

Analisis Diskriminan Menggunakan *Stratified k-Fold Cross Validation* Pada Klasifikasi *Gestational Diabetes Mellitus* (GDM) Ibu Hamil di Rumah Sakit XYZ

Siti Aisyah*, Nur Azizah Komara Rifai

Prodi Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Islam Bandung, Indonesia.

*siti.aisyah.001228@gmail.com, nur.azizah@unisba.ac.id.

Abstract. Discriminant analysis is used for the purpose of classifying an object based on certain criteria and predicting groups for new objects. This study will classify pregnant women affected by GDM and not who are evaluated using stratified k-fold cross validation. In the analysis, several assumptions are needed that must be met, including independent variables must be multivariate normal distributed, homogeneous covariance matrices, and there are average vector differences. The stages of analysis are data exploration, testing discriminant analysis assumptions, determining the number of k-folds, randomizing into training data and testing data gradually (stratified), conducting experiments as many as k times, estimating the discriminant model of each experiment, testing the accuracy of classification and determining the best model. The results of this study the best model was found in the first experiment with the model $Y_1 = 0.0438 \text{ VAT} - 0.1120 \text{ Gestational age} + 0.0887 \text{ Blood sugar} + 0.0535 \text{ BMI}$ with a cut off value of 0.0001739. That is, if there is a new object with a value of $Y_1 \geq 0.0001739$ then the pregnant woman falls into the category of not affected by GDM, and vice versa if $Y_1 < 0.0001739$ then pregnant women fall into the category affected by GDM.

Keywords: *Discriminant Analysis, Classification, Stratified k-Fold Cross Validation.*

Abstrak. Analisis diskriminan digunakan untuk tujuan mengklasifikasikan sebuah objek berdasarkan kriteria tertentu dan memprediksi kelompok untuk objek baru. Penelitian ini akan mengklasifikasikan ibu hamil yang terkena GDM dan tidak yang dievaluasi menggunakan *stratified k-fold cross validation*. Dalam analisisnya, diperlukan beberapa asumsi yang harus dipenuhi diantaranya variabel independen harus berdistribusi normal multivariat, matriks varians kovarian yang homogen, dan ada perbedaan vektor rata-rata. Tahapan analisisnya yaitu eksplorasi data, menguji asumsi analisis diskriminan, menentukan banyaknya *k-fold*, melakukan pengacakan menjadi data latih dan data uji secara bertahap (*stratified*), melakukan eksperimen sebanyak k kali, mengestimasi model diskriminan dari masing-masing eksperimen, menguji ketepatan klasifikasi dan menentukan model terbaiknya. Hasil dari penelitian ini model terbaik terdapat pada eksperimen pertama dengan model $Y_1 = 0.0438 \text{ VAT} - 0.1120 \text{ Usia kehamilan} + 0.0887 \text{ Gula darah} + 0.0535 \text{ BMI}$ dengan nilai *cut off* sebesar 0.0001739. Artinya, jika ada objek baru dengan nilai $Y_1 \geq 0.0001739$ maka ibu hamil tersebut masuk ke dalam kategori tidak terkena GDM, dan sebaliknya apabila $Y_1 < 0.0001739$ maka ibu hamil masuk ke dalam kategori terkena GDM.

Kata Kunci: *Analisis Diskriminan, Klasifikasi, Stratified k-Fold Cross Validation.*

A. Pendahuluan

Analisis diskriminan merupakan salah satu metode pengujian statistik multivariat yang bertujuan untuk mengklasifikasikan objek sehingga memperoleh fungsi yang merupakan kombinasi linear dari variabel sehingga mampu membedakan objek penelitian sesuai dengan kelompoknya. Selain itu, fungsi diskriminan dapat juga digunakan untuk memprediksi kelompok yang cocok untuk suatu objek baru.

Analisis lain yang dapat digunakan untuk memprediksi yaitu analisis regresi, tetapi analisis regresi hanya memprediksi nilai dari variabel dependen, sedangkan analisis diskriminan dapat digunakan untuk mengklasifikasikan objek baru, sehingga lebih cocok digunakan untuk mengklasifikasikan sebuah objek (Maulidya, et. al, 2013). Menurut A. Davidson dan D Hinkley dalam Mardiana, et. al, (2022), analisis diskriminan memerlukan suatu metode untuk mengevaluasi kinerja *classifier*, salah satu yang seringkali digunakan yaitu *k-Fold Cross Validation*, karena metode ini semua objek pada dataset dapat menjadi data latih dan data uji serta bersifat *mutually exclusive* (tidak saling mempengaruhi).

