STATISTIKA Multivariat dengan Program R

Analisis Multivariat adalah suatu metode pengolahan data dalam bentuk variabel yang jumlahnya banyak, dimana tujuannya adalah untuk mencari pengaruh variabel-variabel tersebut terhadap suatu obyek secara simultan atau serentak. Teori dari metode analisis multivariat dalam hal ini sebenarnya telah diketahui sejak lama, hanya saja karena cara perhitungannya yang rumit maka jarang sekali diterapkan. Perhitungan dalam analisis data multivariat lebih kompleks dibandingkan dengan analisis univariat, sehingga penggunaan program statistika akan mempermudah dalam analisis.

Menghitung dan menganalisis metode statistika multivariat dapat dilakukan dengan mudah menggunakan Program R. Program R merupakan sebuah software yang menyajikan variasi pengolahan data yang lengkap, powerful, dan gratis. Topik pembahasan dalam buku ini membahas teori dan contoh analisis statistika multivariat dilengkapi dengan sintax pengolahan data dengan menggunakan Program R.

Semoga buku ini dapat menjadi referensi bagi mahasiswa, dosen, dan peneliti dalam belajar mengajar maupun dalam analisis data suatu penelitian.

ISBN 978-602-498-815-9

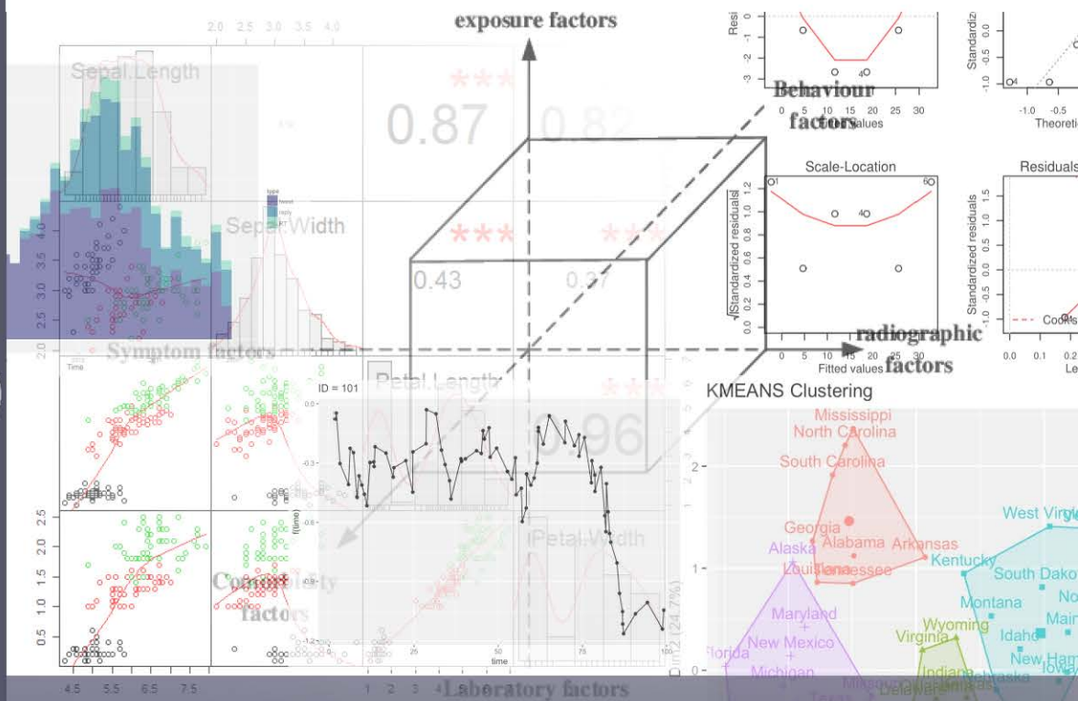
UNY PRESS
Komplek Fakultas Teknik UNY,
Jl. Karangmalang, Karang Gayam, Caturtunggal,
Kec. Depok, Kabupaten Sleman,
Daerah Istimewa Yogyakarta 55281

