Multivariat, MAS328, SD-A2, 2023

Klasifikasi Pasien Jantung Koroner Menggunakan Analisis Diskriminan pada *Cardiovascular Disease Dataset*

Aditya Ananda[a], Ziyan Iffatun Nadhira[b], Stevanus Sembiring[c]

[a]*162112133095 Teknologi Sains Data, Fakultas Teknologi Maju dan Multidisiplin, Universitas Airlangga, Surabaya*
[b]*162112133098 Teknologi Sains Data, Fakultas Teknologi Maju dan Multidisiplin, Universitas Airlangga, Surabaya*
[c]*162112133099 Teknologi Sains Data, Fakultas Teknologi Maju dan Multidisiplin, Universitas Airlangga, Surabaya*

Abstrak

Penelitian ini bertujuan untuk melakukan Analisis Diskriminan terhadap data indikator penyakit jantung pada pria di wilayah Western Cape, Afrika Selatan. *Dataset* diperoleh dari *dataset* "*Cardiovascular Disease*" di Kaggle. Tujuan utama dalam analisis diskriminan yang dilakukan yakni menentukan model prediksi yang dapat membedakan individu yang memiliki penyakit jantung koroner atau tidak berdasarkan variabel-variabel prediktor.

Penelitian ini memanfaatkan Analisis Diskriminan sebagai metode utama. Proses pengolahan data awal dilakukan dengan teknik *winsorizing* dan transformasi Yeo-Johnson. Selanjutnya, dilakukan pengecekan terhadap beberapa asumsi, termasuk uji normalitas variabel prediktor dengan variabel respons, multikolinearitas, dan homogenitas matriks kovarians. Dari sembilan variabel prediktor, enam diantaranya berdistribusi normal. Untuk mengatasi gejala multikolinearitas, satu dari enam variabel tersebut dihapus. Selain itu, matriks kovarians untuk kelima variabel prediktor telah sama. Tiga metode, yaitu *stepwise selection*, *forward elimination*, dan *backward elimination*, digunakan untuk seleksi variabel prediktor yang signifikan dalam model diskriminan.

Hasil penelitian menunjukkan bahwa empat variabel prediktor signifikan terhadap model diskriminan dengan *cutting score* 0,26 yaitu: tekanan darah sistolik, kadar kolesterol LDL, riwayat keluarga, dan tipe kepribadian A. Model diskriminan dalam bentuk transformasi Yeo-Johnson menghasilkan akurasi sebesar 0,7338, presisi 0,748, *recall* 0,894, serta F1-*Score* 0,8145 dengan menganggap pasien pengidap jantung koroner sebagai kelas negatif dan pasien bukan pengidap jantung koroner sebagai kelas positif.

Kata Kunci : Analisis Diskriminan, Jantung Koroner, Variabel Prediktor, Variabel Respons, Transformasi Yeo-Johnson

Abstract

This study aims to conduct a Discriminant Analysis on the indicators of heart disease in men in the Western Cape, South Africa. The dataset was obtained from the "Cardiovascular Disease" dataset on Kaggle. The primary objective of this discriminant analysis is to establish a predictive model that can distinguish individuals with coronary heart disease based on predictor variables.

The study utilizes Discriminant Analysis as the main method. The initial data processing was carried out using winsorizing techniques and Yeo-Johnson transformations. Subsequently, several assumptions were checked, including the normality test of predictor variables with response variables, multicollinearity, and homogeneity of covariance matrices. Out of nine predictor variables, six were normally distributed. To address the symptoms of multicollinearity, one of the six variables was removed. In addition, the covariance matrices for the five predictor variables were the same. Three methods, namely stepwise selection, forward elimination, and backward elimination, were used to select significant predictor variables in the discriminant model.

The study results indicate that four predictor variables are significant to the discriminant model with a cutting score of 0.26, namely: systolic blood pressure, LDL cholesterol levels, family history, and type A personality. The discriminant model in the form of Yeo-Johnson transformation produces an accuracy of 0.7338, precision 0.748, recall 0.894, and F1-Score 0.8145, considering patients with coronary heart disease as the negative class and non-patient individuals as the positive class.

Keywords: Discriminant Analysis, Coronary Heart Disease, Predictor Variables, Response Variables, Yeo-Johnson Transformations

1. Pendahuluan

Analisis Diskriminan adalah metode statistik yang digunakan untuk membedakan atau memprediksi kelompok variabel respons berdasarkan kombinasi variabel prediktor. Dalam konteks penyakit jantung koroner, Analisis Diskriminan dapat digunakan untuk mengklasifikasikan pasien menjadi dua kelompok, yaitu pasien yang menderita penyakit jantung koroner dan pasien yang tidak menderita penyakit jantung koroner, berdasarkan beberapa indikator seperti tekanan darah sistolik, kadar Low-Density Lipoprotein (LDL), riwayat keluarga, tipe kepribadian A, maupun tingkat obesitas.

Peran analisis diskriminan dalam penelitian penyakit jantung koroner sangat penting. Dengan menggunakan analisis diskriminan, peneliti dapat mengetahui variabel-variabel prediktor yang paling berpengaruh terhadap penyakit jantung koroner. Misalnya, analisis diskriminan dapat mengungkap bahwa kadar kolesterol LDL dan riwayat keluarga adalah dua variabel prediktor yang paling signifikan terhadap model diskriminan.

Manfaat melakukan Analisis Diskriminan adalah memungkinkan peneliti untuk menyederhanakan data dan fokus pada indikator-indikator yang paling berpengaruh terhadap penyakit jantung koroner. Dengan demikian, intervensi atau program pencegahan dapat dirancang dengan lebih efektif, dengan menargetkan indikator-indikator yang paling berkontribusi terhadap risiko penyakit. Selain itu, Analisis Diskriminan juga dapat digunakan untuk membuat model prediksi yang akurat, sehingga dapat membantu dalam deteksi dini penyakit jantung koroner.