Penelitian ini menggunakan data sekunder yang didapat dari Rumah Sakit XYZ dengan ukuran sampel sebanyak 133 orang ibu hamil. Tujuan penulisan penelitian ini untuk mendapatkan model diskriminan serta mengukur keakuratan dan konsistensi dari hasil klasifikasi GDM pada ibu hamil di Rumah Sakit XYZ. Pentingnya penelitian ini, untuk mengklasifikasikan ibu hamil tanpa GDM dan dengan GDM agar mendapat penanganan yang tepat untuk meminimalisir komplikasi/efek samping dari GDM baik untuk ibu maupun janin/bayi. Beberapa efek sampingnya bagi ibu yaitu makrosomia (berat badan janin/bayi yang besar) sehingga diharuskan operasi caesar, keguguran, risiko mengalami GDM pada kehamilan selanjutnya, dan berkembang menjadi diabetes tipe 2 setelah melahirkan (Kemenkes, 2022).

B. Metodologi Penelitian

Analisis Diskriminan

Menurut William dan Matthew (1984) dalam Rosyid (2017) analisis diskriminan adalah bagian teknik statistik multivariat dalam mengklasifikasi objek menjadi beberapa atau banyak kelompok yang memiliki keterkaitan. Tujuannya agar objek pengamatan dapat dipisahkan. Selain itu, agar dapat mengalokasikan/memprediksi objek pengamatan yang baru kedalam kelompok yang telah didefinisikan (Supartini, et. al, 2017).

Tantangan utama proses klasifikasi adalah ketepatan dalam menentukan objek pada kelas atau kategori yang sudah ditentukan sebelumnya, Klasifikasi akan membangun model atau aturan klasifikasi agar dapat digunakan untuk tahap prediksi (Fatmawati & Rifai, 2023).

Analisis diskriminan memiliki p variabel independen dengan g kelompok dalam variabel dependen dan n amatan ($\sum_{k=1}^g n_k = n$). Fungsi kepekatatan peluang dari data yang menyebar normal multivariat dengan rata-rata μ dan matriks varians kovarian Σ , sehingga didapatkan fungsi diskriminannya seperti berikut (Johnson & Wichern, 1998):

$$Y = a'X \quad \dots(1)$$

$$Y = a_1X_1 + a_2X_2 + \dots + a_pX_p$$

Dengan Y = Skor diskriminan, a_p = Bobot diskriminan variabel ke- p , X_p = Prediktor/ variabel independen ke- p , p = Banyaknya variabel independen, dan $i = 1, 2, \dots, n$

Untuk mengetahui objek baru termasuk kedalam kelompok tertentu, maka dibutuhkan nilai *cut off* dengan rumus (Pane, 2015):

$$Y_{CU} = \frac{(N_A Y_A + N_B Y_B)}{N_A + N_B} \quad \dots(2)$$

Dengan Y_{CU} = Nilai kritis, Y_A = *Centroid* kelompok A, Y_B = *Centroid* kelompok B, N_A = Jumlah kelompok A, dan N_B = Jumlah kelompok B. Apabila nilai $Y \geq Y_{CU}$ maka objek baru termasuk ke dalam kelompok 1 (tidak terkena GDM), dan sebaliknya bila nilai $Y < Y_{CU}$ maka objek baru termasuk ke dalam kelompok 2 (terkena GDM). Dalam analisis diskriminan

ada pengujian asumsi yang harus dipenuhi, diantaranya:

Uji Normalitas Multivariat, menggunakan statistik uji jarak mahalanobis (d_i^2) (Supartini, et. al, 2017):

Hipotesis:

H_0 : Sebaran data berdistribusi normal multivariat

H_1 : Sebaran data tidak berdistribusi normal multivariat

Taraf nyata: $\alpha = 5\%$

Statistik uji:

$$d_i^2 = (X_i - \bar{X})^t S^{-1} (X_i - \bar{X}) \quad \dots(3)$$

$$q_i = \chi_{p;p_i}^2 = \chi_{p;(i-1)/n}^2 \quad \dots(4)$$

Kriteria uji: H_0 ditolak apabila ada lebih dari 50% nilai $d_i^2 < \chi_{(p;\alpha)}^2$ dengan p adalah banyaknya variabel independen, atau nilai d_i^2 dan q_i diplotkan berada di sekitar garis lurus

Uji Homogenitas Varians, menggunakan statistik uji *M'Box's*, asumsi akan terpenuhi apabila matriks varians kovarian homogen (Kusumawati, 2015):

Hipotesis:

H_0 : $\Sigma_1 = \Sigma_2 = \dots = \Sigma_g$: Kategori dari variabel dependen memiliki matriks varians-kovarian yang homogen.

H_1 : Minimal ada dua kategori dari variabel dependen memiliki matriks varians-kovarian yang berbeda (heterogen)

Taraf nyata: $\alpha = 5\%$

Statistik uji:

$$C = (1 - u)M \quad \dots(5)$$

$$u = \left[\sum_{i=1}^g \frac{1}{(n_i - 1)} - \frac{1}{\sum_{i=1}^g (n_i - 1)} \right] \frac{2p^2 + 3p - 1}{6(p + 1)(g - 1)} \quad \dots(6)$$

$$M = \left[\sum_{i=1}^g (n_i - 1) \right] \ln |S_{pooled}| - \sum_{i=1}^g [(n_i - 1) \ln |S_i|] \quad \dots(7)$$

Kriteria uji: H_0 diterima apabila nilai $C \leq \chi_{\frac{p(p+1)(g-1)}{2}, \alpha}^2$ atau jika $pvalue > \alpha$.