Statistika Multivariat dengan Program R

UNY PRESS

STATISTIKA Multivariat dengan Program R

Teori & Contoh Aplikatif



Heri Retnawati | Samsul Hadi | Kartianom | Krisna Merdekawati | Andi Harpeni Dewantara
Yustina Dwisofiani Lawung | Widayanti | Artina Diniaty | Yessica Mega Aprita
Widinda Normalia Arlianty | Intan Kemala Sari | Okta Alpindo | Mujiyanto
Primanisa Inayati Azizah | Rizqy Cahyo Utomo | Agung Prihantoro
Rizki Fitria Setyaningtyas | Fitriyani Hali | Rina Safitri

Statistika Multivariat dengan Program R: Teori dan Contoh Aplikatif

Penulis:
Heri Retnawati
Samsul Hadi
Kartianom
Krisna Merdekawati
Andi Harpeni Dewantara
Yustina Dwi Sofiani Lawung
Widayanti
Artina Diniaty
Yessica Mega Aprita
Widinda Normalia Arlianty
Intan Kemala Sari
Okta Alpindo
Mujiyanto
Primanisa Inayati Azizah
Rizqy Cahyo Utomo
Agung Prihantoro
Rizki Fitria Setyaningtyas
Fitriyani Hali
Rina Safitri

UNDANG-UNDANG REPUBLIK INDONESIA
NOMOR 28 TAHUN 2014
TENTANG HAK CIPTA

Pasal 2

Undang-undang ini berlaku terhadap:

a. semua Ciptaan dan produk Hak Terkait warga negara, penduduk, dan badan hukum Indonesia;
b. semua Ciptaan dan produk Hak Terkait bukan warga negara Indonesia, bukan penduduk Indonesia, dan bukan badan hukum Indonesia yang untuk pertama kali dilakukan Pengumuman di Indonesia;
c. semua Ciptaan dan/atau produk Hak Terkait dan pengguna Ciptaan dan/atau produk Hak Terkait bukan warga negara Indonesia, bukan penduduk Indonesia, dan bukan badan hukum Indonesia dengan ketentuan:
    1. negaranya mempunyai perjanjian bilateral dengan negara Republik Indonesia mengenai pelindungan Hak Cipta dan Hak Terkait; atau
    2. negaranya dan negara Republik Indonesia merupakan pihak atau peserta dalam perjanjian multilateral yang sama mengenai pelindungan Hak Cipta dan Hak Terkait.

BAB XVII
KETENTUAN PIDANA

Pasal 112

Setiap Orang yang dengan tanpa hak melakukan perbuatan sebagaimana dimaksud dalam Pasal 7 ayat (3) dan/atau pasal 52 untuk Penggunaan Secara Komersial, dipidana dengan pidana penjara paling lama 2 (dua) tahun dan/atau pidana denda paling banyak Rp300.000.000,00 (tiga ratus juta rupiah).

(1) Setiap Orang yang dengan tanpa hak melakukan pelanggaran hak ekonomi sebagaimana dimaksud dalam Pasal 9 ayat (1) huruf i untuk Penggunaan Secara Komersial dipidana dengan pidana penjara paling lama 1 (satu) tahun dan/atau pidana denda paling banyak Rp 100.000.000 (seratus juta rupiah).
(2) Setiap Orang yang dengan tanpa hak dan/atau tanpa izin Pencipta atau pemegang Hak Cipta melakukan pelanggaran hak ekonomi Pencipta sebagaimana dimaksud dalam Pasal 9 ayat (1) huruf c, huruf d, huruf f, dan/atau huruf h untuk Penggunaan Secara Komersial dipidana dengan pidana penjara paling lama 3 (tiga) tahun dan/atau pidana denda paling banyak Rp500.000.000,00 (lima ratus juta rupiah).
(3) Setiap Orang yang dengan tanpa hak dan/atau tanpa izin Pencipta atau pemegang Hak Cipta melakukan pelanggaran hak ekonomi Pencipta sebagaimana dimaksud dalam Pasal 9 ayat (l) huruf a, huruf b, huruf e, dan/atau huruf g untuk Penggunaan Secara Komersial dipidana dengan pidana penjara paling lama 4 (empat) tahun dan/ atau pidana denda paling banyak (1) (21 (3) Rp1.000.000.000,00 (satu miliar rupiah).
(4) Setiap Orang yang memenuhi unsur sebagaimana dimaksud pada ayat (3) yang dilakukan dalam bentuk pembajakan, dipidana dengan pidana penjara paling lama 10 (sepuluh) tahun dan/atau pidana denda paling banyak Rpa.000.000.000,00 (empat miliar rupiah).