2. Landasan Teori

2.1 ***Data Pre-processing***

2.1.1 Konsep Winsorizing untuk Handling Outliers

Winsorizing Tree adalah *robust method* yang direformasi dari CART[1]. Keuntungan dari metode ini adalah dapat menahan kumpulan data yang abnormal dan melindungi informasi asli dari data tersebut. Setiap node melakukan inspeksi *outlier* sebelum menghitung metrik. Nilai *outlier* yang terdeteksi akan diganti dengan nilai persentil tertentu. Proses ini terus dilakukan secara rekursif di setiap node hingga ambang tercapai. Secara umum ada dua jenis *Winsorizing* yaitu *Winsorizing* bawah dan *Winsorizing* atas. Jika nilai *outlier* yang lebih kecil dari persentil tertentu akan diganti dengan nilai persentil tersebut maka disebut *Winsorizing* bawah, dan jika lebih besar maka disebut *Winsorizing* atas[2].

2.1.2 Konsep Yeo-Johnson untuk Transformasi Data

Metode ini memiliki kesamaan dengan transformasi *Box-Cox,* yaitu mengubah data agar mendekati distribusi *Gaussian*, hanya nilai atribut bias sama dengan 0 atau nilai negatif[3]. Transformasi ini memiliki banyak keunggulan dibanding transformasi *Box-Cox*. Transformasi Box-Cox tidak cocok digunakan pada data yang memiliki nilai kurang dari atau sama dengan nol. Sedangkan Transformasi Johnson memungkinkan untuk nilai kurang dari nol. Transformasi Johnson memiliki perbedaan dengan transformasi Box-Cox namun memiliki kesamaan yaitu transformasi pada variabel respon. Transformasi Johnson memungkinkan untuk memilih fungsi dari tiga yang berbeda[4].

2.2 ***Pengujian Asumsi Analisis Diskriminan***

2.2.1 Pengujian Normalitas Variabel Prediktor dengan Faktor

2.2.1.1 Pengujian Kolmogorov-Smirnov

Kolmogorov-Smirnov merupakan salah satu metode pengujian normalitas[5]. Kelebihan dari uji ini adalah sederhana dan tidak menimbulkan perbedaan persepsi di antara satu pengamat dengan pengamat yang lain, yang mana hal ini sering terjadi pada uji normalitas dengan menggunakan grafik. Konsep dasar dari uji ini adalah dengan membandingkan distribusi data yang akan diuji dengan *Z-score*[6]. Jika signifikansi bernilai di atas 0.05 maka terdapat perbedaan yang signifikan antara data yang diuji dengan data normal baku. Artinya data yang diuji bersifat normal.

2.2.2 Pengujian Multikolinearitas dengan Korelasi antar Variabel Prediktor

2.2.2.1 Matriks Korelasi Pooled Within-Groups

Matriks korelasi pooled within-groups adalah matriks korelasi yang menghitung korelasi antara variabel-variabel dalam suatu kelompok, setelah dinormalisasi dengan menggunakan rata-rata dan varians dari masing-masing kelompok[7]. Matriks korelasi pooled within-groups digunakan dalam analisis

diskriminan untuk menguji asumsi homoskedastisitas. Homoskedastisitas adalah asumsi bahwa varians dari variabel-variabel dalam suatu kelompok adalah sama. Jika matriks korelasi pooled within-groups tidak memiliki korelasi yang signifikan, maka asumsi homoskedastisitas terpenuhi. Sebaliknya, jika matriks korelasi pooled within-groups memiliki korelasi yang signifikan, maka asumsi homoskedastisitas tidak terpenuhi.

2.2.2 Pengujian Homogenitas Matriks Kovarians

2.2.3.1 Box's M Test

Salah satu asumsi yang harus dipenuhi dalam membandingkan dua atau lebih vektor rata-rata adalah seluruh populasi memiliki matriks kovarians yang sama. Pengujian homogenitas ini dapat dilakukan dengan melakukan pengujian *Box's M*. Hipotesis yang digunakan adalah sebagai berikut[8]:

$H_0 : \Sigma_1 = \Sigma_2 = \ldots = \Sigma_g = \Sigma$ (matriks varians kovarian homogen)

H_1 : minimal ada satu $\boldsymbol{\Sigma}_l \neq \boldsymbol{\Sigma}$ untuk $l = 1,2, \ldots, g$ (matriks varians kovarian tidak homogen)

Hipotesis nol akan ditolak apabila nilai statistik uji

$$C > \chi^2_{p(p+1)(g-1)/2(\alpha)}$$

2.3 ***Analisis Diskriminan***

2.3.1 Signifikansi Variabel Prediktor terhadap Variabel Respons

Signifikansi variabel prediktor terhadap variabel respons adalah ukuran seberapa besar pengaruh variabel prediktor terhadap variabel respons[9]. Semakin signifikan variabel prediktor, semakin besar pengaruhnya terhadap variabel respons. Hasil uji signifikansi variabel prediktor biasanya dilaporkan dalam bentuk nilai p. Nilai p yang lebih kecil dari nilai kritis menunjukkan bahwa variabel prediktor signifikan secara statistik. Secara umum, signifikansi variabel prediktor terhadap variabel respons dapat dijelaskan sebagai berikut:

- Nilai $p < 0.05$: Variabel prediktor signifikan secara statistik pada tingkat kepercayaan 95%.
- Nilai $p < 0.01$: Variabel prediktor signifikan secara statistik pada tingkat kepercayaan 99%.
- Nilai $p > 0.05$: Variabel prediktor tidak signifikan secara statistik.

2.3.2 Konsep Variable Selection

Variable selection adalah proses memilih variabel-variabel yang paling relevan dan berpengaruh dari sekumpulan variabel yang tersedia untuk digunakan dalam model statistik[10]. Tujuannya adalah untuk membangun model yang lebih akurat, efisien, dan mudah diinterpretasikan. Variable selection dapat membantu mengurangi overfitting dengan menghilangkan variabel-variabel yang tidak relevan dan berisik. Pada penelitian ini metode *variable selection* yang digunakan adalah *stepwise. Stepwise variable selection* adalah metode *iterative* untuk memilih variabel yang paling relevan dan berpengaruh dalam model statistik[10]. Metode ini dimulai dengan model awal yang tidak memiliki variabel, dan kemudian secara bertahap menambahkan atau menghapus variabel berdasarkan kriteria tertentu.