Uji Vektor Rata-rata, asumsi akan terpenuhi apabila ada perbedaan vektor rata-rata rumus (Supartini, et. al, 2017):

Hipotesis:

H_0 : $\mu_1 = \mu_2 = \dots = \mu_g$; Semua kelompok memiliki nilai vektor rata-rata yang sama

H_1 : Minimal ada dua kelompok yang berbeda

Taraf nyata: $\alpha = 5\%$

Statistik uji:

$$V = - \left[(n - 1) - \frac{p + g}{2} \right] \ln (\wedge), \quad \dots(8)$$

$$(\wedge) = \frac{|W|}{|B + W|} \quad \dots(9)$$

Kriteria uji: H_0 ditolak apabila nilai $V > X_{p(g-1), (1-\alpha)}^2$, atau jika $pvalue > \alpha$.

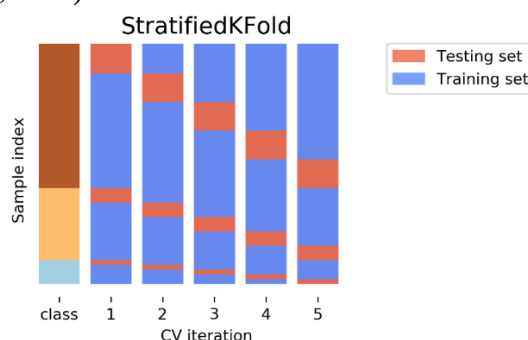
Untuk mengetahui variabel yang berbeda secara signifikan, maka dapat dilakukan pengujian F-Parsial dengan rumus:

$$F = (v - p + 1) \frac{T_p^2 - T_{p-1}^2}{v + T_{p-1}^2} \quad \dots(10)$$

H_0 akan ditolak apabila $F \geq F_{1,v-p+1}$ atau $pvalue < \alpha$, artinya variabel independen tersebut ada perbedaan rata-rata yang signifikan.

Stratified k-Fold Cross Validation

Menurut Davidson & Hinkley dalam Supartini, et. al (2017) Salah satu metode evaluasi *classifier* yaitu *Stratified k-Fold Cross Validation*. Metode ini digunakan untuk memvalidasi keakuratan sebuah model dengan membagi data menjadi data latih untuk membentuk model, dan data uji untuk memvalidasi model. Dalam penelitian ini menggunakan metode *stratified k-fold cross validation*, karena metode ini lebih cocok digunakan pada data yang tidak seimbang proporsi antar kelompoknya (Widodo, et. al, 2022). Metode ini mirip seperti *k-Fold CV* hanya saja pada saat pengacakan, data dibagi kedalam *k fold* dengan komposisi jumlah satu kategori sama untuk setiap *fold*nya dan begitupun untuk kategori yang lainnya. Berikut ilustrasi metode *Stratified k-Fold CV* (Muller, 2020):



Gambar 1. *Stratified k-Fold Cross Validation*

Jumlah k yang seringkali digunakan yaitu 5 dan 10, semakin besar k maka bias dan ukuran perbedaan antara set pelatihan dan subset resampling semakin kecil (Kuhn & Johnson, 2013).

Uji Ketepatan Klasifikasi

Uji ketepatan klasifikasi diantaranya terdiri dari:

Uji Kesalahan Klasifikasi

Menurut Johnson & Wichern dalam Supartini, et. al (2017) menyatakan bahwa nilai besar kecilnya dari jumlah observasi yang salah pengklasifikasian oleh fungsi klasifikasi disebut dengan *Apparent Error Rate* (APER). Dengan tabel kesalahan klasifikasi seperti:

Tabel 1. Kesalahan Klasifikasi

Hasil Observasi (<i>actual class</i>)	Hasil Prediksi (<i>predicted class</i>)	
	Tidak	Ya
Tidak	TN	FP
Ya	FN	TP

Nilai yang bisa didapatkan yaitu:

$$APER = \frac{FP + FN}{TN + FP + FN + TP} \quad \dots(11)$$

$$Akurasi = 1 - APER \quad \dots(12)$$

$$F1\ score = 2 \times \frac{\frac{TP}{TP+FN} \times \frac{TP}{TP+FP}}{\frac{TP}{TP+FN} + \frac{TP}{TP+FP}} \quad \dots(13)$$

Nilai yang biasa digunakan untuk mengetahui seberapa baik informasi yang didapatkan yaitu dengan melihat APER terkecil atau akurasi terbesar, tetapi dalam dunia medis seringkali menggunakan nilai *F1 score* terbesar karena semakin tinggi nilainya (mendekati 1) maka mengindikasikan bahwa klasifikasi yang dihasilkan sudah baik, dan dalam dunia medis kesalahan klasifikasi akan berakibat fatal untuk pasien (Ting, 2011).