Statistika Multivariat dengan Program R: Teori dan Contoh Aplikatif

Penulis:
Heri Retnawati
Samsul Hadi
Kartianom
Krisna Merdekawati
Andi Harpeni Dewantara
Yustina Dwi Sofiani Lawung
Widayanti
Artina Diniaty
Yessica Mega Aprita
Widinda Normalia Arlianty
Intan Kemala Sari
Okta Alpindo
Mujiyanto
Primanisa Inayati Azizah
Rizqy Cahyo Utomo
Agung Prihantoro
Rizki Fitria Setyaningtyas
Fitriyani Hali
Rina Safitri

2024

Statistika Multivariat dengan Program R: Teori dan Contoh Aplikatif

Penulis:

Penulis : Heri Retnawati, Samsul Hadi, Kartianom, Krisna Merdekawati, Andi Harpeni Dewantara, Yustina Dwi Sofiani Lawung, Widayanti, Artina Diniaty, Yessica Mega Aprita, Widinda Normalia Arlianty, Intan Kemala Sari, Okta Alpindo, Mujiyanto, Primanisa Inayati Azizah, Rizqy Cahyo Utomo, Agung Prihantoro, Rizki Fitria Setyaningtyas, Fitriyani Hali, Rina Safitri
Penyunting bahasa : Ibnu Rafi
Desain sampul : Rizqy Cahyo Utomo
Penata letak : Ibnu Rafi

Diterbitkan dan dicetak oleh:
UNY PRESS
Jl. Gejayan, Gg. Alamanda, Komplek Fakultas Tenik UNY
Kampus UNY Karangmalang Yogyakarta 55281
Telp : 0274-589346
E-Mail : unypenerbitan@uny.ac.id
Anggota Ikatan Penerbit Indonesia (IKAPI)
Anggota Asosiasi Penerbit Perguruan Tinggi Indonesia (APPTI)

15 × 23 cm, viii
ISBN: 978-602-498-815-9

Cetakan Pertama, Mei 2024

Kata Pengantar

Alhamdulillah, segala puji syukur hanya untuk Allah Subhanallahu Wa Ta'ala, yang telah melimpahkan anugerah kepada kita semua, dan juga kepada kami yang telah menyelesaikan penulisan buku ini. Buku ini merupakan buku ajar, yang dapat digunakan mahasiswa untuk belajar statistika dengan program R. Pada buku ini, setiap analisis yang dibahas dilengkapi dengan persamaan matematika, hipotesis yang dirumuskan jika ada, analisis dengan menggunakan program R di bawah *integrated development environment* (IDE) RStudio dan interpretasinya. Sintaks dan contoh datanya disertakan juga sehingga pembaca bisa mencobanya sendiri. Bab tentang analisis regresi dan analisis persamaan model struktural tidak disertakan pada buku ini, karena beririsan dengan bahan ajar lainnya.

Buku ini bisa terselesaikan atas dukungan berbagai pihak. Oleh karena itu berkenan kiranya jika kami menyampaikan ucapan terima kasih. Yang pertama, kami ucapkan terima kasih kepada bapak Rektor dan Wakil Rektor Universitas Negeri Yogyakarta atas Dana Hibah Penulisan Buku tahun 2023 yang membiayai penerbitan buku ini. Kami mengucapkan terima kasih juga kepada Bapak Direktur Sekolah Pascasarjana dan Bapak/Ibu Dekan, Ketua Departemen, dan Koordinator Program Studi di lingkungan UNY yang menugaskan kami untuk mengampu mata kuliah Statistik Multivariat yang mana ini telah memotivasi kami untuk menuliskan buku ini.

Terima kasih yang tak terhingga kami ucapkan kepada semua yang telah membantu penerbitan buku ini, termasuk penelaah yang telah memberikan masukan yang konstruktif pada konten buku ini. Kepada anggota tim penulis, terima kasih atas kesolidannya, hingga kita bisa sampai pada tahapan ini. Kepada pembaca pada umumnya, kami tunggu masukannya yang konstruktif. Semoga buku ini memberikan manfaat dan pahala jariah kepada kita semuanya.