2.3.3 Group Centroids and Cutting Score

Group centroids adalah nilai rata-rata dari variabel-variabel dalam suatu kelompok. *Cutting score* adalah nilai yang digunakan untuk membagi data ke dalam dua atau lebih kelompok[11]. *Group centroids* digunakan dalam analisis diskriminan untuk menghitung jarak antara kelompok-kelompok. Jarak ini digunakan untuk menentukan keanggotaan kelompok suatu objek. Cutting score digunakan dalam analisis diskriminan untuk menentukan kelompok mana suatu objek termasuk. Untuk menentukan keanggotaan kelompok suatu objek, analisis diskriminan menghitung jarak antara objek tersebut dan centroids dari masing-masing kelompok. Objek tersebut kemudian dikelompokkan ke dalam kelompok yang memiliki jarak terdekat.

2.3.4 Model atau Persamaan Diskriminan

Model atau persamaan diskriminan adalah suatu model statistik yang digunakan untuk memprediksi keanggotaan suatu objek ke dalam suatu kelompok[11]. Model ini didasarkan pada asumsi bahwa objek-objek dalam suatu kelompok memiliki karakteristik yang serupa, sedangkan objek-objek dalam kelompok yang berbeda memiliki karakteristik yang berbeda. Model diskriminan dapat digunakan

untuk berbagai tujuan, seperti pembagian data ke dalam dua atau lebih kelompok, prediksi hasil suatu penelitian, dan identifikasi objek yang memiliki karakteristik tertentu.

2.3.5 Evaluasi Hasil Klasifikasi

Setelah dilakukan pembangunan model diskriminan, perlu dilakukan evaluasi model terbentuk dengan menggunakan beberapa metrik pengukuran antara lain seperti akurasi, F1-Score, presisi, dan *recall* (senstivitas)[12]. Ukuran kebaikan model tersebut terdapat dalam elemen dari *Confusion Matrix*, berikut ini merupakan detail dari empat elemen yang ada dalam *confusion matrix*: True Positive (TP); elemen ini mengukur kasus ketika model dengan benar memprediksi kelas positif artinya, nilai sebenarnya adalah positif, dan model juga memprediksi sebagai positif, True Negative (TN); elemen kedua dalam *confusion matrix* mengukur kasus ketika model dengan benar memprediksi kelas negatif artinya, nilai sebenarnya adalah negatif, dan model juga memprediksi sebagai negatif, elemen ketiga yakni False Positive (FP): elemen ini adalah kasus ketika model salah memprediksi kelas positif artinya nilai sebenarnya adalah negatif tetapi model memprediksi sebagai positif di mana FP juga dikenal sebagai Kesalahan Tipe I, elemen terakhir yakni False Negative (FN): elemen ini adalah kasus ketika model salah memprediksi kelas negatif artinya, nilai sebenarnya adalah positif, tetapi model memprediksi sebagai negatif, FN juga dikenal sebagai Kesalahan Tipe II.

Adapun keempat ukuran kebaikan model dapat dijelaskan dan diformulasikan sebagai berikut;

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

akurasi merupakan rasio prediksi benar (positif dan negatif) dengan keseluruhan data. Akurasi merupakan metrik evaluasi yang paling intuitif dan mudah dipahami. Namun, akurasi tidak efektif jika data menunjukkan ketidakseimbangan kelas.

$$Presisi = \frac{TP}{TP+FP}$$

presisi merupakan rasio prediksi benar positif dibandingkan dengan total hasil yang diprediksi positif. Presisi juga disebut Nilai Prediktif Positif. Presisi sangat penting ketika biaya False Positive sangat tinggi.

$$Recall = \frac{TP}{TP+FN}$$

recall merupakan rasio prediksi benar positif dibandingkan dengan total aktual positif. Recall juga disebut sebagai True Positive Rate atau Sensitivitas. Recall sangat penting ketika biaya False Negative sangat tinggi.

$$F1\,Score = 2 \times \frac{presisi \cdot recall}{presisi+recall}$$

F1-Score merupakan rata-rata harmonik dari Presisi dan Recall. F1-Score mencoba mencari keseimbangan antara presisi dan recall. F1-Score sangat penting ketika kita ingin mencari keseimbangan antara Presisi dan Recall dan ada ketidakseimbangan kelas yang tidak merata.

3. Sumber Data dan Metodologi

3.1 Sumber Data

Data untuk analisis ini diambil dari dataset "*Cardiovascular Disease Dataset*" di Kaggle, yang disediakan oleh Yassine Hamdaoui[10]. Dataset ini mencakup sampel 462 pria dari wilayah di Western Cape, Afrika Selatan, yang memiliki risiko tinggi terhadap penyakit jantung. Dalam dataset ini, ada sekitar dua insiden jantung koroner setiap harinya. Sebagian besar pasien dengan jantung koroner telah menerima terapi untuk menurunkan tekanan darah dan program lainnya untuk mengurangi faktor risiko penyakit mereka. Dalam beberapa situasi, pengukuran dilakukan setelah pengobatan telah diberikan. Dari sepuluh variabel yang ada, diambil sembilan variabel yang bertindak sebagai prediktor dan satu variabel sebagai respons untuk analisis selanjutnya. Berikut keterangan variabel penelitian;

Tabel 1. Variabel atau Indikator Penelitian

Variabel	Keterangan	Peranan	Jenis
X1	*(sbp) Systolic blood pressure*	Prediktor	Metric
X2	*(tobacco) Cumulative tobacco (kg)*	Prediktor	Metric
X3	*(ldl) Low density lipoprotein cholesterol*	Prediktor	Metric
X4	(*adiposity)*	Prediktor	Metric
X5	*(famhist) Family history of heart disease*	Prediktor	Nonmetric
X6	*(typea) Type-A behavior*	Prediktor	Metric
X7	(*obesity)*	Prediktor	Metric
X8	*(alcohol) Current alcohol consumption*	Prediktor	Metric
X9	*(age) Age at onset*	Prediktor	Metric
Y	*(chd) Coronary heart disease*	Respons	Nonmetric

3.2 Metodologi

Tahapan Analisis Diskriminan dilakukan dengan bantuan software Python dan SPSS di mana Python digunakan untuk *data preprocessing* sedangkan SPSS digunakan untuk analisis diskriminan, adapun tahapan yang dilakukan meliputi:

3.2.1 Data Preprocessing

Sebelum melakukan proses analisis lebih lanjut, diperlukan pra-pemrosesan data terlebih dahulu dengan tujuan menyiapkan data agar dapat memenuhi beberapa asumsi analisis diskriminan.