Uji Keakuratan

Uji keakuratan menggunakan kriteria kemungkinan proporsional (*proportional chance criterion*), hitung nilai Hit_{ratio} , C_{max} , dan C_{pro} dengan rumus (Supartini, et. al, 2017):

Hipotesis:

H_0 : Klasifikasi akurat

H_1 : Klasifikasi tidak akurat

Statistik uji:

$$Hit_{ratio} = \frac{TN + TP}{TN + FP + FN + TP} \quad \dots(14)$$

$$C_{max} = \left(\frac{n_{max}}{N} \right) \times 100\% \quad \dots(15)$$

$$C_{pro} = p^2 + (1 - p)^2 \quad \dots(16)$$

Kriteria uji: Dari pengujian keakuratan, H_0 akan diterima apabila $Hit_{ratio} > C_{max} > C_{pro}$ artinya hasil klasifikasi sudah akurat.

Uji Kekonsistenan

Uji kekonsistenan digunakan untuk mengalokasikan setiap sampel apakah sudah konsisten atau tidak akibat adanya perubahan/perbedaan jumlah sampel yang diteliti. Statistik uji yang digunakan yaitu *Q-Press* dengan rumus sebagai berikut (Supartini, et. al, 2017):

Hipotesis:

H_0 : Klasifikasi konsisten

H_1 : Klasifikasi tidak konsisten

Statistik uji:

$$Q\ Press = \frac{[N - (ng)]^2}{N(g - 1)} \quad \dots(17)$$

Kriteria uji: H_0 akan ditolak apabila nilai $Q\ Press < X^2_{(1,\alpha)}$, artinya keakuratan klasifikasi sudah konsisten, begitupun sebaliknya.

Variabel Penelitian

1. Gestational Diabetes Mellitus (GDM)

GDM merupakan kondisi hiperglikemia dimana nilai glukosa dalam darah ada diatas normal tetapi dibawah diagnostik diabetes dan terjadi selama masa kehamilan. risiko yang ditimbulkan dari GDM ini ialah adanya komplikasi selama kehamilan dan melahirkan (WHO). Menurut Kementerian Kesehatan, istilah GDM merujuk pada keadaan ibu saat masa kehamilan mengidap Diabetes Melitus (DM) tetapi sebelumnya tidak memiliki penyakit diabetes.

2. Visceral Adipose Tissue (VAT)

Visceral fat (VF) disebut juga sebagai lemak *intra abdominal* yang disimpan pada jaringan adiposa di daerah perut. Lemak ini menempel pada organ-organ di rongga perut di bawah kulit dan lemak *intramuscular* pada otot-otot rangka. VF seringkali dikaitkan dengan obesitas, terutama obesitas sentral karena ditandai dengan menumpuknya lemak di daerah abdominal yang berakibat keadaan sindrom metabolik seperti DM (Rahmania, 2022).

3. Usia Kehamilan

Dalam Paputungan, et. al (2016), kehamilan menurut *International Federation of Gynecology and Obstetrics* (FIGO) ialah proses pembuahan dan dilanjutkan dengan proses nidasi yaitu bersarangnya zigot/sel telur yang telah dibuahi di rahim. Pada kasus ibu hamil penderita GDM terjadi pada saat usia kehamilan 24 minggu dan sebagian dari mereka kembali normal setelah melahirkan (Rahayu, 2016).

4. Kadar Gula Darah

Kadar gula darah yang disertai dengan gangguan metabolisme karbohidrat dapat menjadi tanda adanya gangguan metabolisme kronis suatu penyakit seperti DM (Syahputra, 2020). Kadar gula darah yang normal berkisar antara 70-110 mg/dl, apabila terus mengalami peningkatan, maka akan menjadi pemicu pankreas untuk memproduksi insulin, sehingga sebagian dari gula darah tersebut masuk ke berbagai macam sel untuk digunakan sebagai bahan energi dan sisanya diubah menjadi protein dan lemak. Apabila tubuh kekurangan insulin maka akan timbul hiperglikemia/DM.