Yogyakarta, 4 November 2023

Daftar Isi

Bab 1
Pengantar Analisis Multivariat

Pada suatu penelitian, terkadang peneliti tidak hanya menggunakan satu variabel saja untuk memecahkan masalahnya. Bisa saja peneliti mengelaborasi banyak variabel, lebih dari dua, bahkan sampai ratusan variabel. Banyaknya variabel yang peneliti gunakan ini akan menentukan jenis analisis data yang nantinya digunakan. Analisis data terkait dengan banyaknya variabel ini dapat dikategorikan menjadi analisis data univariat, bivariat, dan multivariat. Selain itu, ada dua jenis teknik analisis, yaitu deskriptif dan inferensial.

Analisis deskriptif dilakukan untuk merangkum dan mendeskripsikan ciri-ciri suatu kelompok atau membuat perbandingan ciri-ciri antar kelompok disebut statistik deskriptif. Statistik inferensial digunakan untuk membuat generalisasi atau kesimpulan tentang suatu populasi berdasarkan temuan dari suatu sampel, atau dengan kata lain, karakteristik populasi bisa dipahami melalui karakteristik sampel.

Analisis univariat digunakan ketika analisis masalah penelitian hanya pada satu variabel pada suatu waktu (misalnya tingkat pendidikan penduduk di suatu provinsi). Analisis bivariat dilakukan ketika penelitian terfokus pada dua variabel secara bersamaan (misalnya, hubungan antara waktu bekerja dan pendapatan seseorang atau waktu belajar dengan capaian prestasinya). Sementara itu analisis multivariat dilakukan ketika analisis data penelitian dilakukan pada lebih dari dua variabel secara bersamaan (misalnya, hubungan antara waktu bekerja, jenis kelamin dan pendapatan seseorang).

Analisis univariat melibatkan pemrosesan sebuah variabel saja dalam satu waktu. Analisis yang biasa terkait adalah distribusi frekuensi, baik yang dikelompokkan maupun yang tidak dikelompokkan, analisis proporsi, analisis ukuran tendensi sentral yang meliputi rerata, modus, median, dan ukuran penyebaran. Ukuran mencermin-

kan penyebaran atau sebaran distribusi, menggunakan parameter jangkauan (*range*), variansi atau ragam, dan deviasi standar atau simpangan baku. Jangkauan (*range*) adalah selisih antara skor terbesar dan terkecil atau skor tertinggi dan terendah. Sementara itu, variansi atau ragam adalah rata-rata selisih kuadrat antara setiap observasi dan rata-rata (*mean*). Akar kuadrat dari variansi tersebut kemudian disebut sebagai deviasi standar atau simpangan baku.

Analisis pada data penelitian yang melibatkan dua variabel disebut analisis bivariat. Contoh analisis bivariat adalah analisis korelasi dan analisis tabel kontingensi. Analisis korelasi bisa mengkorelasikan kedua variabel tersebut. Jika salah satu variabel merupakan variabel penyebab, dan salah satu variabel akibat, analisis dapat menggunakan regresi sederhana. Jika datanya tidak kontinu, tabel kontingensi dapat digunakan.

Analisis data multivariat memungkinkan analisis dengan melibatkan banyak variabel. Wilayah analisis multivariat meliputi pengenalan variabel dan data, regresi linear sederhana, korelasi, analisis regresi ganda, analisis jalur (*path analysis*), korelasi kanonis, analisis komponen utama (*principal component analysis*, PCA), analisis faktor, analisis klaster (*cluster analysis*), analisis korespondensi, regresi logistik, analisis data survival, analisis penskalaan multidimensional, dan persamaan model struktural (*structural equation modeling* atau disingkat SEM).

Regresi berganda cocok jika permasalahan penelitian melibatkan satu variabel terikat metrik yang dianggap terkait dengan dua atau lebih variabel bebas metrik (terkadang non-metrik). Tujuan analisis ini yaitu untuk memprediksi perubahan variabel terikat sebagai respons terhadap perubahan variabel bebas.