3.2.2 Pengujian Asumsi Analisis Diskriminan

3.2.1 Pengujian Normalitas Variabel Prediktor dengan Faktor

Sebelum dilakukannya analisis diskriminan, dilakukan pengujian asumsi normalitas variabel prediktor dengan faktor/ variabel respons menggunakan uji Kolmogorov-Smirnov.

3.2.2 Pengujian Multikolinearitas

Pengujian asumsi selanjutnya yang dilakukan yaitu melihat korelasi antar variabel prediktor atau pengujian multikolinearitas dengan melihat hasil matriks korelasi *pooled within-groups*.

3.2.3 Pengujian Homogenitas Matriks Kovarians

Pengujian asumsi terakhir yang akan diuji yaitu kesamaan matriks varians-kovarians variabel-variabel prediktor dengan menggunakan metode Box's M Test.

3.2.3 Analisis Diskriminan

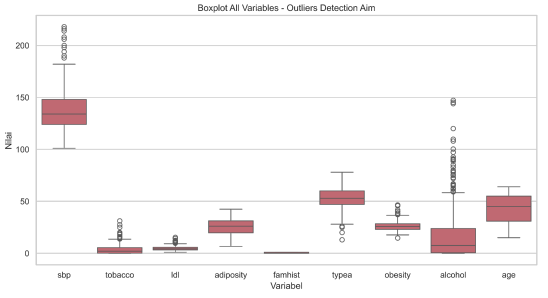
Tahapan pertama dalam melakukan analisis diskriminan yaitu melihat bagaimana kepengaruhan atau signifikansi yang terjadi antara variabel prediktor dengan kelas dalam variabel respon/ level dari faktor, kemudian melihat hasil penyeleksian variabel dengan menggunakan metode *Stepwise* sehingga dapat terlihat variabel prediktor mana yang signifikan terhadap variabel respons yang nantinya digunakan untuk membentuk model atau persamaan diskriminan, langkah berikutnya yaitu menentukan *cutting score* yaitu sebagai suatu nilai; penentu atau pembanding dengan menggunakan informasi dari *group centroids* serta jumlah tiap anggota grup atau level faktor. Tahapan berikutnya yakni menentukan model atau persamaan diskriminan, di mana model atau persamaan ini akan digunakan untuk menghitung dan membandingkan nilai diskriminan setiap observasi dengan *cutting score* sehingga suatu observasi dapat dikelompokkan/ diklasifikasikan. Langkah terakhir yakni melakukan evaluasi terhadap model diskriminan atau hasil klasifikasi (seperti penentuan akurasi dll).

4. Analisis dan Pembahasan

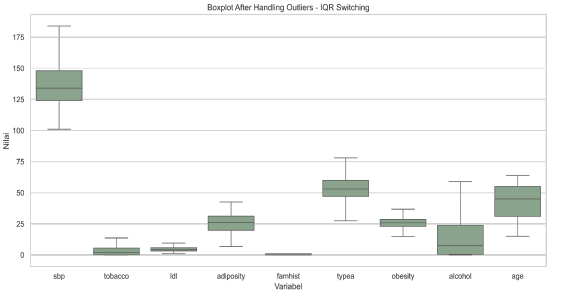
4.1 Data Preprocessing

4.1.1 Handling Outliers dengan Winsorizing Method

Metode penanganan outlier yang digunakan adalah dengan mengganti nilai-nilai ekstrem (*outlier*) dengan nilai batas atas (Q3 + 1.5 IQR) atau batas bawah (Q1 - 1.5 IQR), atau sering disebut dengan "winsorizing". Berikut adalah gambar box plot sebelum handling outlier dan sesudahnya:



Gambar 1. Box Plot semua variabel sebelum handling outlier

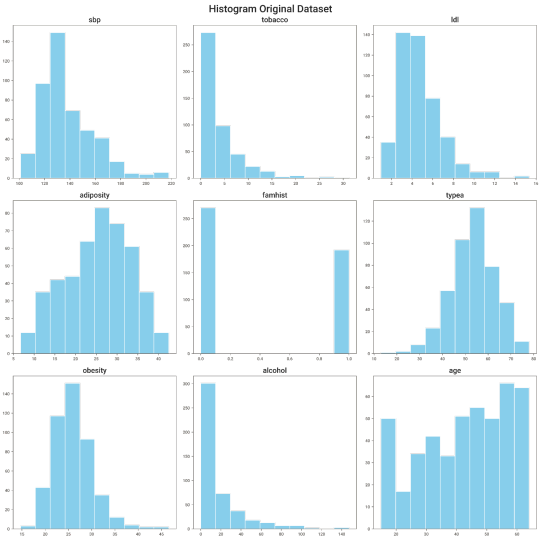


Gambar 2. Box Plot semua variabel setelah handling outlier

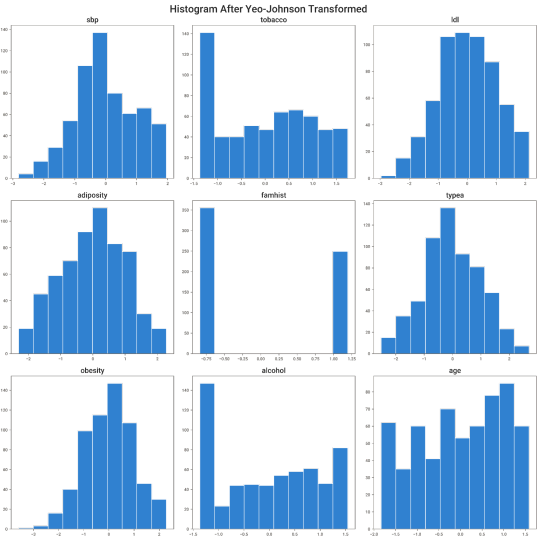
setelah dilakukan penanganan *outliers*, selanjutnya akan dilakukan transformasi terhadap data. Transformasi dilakukan sebagai upaya pengubahan bentuk distribusi menuju distribusi normal.