5. Body Mass Index (BMI)

Indeks Massa Tubuh atau Body Mass Index (BMI) digunakan untuk menyatakan pembagian berat badan (kg) dengan tinggi badan (m) yang dapat digunakan untuk penentuan status gizi. BMI berkaitan dengan kadar gula darah pada penderita DM (Adnan, et. al, 2013). BMI dapat dihitung menggunakan rumus di bawah ini (Kemenkes, 2023):

$$BMI = \frac{\text{Berat badan (kg)}}{\text{Tinggi badan (m)} \times \text{Tinggi badan (m)}} \quad \dots(18)$$

Data Penelitian

Data diambil dari website www.physionet.org mengenai kasus GDM ibu hamil di Rumah Sakit XYZ, dengan variabel yang diteliti seperti berikut:

Tabel 2. Variabel Penelitian

Variabel	Notasi	Jenis data	Nama Variabel
Independen	X ₁	Rasio	<i>Visceral Adipose Tissue (VAT)</i>
	X ₂	Rasio	Usia kehamilan
	X ₃	Rasio	Kadar gula darah
	X ₄	Rasio	<i>Body Mass Index (BMI)</i>
Dependen	Y	Nominal	<i>Status Gestational Diabetes Mellitus (GDM)</i>

Kelima variabel tersebut diambil dari setiap ibu hamil sebanyak 133 orang. Berikut metode analisis untuk mencapai tujuan dari penelitian:

1. Eksplorasi data,
2. Menguji asumsi dasar yang harus terpenuhi terdiri dari:
 - a. Uji normalitas multivariat,
 - b. Uji homogenitas varians,
 - c. Uji Vektor Rata-rata,
3. Tentukan banyaknya k yang akan digunakan, biasanya k yang digunakan adalah 5 dan 10. Dalam penelitian ini menggunakan $k = 5$,
4. Melakukan pengacakan secara bertahap menggunakan *Stratified k-Fold Cross Validation* menjadi data latih dan data uji, pengacakan dilakukan dengan membagi data sebanyak k fold pada masing-masing kelompok tanpa pengembalian,
5. Melakukan eksperimen sebanyak k kali (5 kali), dengan 1 fold sebagai data uji dan $k - 1$ fold sebagai data latih,
6. Mengestimasi fungsi/model diskriminan untuk setiap eksperimen,
7. Menguji ketepatan klasifikasi yang terdiri dari:
 - a. Uji kesalahan klasifikasi dengan menggunakan nilai APER, dan F1 score,
 - b. Uji keakuratan menggunakan analisis kriteria kemungkinan proporsional

(*proportional chance criterion*),

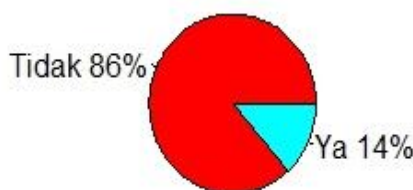
c. Uji kekonsistenan menggunakan *Q Press*.

Menentukan model terbaik menggunakan nilai APER yang paling kecil/jumlah kesalahan klasifikasi paling sedikit dan nilai *F1 score* yang paling besar.

C. Hasil Penelitian dan Pembahasan

Ekplorasi Data

Untuk mengetahui proporsi dari masing-masing kelompok dalam variabel dependen dapat dilihat dari Gambar 2.



Gambar 2. Ukuran Sampel setiap Kelompok

Ibu hamil dengan penyakit GDM lebih sedikit dibandingkan dengan yang tidak terkena GDM, dan terlihat perbedaan yang begitu jelas antara komposisi data dengan status status ibu hamil pengidap GDM dan tidak, oleh karena itu data yang digunakan dapat dikatakan tidak seimbang (*imbalance*). Sehingga cocok menggunakan *stratified k-fold cross validation*.

Pengujian Asumsi

1. Uji Normalitas Multivariat

Dari hasil perhitungan menggunakan jarak mahalanobis (d_i^2), banyaknya nilai $d_i^2 < \chi^2_{(p;\alpha)}$ dengan nilai $\chi^2_{(p;\alpha)} = \chi^2_{(4;0.05)} = 9.488$ yang persentasenya melebihi 50% dataset. Dari hasil perhitungan jarak mahalanobis ada sebanyak 106 dari 115 atau 92.17% pengamatan dengan nilai $d_i^2 < 9.488$. Sehingga dapat disimpulkan bahwa keempat variabel independen berdistribusi normal multivariat, maka asumsi terpenuhi.

2. Uji Homogenitas Varians

Dari hasil perhitungan menggunakan statistik uji *M-Box's* didapatkan nilai:

Tabel 3. Uji Homogenitas Varians

Statistik uji	C	$\chi^2_{10;0.05}$	$p\text{-value}$	Kesimpulan
<i>M-Box's</i>	16.9201	18.307	0.08166	H_0 diterima (Homogen)

Dari **Tabel 3**, nilai $16.9201 < 18.307$ dan nilai $0.08166 > 0.05$, maka H_0 diterima. Artinya matriks varians kovarian pada kategori/kelompok dalam variabel dependen sama (homogen) sehingga asumsi terpenuhi.