Analisis diskriminan membantu dalam membedakan dua atau lebih kumpulan objek atau orang berdasarkan pengetahuan tentang beberapa karakteristiknya. Analisis ini adalah teknik untuk menganalisis data ketika kriteria atau variabel terikat bersifat kategoris (karenanya non-metrik) dan prediktor atau variabel bebas bersifat metrik. Tujuan utama analisis diskriminan adalah untuk memahami perbedaan kelompok dan memprediksi kemungkinan bahwa suatu entitas (individu atau objek) akan menjadi bagian dari kelompok tertentu

berdasarkan beberapa variabel independen metrik. Contoh dari analisis ini yaitu membagi orang menjadi berpotensi termotivasi dan tidak termotivasi, mengklasifikasikan individu ke dalam prestasi siswa yang baik atau buruk, mengklasifikasikan sekolah sebagai layanan yang baik atau buruk.

Analisis berikutnya yaitu analisis konjoin (atau *conjoint analysis*). Analisis ini merupakan teknik analisis hubungan untuk menilai tingkat pemanfaatan konsumen terhadap atribut produk tertentu dan tingkatnya. Konsumen diharuskan mengevaluasi hanya pada beberapa profil produk yang merupakan kombinasi tingkat produk. Analisis ini dapat menjawab pertanyaan seperti kemanfaatan apa yang dilihat konsumen dalam tingkat harga, tingkat layanan purna jual, fitur produk, dan lainya. Analisis konjoin dapat digunakan dalam evaluasi produk atau layanan baru maupun yang sudah ada.

Analisis varians multivariat (sering disebut *multivariat analysis of varians*, MANOVA) tepat digunakan jika permasalahan penelitian melibatkan beberapa variabel terikat metrik yang dianggap bergantung pada satu atau lebih variabel bebas non-metrik (biasanya disebut perlakuan). Dengan MANOVA, uji signifikansi perbedaan rata-rata (*mean)* antar kelompok dapat dilakukan secara bersamaan untuk dua atau lebih variabel dependen. Sebagai contohnya yaitu pada penelitian eksperimen pembelajaran berbasis masalah (*problem-based learning*, PBL) dibandingkan dengan pembelajaran berbasis proyek (*project-based learning*, PjBL). Dengan memanipulasi PBL dan PjBL, peneliti dapat mengetahui pengaruhnya terhadap sikap, pengetahuan, dan keterampilan siswa. Demikian pula ketika mempertimbangkan variabel-variabel yang mempengaruhi pembelajaran di perguruan tinggi. Pengaruh jenis mutu perguruan tinggi terhadap persepsi, seperti biaya terjangkau, kebanggaan, dan motivasi calon mahasiswa untuk berprestasi.

Jenis analisis lain yang termasuk dalam analisis multivariat ialah analisis korelasi. Korelasi yang melibatkan banyak variabel ini disebut dengan analisis korelasi kanonis. Korelasi ini mengkorelasikan vektor dengan vektor, dimana elemen di tiap vektor boleh tidak sama. Sebagai contohnya yaitu hubungan beberapa kemampuan personal yang dikaitkan dengan kemampuan belajar bahasa. Kemampuan

personal seperti kedisiplinan, kemampuan awal, dan keterpaparan bahasa dikaitkan dengan kemampuan mendengarkan (*listening*), membaca (*reading*), berbicara (*speaking*), dan menulis (*writting*).

Analisis komponen utama (*principal component analysis*, PCA) ditujukan untuk menyederhanakan deskripsi dari sekumpulan variabel yang saling berhubungan. Pada analisis ini, semua variabel diperlakukan sama. Hasil analisis berupa variabel baru yang tidak berkorelasi yang disebut komponen utama. Masing-masing variabel baru merupakan kombinasi linier dari variabel aslinya. Ukuran informasi yang disampaikan masing-masing kombinasi linear ini adalah variansinya.