4.1.2 Yeo-Johnson Transformation Method

Transformasi data yang diterapkan yaitu metode Yeo-Johnson karena metode ini dapat mengubah bentuk distribusi beberapa variabel prediktor menjadi *bell-shaped*. Berikut ini komparasi distribusi variabel prediktor sebelum dan setelah dilakukan transformasi:



Gambar 3. Histogram Data Sebelum Transformasi



Gambar 4. Histogram Yeo-Johnsons Transformed

dengan menggunakan metode Yeo-Johnson ini dapat memperbaiki bentuk distribusi variabel prediktor menjadi *bell-shaped looks like*. Dapat dilihat bahwa enam distribusi variabel *sbp, ldl, adiposity, typea, obesity,* dan *age* setelah transformasi lebih berbentuk *bell-shaped*.

4.2 Pengujian Asumsi Analisis Diskriminan

4.2.1 Pengujian Asumsi Normalitas Variabel Prediktor dengan Faktor menggunakan Kolmogorov-Smirnov

H_0: *Data Berdistribusi Normal*

H_1: *Data Tidak mengikuti Distribusi Normal*

$\alpha : 0,05$

Tests of Normality

	chd	Kolmogorov-Smirnov[a] Statistic	df	Sig.	Shapiro-Wilk Statistic	df	Sig.
sbp	0	.062	302	.006	.991	302	.067
	1	.065	160	.099	.977	160	.008
ldl	0	.028	302	.200*	.995	302	.361
	1	.035	160	.200*	.986	160	.106
adiposity	0	.056	302	.024	.982	302	.001
	1	.048	160	.200*	.989	160	.216
typea	0	.041	302	.200*	.994	302	.227
	1	.042	160	.200*	.993	160	.671
obesity	0	.042	302	.200*	.993	302	.205
	1	.036	160	.200*	.989	160	.218
tobacco	0	.183	302	.000	.866	302	.000
	1	.108	160	.000	.927	160	.000
famhist	0	.434	302	.000	.586	302	.000
	1	.392	160	.000	.622	160	.000
alcohol	0	.142	302	.000	.903	302	.000
	1	.150	160	.000	.895	160	.000
age	0	.075	302	.000	.944	302	.000
	1	.120	160	.000	.934	160	.000

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

Gambar 5. Output SPSS - Uji Kolmogorov-Smirnov

Terlihat bahwa variabel prediktor ***ldl, typea,*** dan ***obesity*** memiliki p-value > 0.05 untuk kedua kelas variabel respons artinya ketiga variabel ini berdistribusi normal. Kemudian untuk kedua variabel; ***sbp*** dan ***adiposity*** hanya menunjukkan signifikan untuk kelas 1 atau pasien yang mengalami penyakit kardiovaskular namun keduanya dapat dianggap telah berdistribusi normal. Adapun untuk ketiga variabel *tobacco, alcohol,* maupun *age* tidak berdistribusi normal. ***Famhist*** tidak dapat ditentukan apakah berdistribusi normal atau tidak, karena variabel ini kualitatif. Dengan demikian, didapatkan enam (6) variabel prediktor yang akan dimasukkan dalam analisis diskriminan berikutnya.

4.2.2 Pengujian Multikolinearitas

Multikolinieritas terjadi ketika korelasi antar variabel prediktor melebihi 0,5. Berikut korelasi antar ke-enam variabel prediktor:

Pooled Within-Groups Matrices

		sbp	ldl	adiposity	famhist	typea	obesity
Correlation	sbp	1.000	.158	.349	.062	-.073	.272
	ldl	.158	1.000	.436	.116	.018	.377
	adiposity	.349	.436	1.000	.120	-.073	.759
	famhist	.062	.116	.120	1.000	.015	.104
	typea	-.073	.018	-.073	.015	1.000	.058
	obesity	.272	.377	.759	.104	.058	1.000

Gambar 6. Output SPSS - Pooled Within-Groups Matriks

Terdapat korelasi antara variabel prediktor *obesity* dan *adiposity* yang melebihi 0,5 yakni 0,759 sehingga dalam data ini terjadi kasus multikolinieritas, sehingga salah satu akan dihapus dari analisis yakni variabel prediktor *adiposity* karena variabel ini hanya berdistribusi normal terhadap data pasien kelas 1 atau pasien pengidap *chd*. Sehingga proses analisis diskriminan selanjutnya menggunakan lima (5) variabel prediktor; *sbp, ldl, famhist, typea,* serta *obesity*.

4.2.3 Pengujian Homogenitas Matriks Kovarians dengan Box's M Test

H_0: $\sum_1 = \sum_2 = \sum_3 = \ldots = \sum_5$

H_1: *Paling tidak ada dua (sepasang) matriks kovariansi yang tidak sama*

$\alpha : 0,05$

Test Results

Box's M		11.853
F	Approx.	1.173
	df1	10
	df2	501012.095
	Sig.	.304

Tests null hypothesis of equal population covariance matrices.

Gambar 7. Output SPSS - Box's M Test

Didapatkan P-Value(0,304) > α(0,05) sehingga Gagal Tolak H_0 artinya asumsi homogenitas matriks kovarians variabel-variabel prediktor terpenuhi atau dalam artian matriks varians-kovarians ke-lima variabel prediktor adalah sama.

4.3 Analisis Diskriminan

4.3.1 Signifikansi Variabel Prediktor terhadap Variabel Respons

4.3.1.1 Group Statistics

Group Statistics

chd		Mean	Std. Deviation	Valid N (listwise) Unweighted	Weighted
0	sbp	-.124310652	.9401923920	302	302.000
	ldl	-.199028464	.9769733513	302	302.000
	famhist	-.198253169	.9464376021	302	302.000
	typea	-.078210213	.9596557725	302	302.000
	obesity	-.072242112	.9833969327	302	302.000
1	sbp	.2346363555	1.071212783	160	160.000
	ldl	.3756662259	.9388856727	160	160.000
	famhist	.3742028565	.9971868187	160	160.000
	typea	.1476217778	1.062402833	160	160.000
	obesity	.1363569860	1.022886043	160	160.000
Total	sbp	.0000000000	1.001084011	462	462.000
	ldl	.0000000000	1.001084011	462	462.000
	famhist	.0000000000	1.001084011	462	462.000
	typea	.0000000000	1.001084011	462	462.000
	obesity	.0000000000	1.001084011	462	462.000