3. Uji Vektor Rata-rata

Hasil pengujian menggunakan statistik uji Wilks Lambda didapatkan nilai:

Tabel 4. Uji Vektor Rata-rata

Statistik uji	Λ	V	$\chi^2_{4;0.95}$	p -value	Kesimpulan
<i>Wilks Lambda</i>	0.81927	22.1268	0.711	0.0001895	H ₀ ditolak (Ada perbedaan secara signifikan)

Dari **Tabel 4**, nilai 22.1268 > 0.711 dan nilai 0.0001895 < 0.05, maka H₀ ditolak, artinya ada perbedaan vektor rata-rata pada variabel independen dalam mengelompokkan variabel dependen secara simultan sehingga asumsi terpenuhi.

Untuk mengetahui variabel mana yang signifikan dalam membedakannya maka digunakan statistik uji F-parsial. Dari hasil perhitungannya didapatkan nilai:

Tabel 5. Kesamaan vektor rata-rata masing-masing variabel

Variabel	F-parsial	p -value	Kesimpulan
X ₁	12.114	0.00071	H ₀ ditolak (Berbeda secara signifikan)
X ₂	0.041	0.83954	H ₀ diterima (Tidak berbeda secara signifikan)
X ₃	12.806	0.00051	H ₀ ditolak (Berbeda secara signifikan)
X ₄	11.923	0.00078	H ₀ ditolak (Berbeda secara signifikan)

Dari **Tabel 5**, dapat disimpulkan bahwa variabel yang tidak memiliki perbedaan rata-rata secara signifikan hanya X₂ (usia kehamilan), sedangkan variabel sisanya atau X₁ (VAT), X₃ (Gula darah), dan X₄ (BMI) memiliki perbedaan rata-rata yang signifikan.

Analisis Diskriminan dengan Metode Stratified 5 Fold Cross Validation

Pada penelitian ini banyaknya *fold* yang digunakan sebanyak 5 dengan masing-masing *fold* berisikan 23 pengamatan yang terpilih secara acak dengan banyaknya pengamatan dari kelompok 0 (tidak) ada 20 pengamatan dan dari kelompok 1 (ya) ada 3 pengamatan, hasilnya:

Tabel 6. Pengacakan & Pembagian Fold

<i>Fold</i>	Pengacakan Pengamatan	
	Kelompok 1	Kelompok 2
1	7, 10, 14, 17, 21, 22, 28, 31, 40, 54, 56, 60, 73, 75, 78, 81, 86, 87, 89, 93	107, 109, 110
2	8, 12, 13, 20, 29, 30, 33, 47, 48, 62, 63, 68, 69, 72, 77, 92, 94, 95, 98, 99	108, 113, 114
3	2, 9, 11, 18, 19, 24, 26, 32, 43, 45, 50, 53, 58, 65, 70, 71, 74, 76, 83, 97	101, 106, 112
4	1, 3, 4, 5, 6, 16, 34, 36, 41, 46, 49, 51, 55, 61, 64, 66, 79, 82, 88, 91	104, 105, 111
5	15, 23, 25, 27, 35, 37, 38, 39, 42, 44, 52, 57, 59, 67, 80, 84, 85, 90, 96, 100	102, 103, 115

Selanjutnya melakukan 5 kali eksperimen untuk mendapatkan model diskriminan dan nilai *cut off*. Dengan eksperimen pertama *fold* ke-1 sebagai data uji dan *fold* sisanya menjadi data latih, begitupun seterusnya hingga eksperimen kelima. Hasilnya didapatkan:

Tabel 7. Hasil Eksperimen

Eks	Model	Centroid		Cut Off
		Tidak	Ya	
1	$Y_1 = 0.0438X_1 - 0.1120X_2 + 0.0887X_3 + 0.0535X_4$	-0.184	1.228	0.000174
2	$Y_2 = 0.0327X_1 - 0.0909X_2 + 0.1078X_3 + 0.0561X_4$	-0.196	1.308	0.000174
3	$Y_3 = 0.0115X_1 + 0.0216X_2 + 0.0992X_3 + 0.0100X_4$	0.173	-1.155	-0.00022
4	$Y_4 = 0.0516X_1 - 0.0659X_2 + 0.0450X_3 + 0.0841X_4$	-0.182	1.210	-0.00043
5	$Y_5 = 0.0471X_1 - 0.1321X_2 + 0.0691X_3 + 0.0717X_4$	0.187	-1.249	-0.00030

Dengan interpretasinya yaitu pada eksperimen 1 mendapatkan model Y_1 , artinya jika variabel X_1 naik sebesar 1 mm maka akan menaikkan nilai Y_1 sebesar 0.0438 satuan jika variabel lain bernilai konstan. Jika variabel X_2 naik sebesar 1 minggu maka akan menurunkan nilai Y_1 sebesar 0.1120 satuan jika variabel lain konstan. Jika variabel X_3 naik sebesar 1 mg/dl maka akan menaikkan nilai Y_1 sebesar 0.0887 satuan jika variabel lain bernilai konstan. Dan jika variabel X_4 naik sebesar 1 kg/m maka akan menaikkan nilai Y_1 sebesar 0.0535 satuan jika variabel lain bernilai konstan. Nilai *cut off* = 0.000174 artinya jika nilai $Y_1 \geq -0.0003043$ maka pengamatan baru pada masuk ke dalam kategori tidak terkena GDM, dan sebaliknya. Begitupun interpretasi untuk model dan *cut off* pada eksperimen lainnya.