Salah satu penerapan dari PCA adalah pada analisis faktor. Analisis faktor ini digunakan untuk membedakan dimensi atau keteraturan yang mendasari fenomena. Tujuan umumnya adalah untuk meringkas informasi yang terkandung dalam sejumlah besar variabel menjadi sejumlah faktor yang lebih kecil. Sebagai contohnya yaitu kita mengidentifikasi kandungan bahan makanan yang ada di kota kita masing-masing. Tentu saja ada banyak sekali bahan makanan yang ada, misalnya gandum, wortel, telur, daging, beras/nasi, bayam, dan lain-lain. Kandungan tiap bahan makanan per 100 gram yang teridentifikasi selanjutnya disebut sebagai variabel. Dengan menggunakan variabel-variabel yang telah teridentifikasi, seperti karbohidrat, lemak, vitamin, protein, dan kandungan lain, selanjutnya dapat diketahui kandungan makanan dominan yang kemudian digunakan untuk mengelompokkan subjek. Hasil dari analisis ini yaitu bahwa variabel dominan yang ada yaitu karbohidrat, protein, dan serat. Setelah itu, bahan-bahan makanan tersebut dapat dikelompokkan menjadi bahan makanan yang dominan karbohidrat seperti gandum, kentang, nasi, dan lain-lain. Sama halnya untuk bahan-bahan makanan yang dominan mengandung protein, lemak, serat, dan yang lainnya.

Jika pada analisis faktor peneliti mengelompokkan variabelnya sedemikian sehingga menjadi lebih ringkas, melalui analisis klaster (*cluster analysis*) peneliti dapat mengelompokkan subjek penelitian menjadi beberapa kelompok memperhatikan variabel tertentu. Analisis klaster ini merupakan serangkaian teknik dengan tujuan mengklasifikasikan individu atau objek ke dalam sejumlah kecil kelompok

yang saling eksklusif, memastikan bahwa terdapat sebanyak mungkin persamaan dalam kelompok dan sebanyak mungkin perbedaan di antara kelompok. Analisis ini cocok untuk diterapkan dalam pengelompokan sekolah, mengelompokkan subjek berdasarkan kemiripannya. Pengelompokan tersebut melibatkan identifikasi indikator kualitas sekolah, misal kualitas guru, sarana-prasarana, pembelajaran, tenaga kependidikan, kultur, kepemimpinan, dan lain-lain.

Multidimensional scaling (MDS) merupakan suatu teknik statistik yang mengukur objek dalam ruang multidimensi berdasarkan penilaian responden terhadap kesamaan objek. Pada suatu penilaian, jika objek A dan B dinilai oleh responden sebagai yang paling mirip dibandingkan dengan semua kemungkinan pasangan objek lainnya, MDS (*perceptual mapping*) akan menempatkan objek tersebut paling dekat satu sama lain dalam hal jarak dalam peta multidimensi. Pengungkapan atribut setiap objek digunakan untuk membantu memahami mengapa objek dinilai serupa atau berbeda. Contohnya pemilik universitas A ingin mengetahui apakah pesaing terkuatnya adalah universitas B atau universitas C. Sampel pelanggan diminta menilai pasangan universitas dari yang paling mirip hingga yang paling tidak mirip. Hasil MDS menunjukkan B atau C termirip dengan A.

Teknik multivariat lain yang dapat digunakan yaitu analisis korespondensi berbeda dengan teknik saling ketergantungan lainnya dalam kemampuannya mengakomodasi data non metrik. Analisis ini menggunakan tabel kontingensi yang merupakan tabulasi silang dari dua variabel kategori. Contoh analisis ini adalah preferensi sekolah responden dapat ditabulasi silang berdasarkan variabel demografi (jenis kelamin, kategori pendapatan, dan pekerjaan) dengan menunjukkan berapa banyak orang yang lebih memilih sekolah yang termasuk dalam setiap kategori variabel demografis. Dengan menggunakan teknik analisis korespondensi, sekolah dan karakteristik sekolah yang disukai ditampilkan dalam peta dua dimensi atau tiga dimensi, baik merek dan karakteristik responden. Sekolah yang dianggap serupa letaknya berdekatan satu sama lain. Demikian pula karakteristik responden yang memilih sekolah tertentu juga ditentukan oleh kedekatan kategori variabel demografi dengan posisi sekolah.

Analisis lain berikutnya yaitu regresi logistik. Dengan menggunakan berbagai variabel, analisis regresi logistik dapat digunakan untuk mengklasifikasikan individu dalam salah satu dari dua populasi berdasarkan serangkaian kriteria. Analisis ini digunakan dengan input berupa kombinasi variabel diskrit atau kontinu. Proses analisis dapat menggunakan algoritma estimasi kemungkinan maksimum (*maximum likelihood*) untuk mengklasifikasikan individu berdasarkan daftar variabel independen. Contoh analisis ini yaitu memprediksi peluang hidup penderita COVID-19, misalnya. Beberapa gejala dan komorbid diidentifikasi, kemudian dimasukkan ke dalam fungsi untuk mengklasifikasikan kategori keselamatan peserta.