Gambar 8. Output SPSS - Group Statistics

Terlihat bahwa nilai rata-rata kelima variabel prediktor antara kelas 0 (tidak mengidap *chd*) dan kelas 1 (pengidap *chd*) berbeda secara signifikan, di mana seluruh nilai rata-rata pada kelas 0 bernilai negatif, sedangkan untuk kelas 1 seluruhnya bernilai positif. Hal ini menjadi dugaan awal bahwa terdapat perbedaan antara pasien tidak mengidap *chd* dan pengidap dari segi *ldl, famhist, typea,* dan *obesity.*

4.3.1.2 Tests of Equality Group Means

Tests of Equality of Group Means

	Wilks' Lambda	F	df1	df2	Sig.
sbp	.971	13.820	1	460	.000
ldl	.925	37.173	1	460	.000
famhist	.926	36.861	1	460	.000
typea	.988	5.373	1	460	.021
obesity	.990	4.576	1	460	.033

Gambar 9. Output SPSS - Tests of Equality Group Means

Dari hasil output di samping, terbukti bahwa ke-lima variabel prediktor signifikan terhadap variabel respons *chd* atau dalam artian ke-lima faktor *sbp, ldl, famhist, typea,* dan *obesity* berpengaruh dalam membedakan antara pasien pengidap jantung koroner dengan tidak.

4.3.2 Variable Selection

4.3.2.1 Metode Stepwise

Variables Entered/Removed[a,b,c,d]

Step	Entered	Wilks' Lambda Statistic	df1	df2	df3	Exact F Statistic	df1	df2	Sig.
1	ldl	.925	1	1	460.000	37.173	1	460.000	.000
2	famhist	.874	2	1	460.000	33.105	2	459.000	.000
3	sbp	.863	3	1	460.000	24.182	3	458.000	.000
4	typea	.854	4	1	460.000	19.461	4	457.000	.000

At each step, the variable that minimizes the overall Wilks' Lambda is entered.

a. Maximum number of steps is 10.

b. Maximum significance of F to enter is .05.

c. Minimum significance of F to remove is .10.

d. F level, tolerance, or VIN insufficient for further computation.

Gambar 9. Output SPSS - Stepwise Selection

Berdasarkan metode *stepwise selection*, didapatkan bahwa urutan variabel prediktor yang paling berpengaruh yaitu *ldl, famhist, sbp,* dan *typea*. Variabel *obesity* ternyata berdasarkan metode *stepwise* tidak signifikan terhadap model diskriminan karena p-valuenya melebihi 0,1. Selanjutnya akan dilihat dengan metode lain yakni *forward* dan *backward elimination*.

4.3.2.2 Metode Forward Elimination

Variables in the Analysis

Step		Tolerance	Sig. of F to Remove	Wilks' Lambda
1	ldl	1.000	.000	
2	ldl	.987	.000	.926
	famhist	.987	.000	.925
3	ldl	.964	.000	.905
	famhist	.985	.000	.911
	sbp	.973	.018	.874
4	ldl	.963	.000	.895
	famhist	.984	.000	.900
	sbp	.967	.012	.866
	typea	.994	.030	.863

Gambar 10. Output SPSS - Variables in the Analysis

Berdasarkan metode *forward elimination*, terlihat bahwa variabel prediktor pertama kali yang dimasukkan dalam model yaitu *ldl* hingga berlanjut ke-*step* 4 yang sama dengan hasil dari *stepwise selection* yakni terbentuk 5 variabel prediktor yang signifikan terhadap model diskriminan, antara lain; *ldl, famhist, sbp,* dan *typea*. Selanjutnya akan dilihat terkait proses bagaimana variabel *obesity* tidak termasuk dalam model diskriminan.

4.3.2.3 Metode Backward Elimination

Variables Not in the Analysis

Step		Tolerance	Min. Tolerance	Sig. of F to Enter	Wilks' Lambda
0	sbp	1.000	1.000	.000	.971
	ldl	1.000	1.000	.000	.925
	famhist	1.000	1.000	.000	.926
	typea	1.000	1.000	.021	.988
	obesity	1.000	1.000	.033	.990
1	sbp	.975	.975	.008	.911
	famhist	.987	.987	.000	.874
	typea	1.000	1.000	.034	.916
	obesity	.858	.858	.867	.925
2	sbp	.973	.964	.018	.863
	typea	1.000	.986	.047	.866
	obesity	.854	.852	.620	.873
3	typea	.994	.963	.030	.854
	obesity	.809	.809	.288	.861
4	obesity	.804	.804	.222	.852

Gambar 11. Output SPSS - Variables Not in the Analysis

pada output di samping, metode tersebut sama halnya dengan *backward elimination* yaitu menaruh seluruh variabel prediktor dalam model hanya saja biasanya metode *backward elimination* membuang variabel yang tidak signifikan pada model namun pada output tersebut malah sebaliknya di mana variabel signifikan yang dihapus dari tampilan. Dapat diketahui bahwa mengapa variabel *obesity* tidak signifikan terhadap model diskriminan yakni P-Value pada *step* ke-4 menunjukkan angka 0,288 di mana angka ini melebihi nilai signifikansi yang dapat masuk dalam model yakni lebih kecil dari 0,05.

4.3.3 Group Centroids and Cutting Score

Functions at Group Centroids

chd	Function 1
0	-.300
1	.566

Unstandardized canonical discriminant functions evaluated at group means

Gambar 12. Output SPSS - Functions at Group Centroids

Dari output di samping dapat diketahui bahwa nilai *centroid* untuk grup pasien yang tidak mengidap jantung koroner adalah -0,3 sedangkan untuk pasien pengidap jantung koroner adalah 0,566. Selanjutnya nilai ini disebut $Z_{A,B}$. Selanjutnya akan dihitung *cutting score* sebagai suatu nilai patokan dalam pengelompokkan objek atau dalam kasus ini adalah pasien;

$$Z_{CU} = \frac{N_A Z_B + N_B Z_A}{N_A + N_B}$$

di mana;
Z_{CU} : *cutting score untuk grup dengan ukuran sampel berbeda*
$N_{A,B}$: *jumlah anggota grup A atau B*
$Z_{A,B}$: *centroid anggota grup A atau B*

Perhitungan *cutting score* dilakukan di bawah: di mana jumlah pasien grup A adalah 302 sedangkan pasien grup B sebanyak 160 sehingga total 462 pasien.