Uji Ketepatan Klasifikasi

1. Uji Kesalahan Klasifikasi

Untuk mengetahui ketepatan klasifikasi dapat dilihat dari nilai APER terkecil atau F1 score terbesar. Pada eksperimen 1 didapatkan *confusion matrix* untuk data latih seperti:

Tabel 8. Kesalahan Klasifikasi Data Latih Eksperimen 1

Hasil Observasi (<i>actual class</i>)	Hasil Prediksi (<i>predicted class</i>)		Jumlah Observasi
	Tidak	Ya	
Tidak	79	1	80
Ya	1	11	12

Dari **Tabel 8** didapatkan nilai APER, akurasi, dan F1 score hasilnya seperti berikut:

Tabel 9. Nilai APER, Akurasi, dan F1 score Semua Eksperimen

Eksperimen	APER	Akurasi	F1 score
1	0.0217	0.9783	0.8571
2	0.0761	0.9239	0.7952
3	0.0978	0.9022	0.7952

Eksperimen	APER	Akurasi	F1 score
4	0.0978	0.9022	0.7952
5	0.0978	0.9022	0.7952

Dari **Tabel 9** di atas, terlihat bahwa nilai APER terkecil serta nilai akurasi dan F1 score terbesar terdapat pada eksperimen 1, maka dari itu model dan hasil klasifikasi terbaik terdapat pada eksperimen 1.

1. Uji Keakuratan

Statistik uji keakuratan dapat dihitung dengan persamaan (14) sampai (16) hasilnya seperti:

Tabel 10. Nilai Hit_{ratio} , C_{max} , dan C_{pro} Semua Eksperimen

Eksperimen	Hit_{ratio}	C_{max}	C_{pro}
1	97.83%	86.9565%	77.3157%
2	92.39%	86.9565%	77.3157%
3	90.22%	86.9565%	77.3157%
4	90.22%	86.9565%	77.3157%
5	90.22%	86.9565%	77.3157%

H_0 akan diterima apabila nilai $Hit_{ratio} > C_{max} > C_{pro}$, karena keseluruhan nilai $Hit_{ratio} > C_{max} > C_{pro}$ maka H_0 diterima untuk setiap eksperimen untuk data latih dan uji. Artinya klasifikasi pada data latih dan data uji dari kelima eksperimen sudah akurat.

2. Uji Kekonsistenan

Uji kekonsistenan menggunakan statistik uji *Q-Press* yang dihitung menggunakan persamaan (17). Didapatkan nilai *Q-Press* untuk masing-masing eksperimen seperti berikut:

Tabel 11. Nilai Statistik Uji *Q-Press*

Eksperimen	<i>Q Press</i>
1	103.3130
2	85.2261
3	78.4782
4	78.4782
5	78.4782

Kriteria pengujiannya yaitu H_0 akan ditolak apabila nilai $Q\text{ Press} < \chi^2_{(1,\alpha)}$, dengan $\chi^2_{(1;0.05)} = 3.841$. Karena seluruh nilai $Q\text{ Press} > 3.841$, maka H_0 diterima untuk seluruh eksperimen. Artinya hasil klasifikasi ibu hamil untuk data latih dan data uji pada kelima eksperimen sudah konsisten.

D. Kesimpulan

Hasil penelitian mengenai klasifikasi *Gestational Diabetes Mellitus* (GDM) menggunakan analisis diskriminan dengan metode *stratified k-fold cross validation* pada ibu hamil di Rumah Sakit XYZ didapatkan 5 model diskriminan. Model terbaik yang didapat yaitu: $Y_1 = 0.0438 \text{ VAT} - 0.1120 \text{ Usia kehamilan} + 0.0887 \text{ Gula darah} + 0.0535 \text{ BMI}$. Jika ada objek baru dengan nilai $Y_1 \geq 0.0001739$ maka ibu hamil tersebut masuk ke dalam kategori tidak terkena GDM, dan sebaliknya apabila $Y < 0.0001739$ maka ibu hamil masuk ke dalam kategori terkena GDM.

Akurasi pada data latih terbesar terdapat pada eksperimen 1 sebesar 97.83%. Hasil pengklasifikasian dengan menggunakan $k = 5$ untuk semua eksperimen akurasi sudah sangat baik dan konsisten.

Acknowledge

Penulis ucapkan kepada keluarga dan Ibu Nur Azizah Komara Rifai, S. Si., M. Si selaku dosen pembimbing yang telah memberikan bimbingan dan arahan dalam menulis penelitian.

