyakit diabetes, khususnya ketika dikombinasikan dengan algoritma *Random Forest*. Model dengan 13 fitur terpilih menunjukkan akurasi tertinggi sebesar 90,43%, dibandingkan dengan model yang hanya menggunakan lima fitur (72,40%) dan sepuluh fitur (89,23%). *ANOVA F-test* memiliki keunggulan dalam mengidentifikasi fitur-fitur yang paling relevan, sehingga dapat menurunkan kompleksitas komputasi serta meminimalkan risiko *overfitting* tanpa menurunkan performa prediksi. Penelitian ini dapat dikembangkan lebih lanjut dengan menerapkan metode seleksi fitur lain, seperti *Recursive Feature Elimination* (RFE) atau Mutual Information, yang belum diuji dalam penelitian ini. Selain itu, pengujian konsistensi performa terhadap algoritma klasifikasi lain seperti *XGBoost* atau *Neural Network* juga perlu dilakukan. Oleh karena itu, disarankan agar metode seleksi fitur *ANOVA F-test* dikombinasikan dengan berbagai algoritma klasifikasi dan diuji pada dataset dari beragam sumber guna membangun model yang lebih kuat, adaptif, dan aplikatif dalam konteks nyata.

Daftar Pustaka

[1] T. Gita Miranda, L. Yovita Sari, Stik. Bhakti Husada Bengkulu, and U. Dehasen, "Faktor Risiko Kejadian Diabetes Melitus (dm) Tipe 2 Di Poliklinik Penyakit Dalam Rs Kota Bengkulu," 2023.

[2] I. Maulida, A. Suyatno, H. Rahmania Hatta, and U. Mulawarman, "Seleksi Fitur Pada Dokumen Abstrak Teks Bahasa Indonesia Menggunakan Metode Information Gain," *OKTOBER 2016 IJCCS*, vol. 17, pp. 1–5, 2016.

[3] S. Ahmed, M. Shamim Kaiser, M. S. Hossain, and K. Andersson, "A Comparative Analysis of LIME and SHAP Interpreters with Explainable ML-Based Diabetes Predictions," *IEEE Access*, 2024, doi: 10.1109/ACCESS.2024.3422319.

[4] N. F. Cleymans, M. Van De Casteele, J. Vandewalle, A. K. Desouter, F. K. Gorus, and K. Barbe, "Analyzing Random Forest's predictive capability for Type 1 Diabetes progression," *IEEE Open Journal of Instrumentation and Measurement*, 2025, doi: 10.1109/OJIM.2025.3551837.

[5] D. R. Manjunath, J. J. Lohith, S. Selva Kumar, and A. Das, "Predicting Diabetic Retinopathy and Nephropathy Complications Using Machine Learning Techniques," *IEEE Access*, 2025, doi: 10.1109/ACCESS.2025.3562483.

[6] M. D. Purbolaksono, M. Irvan Tantowi, A. Imam Hidayat, and A. Adiwijaya, "Perbandingan Support Vector Machine dan Modified Balanced Random Forest dalam Deteksi Pasien Penyakit Diabetes," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 5, no. 2, pp. 393–399, Apr. 2021, doi: 10.29207/resti.v5i2.3008.

[7] C. Z. V. Junus, T. Tarno, and P. Kartikasari, "Klasifikasi Menggunakan Metode Support Vector Machine Dan Random Forest Untuk Deteksi Awal Risiko Diabetes melitus," *Jurnal Gaussian*, vol. 11, no. 3, pp. 386–396, Jan. 2023, doi: 10.14710/j.gauss.11.3.386-396.

[8] R. B. Prasetyo, "Prediksi Dini Penyakit Diabetes Pada Ibu Hamil Dengan Algoritma Random Forest," 2024.

[9] T. Z. Jasman, E. Hasmin, Sunardi, C. Susanto, and W. Musu, "Perbandingan Logistic Regression, Random Forest, dan Perceptron pada Klasifikasi Pasien Gagal Jantung," *CSRID (Computer Science Research and Its Development Journal)*, vol. 14, no. 3, pp. 271–286, Dec. 2022, doi: 10.22303/csrid.14.3.2022.271-286.

[10] F. Putra, H. F. Tahiyat, R. M. Ihsan, R. Rahmaddeni, and L. Efrizoni, "Penerapan Algoritma K-Nearest Neighbor Menggunakan Wrapper Sebagai Preprocessing untuk Penentuan Keterangan Berat Badan Manusia," *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 4, no. 1, pp. 273–281, Jan. 2024, doi: 10.57152/malcom.v4i1.1085.

[11] A. Agung, A. Daniswara, I. Kadek, and D. Nuryana, "Data Preprocessing Pola Pada Penilaian Mahasiswa Program Profesi Guru," *Journal of Informatics and Computer Science*, vol. 05, 2023.

[12] Y. A. Prasetyo, E. Utami, and A. Yaqin, "Pengaruh Komposisi Split Data Terhadap Performa

Akurasi Analisis Sentimen Algoritma Naïve Bayes dan SVM," *Journal homepage: Journal of Electrical Engineering and Computer (JEECOM)*, vol. 6, no. 2, 2024, doi: 10.33650/jeecom.v4i2.

[13] E. Hariyanti, D. Pramana Hostiadi, Y. Priyo Atmojo, I. Made Darma Susila, and I. Tangkawarow, "Analisis Perbandingan Metode Seleksi Fitur pada Model Klasifikasi Decission Tree untuk Deteksi Serangan di Jaringan Komputer".

[14] Hasna Marhamah Auliya, "Seleksi Fitur Pada Klasifikasi Kesejahteraan Janin Berdasarkan Data Kardiotokografi (KTG) Berbasis Machine Learning," 2025.

[15] D. Haganta Depari *et al.*, "Perbandingan Model Decision Tree, Naive Bayes dan Random Forest untuk Prediksi Klasifikasi Penyakit Jantung," *JURNAL INFORMATIK Edisi ke*, vol. 18, p. 2022.

[16] S. Devella and F. Novia Rahmawati, "Implementasi Random Forest Untuk Klasifikasi Motif Songket Palembang Berdasarkan SIFT," *Jurnal Teknik Informatika dan Sistem Informasi*, vol. 7, no. 2, 2020, [Online]. Available: http://jurnal.mdp.ac.id

[17] A. U. Zailani and N. L. Hanun, "Penerapan Algoritma Klasifikasi Random Forest Untuk Penentuan Kelayakan Pemberian Kredit Di Koperasi Mitra sejahtera," *Infotech: Journal of Technology Information*, vol. 6, no. 1, pp. 7–14, Jun. 2020, doi: 10.37365/jti.v6i1.61.

[18] 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, Dec. 2022, doi: 10.35580/variansiunm31.

[19] M. Fadli and R. A. Saputra, "Klasifikasi Dan Evaluasi Performa Model Random Forest Untuk Prediksi Stroke Classification And Evaluation Of Performance Models Random Forest For Stroke Prediction," vol. 12, 2023, [Online]. Available: http://jurnal.umt.ac.id/index.php/jt/index

[20] N. Sulistianingsih and G. H. Martono, "Feature Selection versus Feature Extraction Models for IoT Intrusion Detection System Using Convolutional Neural Network," in *COMNETSAT 2024 - IEEE International Conference on Communication, Networks and Satellite*, Institute of Electrical and Electronics Engineers Inc., 2024, pp. 35–42. doi: 10.1109/COMNETSAT63286.2024.10862350.