$Z_{CU} = \frac{302(0{,}566)+160(-0{,}3)}{302+160} = \frac{170{,}932-48}{462} = \frac{122{,}932}{462} = 0,26608658$ (nilai ini masih dalam bentuk transformasi Yeo-Johnson)

4.3.4 Model atau Persamaan Diskriminan

Canonical Discriminant Function Coefficients

	Function 1
sbp	.317
ldl	.587
famhist	.619
typea	.267
(Constant)	.000

Unstandardized coefficients

Gambar 13. Output SPSS - Canonical Discriminant Function Coefficients

Persamaan diskriminan untuk menentukan seorang pasien pengidap jantung koroner atau tidak berdasarkan data yang dianalisis didapatkan;

$$D = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3 + b_4 X_4$$

$$D_{Yeo-Johnson} = 0 + 0,317\, sbp + 0,587\, ldl + 0,619\, famhist + 0,267\, typea$$

model tersebut dalam bentuk transformasi Yeo-Johnson.

4.3.4 Evaluasi Kebaikan Model Diskriminan

Classification Results[a]

			Predicted Group Membership		
		chd	0	1	Total
Original	Count	0	270	32	302
		1	91	69	160
	%	0	89.4	10.6	100.0
		1	56.9	43.1	100.0

a. 73.4% of original grouped cases correctly classified.

Gambar 14. Output SPSS - Classification Results

Berdasarkan *confusion matrix* di samping, didapatkan hasil kebaikan model diskriminan yang telah dibangun sebagai berikut (kelas positif adalah bukan pasien jantung koroner sedangkan kelas negatif adalah pasien jantung koroner):

$$accuracy = \frac{TP+TN}{TP+FP+FN+TN} = \frac{270+69}{270+91+32+69} = 0,7338$$

$$precision = \frac{TP}{TP+FP} = \frac{270}{270+91} = 0,7480$$

$$recall = \frac{TP}{TP+FN} = \frac{270}{270+32} = 0,8940$$

$$F1\ Score = 2 \times \frac{(prec \times rec)}{(prec+rec)} = 0,8145$$

4.3.4.1 Contoh Prediksi Data dengan Model Diskriminan

Akan dilihat bagaimana model diskriminan dalam menentukan setiap pasien termasuk memiliki penyakit jantung koroner atau tidak berdasarkan 5 data pasien teratas:

Tabel 2. Lima Data Pasien Teratas berdasarkan Variabel Prediktor Model

pasien	***sbp***	***ldl***	***famhist***	***typea***	***chd***
1	160	5,73	1	49	1
2	144	4,41	0	55	1
3	118	3,48	1	52	0
4	170	6,41	1	51	1
5	134	3,5	1	60	1

Tabel 3. Lima Data Pasien Teratas Hasil Transformasi Yeo-Johnson dan Hasil Prediksi

pasien	***sbp***	***ldl***	***famhist***	***typea***	***chd*** **(*true* label)**	**prediksi model**	**nilai diskriminan**
1	1.193790050443 21530	.673807583482 003200	1.1858541225 631425	-.46199763223 974720	1	1	1.38479
2	.4856660656812 41270	.005421097044 962458	-.8432740427 115679	.162002131055 05898	1	0	-0.32189
3	-1.15476410226 8325700	-.56628158698 7493800	1.1858541225 631425	-.15309394915 107846	0	0	-0.00463
4	1.558612747163 524700	.970775817950 641400	1.1858541225 631425	-.25676122319 221617	1	1	1.72932
5	-0.05808225888 7346150	-.55284214087 3777100	1.1858541225 631425	.700340248211 53270	1	0	0.57836

dari kelima data tersebut, dapat dilihat bahwa model diskriminan memprediksi kondisi pasien sesuai ke-empat variabel prediktor dengan benar sebanyak 2 dari 3 data. Diketahui bahwa cutting score model diskriminan yang terbentuk adalah $0,26608658$ sehingga pasien 1 dan 4 diidentifikasi oleh model sebagai pasien pengidap jantung koroner (nilai diskriminan$_{1.38;1.72} \geq$ cutting score), sedangkan pasien 2, 3, serta 5 diidentifikasi sebagai pasien tidak memiliki masalah kesehatan jantung koroner (nilai diskriminan$_{-.32;-.004;.57}$ < cutting score).

Kesimpulan

Sebelum dilakukan Analisis Diskriminan, terlebih dahulu melakukan *preprocessing* data dengan *winsorizing* dan transformasi Yeo-Johnson serta melakukan pengecekan asumsi - antara lain; normalitas variabel prediktor dengan faktor, multikolinearitas, dan homogenitas matriks kovarians setiap variabel prediktor dalam data indikator penyakit jantung koroner terseleksi. Didapatkan bahwa lima dari sembilan variabel prediktor berdistribusi normal yaitu: *sbp, ldl, famhist, typea,* serta *obesity*. Terdapat gejala multikolinearitas antara *obesity* dan *adiposity* kemudian diatasi dengan menghapus *adiposity*. Matriks kovarians telah sama untuk ke-empat variabel prediktor.

Setelah pengecekan asumsi Analisis Diskriminan dilakukan, analisis yang pertama yaitu melihat signifikansi variabel prediktor terhadap respons, didapatkan ke-empatnya signifikan. Kemudian dalam proses seleksi variabel didapati bahwa dari ketiga metode; *stepwise selection, forward elimination,* dan *backward elimination* menunjukkan variabel prediktor *obesity* tidak signifikan terhadap model diskriminan sehingga tersisa empat variabel. Berdasarkan *centroid* dan jumlah sampel grup pasien tidak memiliki masalah kesehatan jantung koroner dan pasien pengidap jantung koroner, didapatkan *cutting score* sebesar $0,26608658$ (dalam bentuk transformasi Yeo-Johnson). Adapun model diskriminan yang terbentuk yaitu:

$$D_{Yeo-Johnson} = 0 + 0,317\ sbp + 0,587\ ldl + 0,619\ famhist + 0,267\ typea$$

di mana model dalam bentuk transformasi tersebut menghasilkan akurasi sebesar 0,7338; presisi 0,748; *recall* 0,894; F1-*Score* 0,8145 dengan menganggap pasien pengidap jantung koroner sebagai kelas negatif dan pasien bukan pengidap sebagai kelas positif.