Daftar Pustaka

- [1] Adnan, Miftahul., Mulyati, Tatik., Isworo, Joko Teguh. (2013). Hubungan Indeks Massa Tubuh (IMT) dengan Kadar Gula Darah Penderita Diabetes Mellitus (DM) Tipe 2 Rawat Jalan di RS Tugurejo Semarang, **2** (1), 18-24.
- [2] Fatmawati., Rifai, Nur Azizah Komara. 2023. Klasifikasi Penyakit Diabetes Retinopati menggunakan Support Vector Machine dengan Algoritma Grid Search Cross-Validation, **3** (1), 79-86.
- [3] Johnson, RA., Winchern, DW. (1998). *Applied to Multivariate Analysis Sixth Edition*. New York: John Willey & Sons.
- [4] Kementerian Kesehatan. (2022). Mengenal Diabetes pada Kehamilan, (Online), (https://yankes.kemkes.go.id/view_artikel/1135/mengenal-diabetes-pada-kehamilan, diakses 9 Juni 2023).
- [5] Kementerian Kesehatan. (2023). IMT (Indeks Massa Tubuh = BMI (Body Mass Index)), (Online), ([https://www.kemkes.go.id/index.php?txtKeyword=status+gizi&act=search-by-map&pgnumber=0&charindex=&strucid=1280&fullcontent=1&C-ALL=1#:~:text=Salah%20satu%20cara%20untuk%20menentukan,%20FTB2%20\(dalam%20meter,](https://www.kemkes.go.id/index.php?txtKeyword=status+gizi&act=search-by-map&pgnumber=0&charindex=&strucid=1280&fullcontent=1&C-ALL=1#:~:text=Salah%20satu%20cara%20untuk%20menentukan,%20FTB2%20(dalam%20meter,) diakses 26 Juni 2023).
- [6] Mardiana, Linda., Kusnandar, D., Satyahadewi, Neva. (2022). Analisis Diskriminan dengan K-Fold Cross Validation untuk Klasifikasi Kualitas Air di Kota Pontianak, **11**, 97-102.
- [7] Maulidya., Susanto, H.T., Oktaviarina, Affiati. (2013). Perbandingan Analisis Diskriminan dan Regresi Logistik.
- [8] Muller, Andreas C. (2020). *Applied Machine Learning in Python*, (Online), (<https://amueller.github.io/aml/04-model-evaluation/1-data-splitting-strategies.html>, diakses 1 Juni 2023).
- [9] Pane, Rosmadewi Ayuningtyas. (2015). Analisis Diskriminan untuk Memprediksi Kebangkrutan Perusahaan (Studi pada Perusahaan Manufaktur yang Terdaftar di Bursa Efek Indonesia Tahun 2011-2013), **27** (2), 1-8.
- [10] Paputungan, Tiara V., Wagey, Freddy W., Lengkong, Rudy A. (2016). Profil Penderita Mola Hidatidosa di RSUP Prof. Dr. R. D. Kandou Manado. **4** (1), 215-222.

- [11] Rahayu, Anita., Rodiani. (2016). Efek Diabetes Melitus Gestasional terhadap Kelahiran Bayi Makrosomia, *5* (4), 17-22.
- [12] Rahmania, Puja. (2022). Hubungan Persentase Visceral Fat dengan Kejadian Diabetes Melitus Tipe 2 di Rumah Sakit Umum Haji Medan. Medan: Fakultas Kedokteran, Universitas Muhammadiyah Sumatera Utara.
- [13] Rosyid, Harun Al. (2017). Analisis Regresi dan Diskriminan untuk Menentukan Rapor serta Klasifikasi Penyelenggaraan Usaha Jasa Pengiriman. Surabaya: Program Magister, Departemen Teknik Elektro, Fakultas Teknologi Elektro, Insitut Teknologi Sepuluh September.
- [14] Supartini, I.A.M., Sukarsa, I K.M., Srinadi, I G.A.M. (2017). Analisis Diskriminan pada Klasifikasi Desa di Kabupaten Tabanan menggunakan Metode K-FoldCross Validation, *6*, 106-115.
- [15] Syahputra, Ismadi. (2020). Klasifikasi Penyakit Diabetes Melitus Tipe I dan Diabetes Melitus Tipe II dengan Menggunakan Metode Analisis Diskriminan. Medan: Program Studi Matematika, Fakultas Sains dan Teknologi, Universitas Islam Negeri Sumatera Utara.
- [16] Ting, Kai Ming. (2011). Precision and Recall. In: Sammut, C., Webb, G.I. (eds) *Encyclopedia of Machine Learning*. Springer, Boston, MA.
- [17] Widodo, Slamet., Brawijaya, Herlambang., Samudi. (2022). Stratified K-FoldCross Validation Optimization on Machine Learning for Prediction, *7*, 2407-2414.