Daftar Pustaka

[1] A. Horsch, “Detecting and Treating Outliers In Python — Part 3,” Medium. Accessed: Dec. 08, 2023. [Online]. Available: https://towardsdatascience.com/detecting-and-treating-outliers-in-python-part-3-dcb54abaf7b0

[2] C. K. Ch’ng, “Comparing the Performance of Winsorize Tree to Other Data Mining Techniques for Cases Involving Outliers,” vol. 8, pp. 197–201, Aug. 2019, doi: 10.35940/ijrte.B1036.0782S219.

[3] K. S. V. Muralidhar, “How to transform the data to look like Gaussian Distribution?,” Medium. Accessed: Dec. 08, 2023. [Online]. Available: https://medium.datadriveninvestor.com/how-to-transform-the-data-to-look-like-gaussian-distribution-c50ab3fdada5

[4] S. Sumin, “Perbandingan Transformasi Box Cox Dan Transformasi Johnson Untuk Mengatasi Pelanggaran Asumsi Normalitas,” *Transformasi Box Cox Dan Transformasi Johnson*, Dec. 2015, Accessed: Dec. 08, 2023. [Online]. Available: https://digilib.iainptk.ac.id/xmlui/handle/123456789/2766

[5] “STATISTIK_PARAMETRIK_DENGAN_SPSS_DAN_INTERPRETASI_KELUARANNYA-libre.pdf.” Accessed: Dec. 08, 2023. [Online]. Available: https://d1wqtxts1xzle7.cloudfront.net/55637578/STATISTIK_PARAMETRIK_DENGAN_SPSS_DAN_INTERPRETASI_KELUARANNYA-libre.pdf?1516940622=&response-content-disposition=inline%3B+filename%3DStatistik_Parametrik_Menggunakan_Sofwere.pdf&Expires=1702014491&Signature=DA-yTLNtH3KDOtrSzahrKI-eeLWDSWOw0rC9zqzxDSbS8HS241irkAhPv5cGYztL2n-eVrYu5jc6KZUq4ILaIsC9Bhc~cJuaGGNF9PxBXpM64~K0kB5te~IIYtXEj~J2GBhPWF8EzvsBq7Bs5H8NhJCT9jjzF-Vv3K475cC8mO0BFS9oVuPZvKQR4EjBPksNkgDOKhV7~ZyvdA6nGAXC2pRpgcRxynS5eYgDCsNbCnlutad-r3wWNtEAooun-doOuw3BGMy1ASMqQbrepXKMOz4pOHQZYf5t9J9mm4wTS042Agl6Mu0rwEqWWwKHxteOuwZvNgfi4yoK2uhpQa~vCg__&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA

[6] A. Hidayat, “Penjelasan Rumus Kolmogorov Smirnov Uji Normalitas,” Uji Statistik. Accessed: Dec. 08, 2023. [Online]. Available: https://www.statistikian.com/2013/01/rumus-kolmogorov-smirnov.html

[7] E. Ardjmand, D. F. Millie, I. Ghalehkhondabi, W. A. Young Ii, and G. R. Weckman, “A State-Based Sensitivity Analysis for Distinguishing the Global Importance of Predictor Variables in Artificial

Neural Networks," *Adv. Artif. Neural Syst.*, vol. 2016, pp. 1–11, Aug. 2016, doi: 10.1155/2016/2303181.

[8] K. Khotimah, V. Ratnasari, and M. Ratna, "Pengelompokkan Kelurahan di Kota Surabaya Berdasarkan Kriteria Pembentukan Kampung Keluarga Berencana," *J. Sains Dan Seni ITS*, vol. 7, no. 2, Art. no. 2, Feb. 2019, doi: 10.12962/j23373520.v7i2.35272.

[9] R. A. Pane, "ANALISIS DISKRIMINAN UNTUK MEMPREDIKSI KEBANGKRUTAN PERUSAHAAN (Studi pada Perusahaan Manufaktur yang Terdaftar di Bursa Efek Indonesia".

[10] J. M. Wagner and D. G. Shimshak, "Stepwise selection of variables in data envelopment analysis: Procedures and managerial perspectives," *Eur. J. Oper. Res.*, vol. 180, no. 1, pp. 57–67, Jul. 2007, doi: 10.1016/j.ejor.2006.02.048.

[11] B. Simamora, "Analisis Diskriminan," Bilson Simamora Marketing and Research Center. Accessed: Dec. 08, 2023. [Online]. Available: https://www.bilsonsimamora.com/analisis-diskriminan/

[12] "Confusion Matrix," School of Computer Science. Accessed: Dec. 08, 2023. [Online]. Available: https://socs.binus.ac.id/2020/11/01/confusion-matrix/

LAMPIRAN

```python
# Written by Aditya Ananda
import pandas as pd
df2 = pd.read_csv('cardiovascular.csv',sep=';')
df2 = df2.drop(['ind'],axis=1)
df2

# Handling Outliers - IQR Detection and Switching Based on Outlier Location it's self)
# Import library
import numpy as np
# df2_ = df2.drop('chd',axis=1)
# Create a new DataFrame to store the results
df2_handled = df2.copy()

# Loop through each column in df2
for col in df2.columns:
    # Define Q1, Q3, and IQR
    Q1 = df2[col].quantile(0.25)
    Q3 = df2[col].quantile(0.75)
    IQR = Q3 - Q1

    # Define the upper and lower bounds
    lower_bound = Q1 - 1.5 * IQR
    upper_bound = Q3 + 1.5 * IQR

    # Replace outliers with the upper or lower bound in the new DataFrame
    df2_handled.loc[df2[col] < lower_bound, col] = lower_bound
    df2_handled.loc[df2[col] > upper_bound, col] = upper_bound

# Yeo-Johnson Transformation
from sklearn.preprocessing import PowerTransformer
pt = PowerTransformer(method='yeo-johnson')
df_yeo_johnson1 = pt.fit_transform(df2_handled.drop('chd',axis=1))
df_yeo_johnson1 = pd.DataFrame(df_yeo_johnson1, columns=df2_handled.drop('chd',axis=1).columns)
df_yeo_johnson1['chd'] = df2['chd']
df_yeo_johnson1

# Quick EDA
import sweetviz as sv
report = sv.analyze(df_yeo_johnson1)
report.show_notebook
```