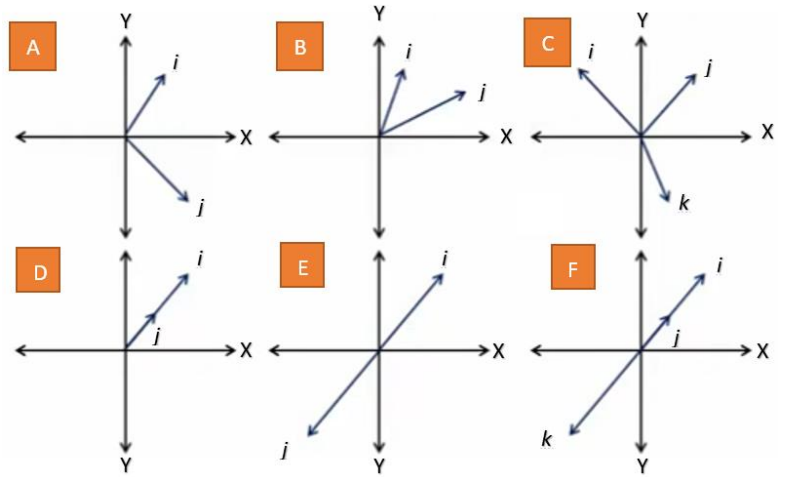
Analisis data ketahanan hidup (*survival data analysis*) juga merupakan salah satu teknik statistik multivariat. Analisis ini digunakan untuk keperluan prediksi. Analisis dilakukan dengan memperkirakan lamanya waktu yang diperlukan untuk terjadinya peristiwa tertentu. Sebagai contoh, dengan banyak variabel yang dijadikan prediktor, misalnya kebiasaan tidak sehat, merokok, adanya komplikasi, dan lain-lain, dapat diperkirakan saatnya kegagalan organ seseorang. Variabel yang dimasukkan pada fungsi ini disebut penjelas (*covariate*).

Ada banyak jenis analisis yang bisa digunakan peneliti. Pada implementasinya, tidak semua bisa diterapkan semuanya. Pemilihan jenis analisis didasarkan pada pertanyaan penelitian. Selain itu, jenis data yang dikumpulkan juga menentukan jenis analisisnya. Lebih lanjut, pengguna laporan dan yang akan menerima hasil penelitian menentukan jenis analisis mana yang digunakan. Untuk mempermudah dalam melakukan analisis multivariat, berbagai perangkat lunak (*software*) atau program dapat dimanfaatkan. *Software* tersebut ada yang berbayar dan ada yang gratis. Untuk yang berbayar, dua *software* yang dapat digunakan yaitu SPSS dan LISREL, dimana dengan menggunakan kedua *software* kita dapat melakukan analisis faktor, analisis jalur, dan model persamaan struktural. Untuk yang tidak berbayar, berbagai analisis tersebut dapat dilakukan dengan menggunakan R melalui beberapa *Integrated Development Environment* (IDE) seperti RStudio dengan memanfaatkan paket-paket analisis yang bisa diunduh. Pada bagian akhir Bab 2, kami memberikan pendahuluan mengenai program R dan penggunaannya.

Bab 2

Vektor Rerata, Matriks Korelasi dan Kovariansi, dan Program R

Dahulu kita pernah belajar sumbu koordinat dan istilah bebas linear. Apakah itu bebas linear (*linear independence*)? Sebelum membahas lebih jauh mengenai penerapan dari bebas linear, terlebih dahulu dibahas makna dari bebas linear melalui penyajian Gambar 2.1.



Gambar 2.1 Variasi hubungan antara dua vektor

Misal variabel x dan y berturut-turut disajikan pada sumbu horizontal dan vertikal yang berturut-turut merepresentasikan sumbu X dan Y. Sumbu X dan Y tidak dapat disatukan atau dijumlahkan, yang kemudian dikatakan bahwa sumbu X dan Y saling tegak lurus atau juga dapat disebut dengan istilah ortogonal. Ortogonal berarti bahwa arah dua vektor itu saling bebas. Ada garis yang berarah, yang kemu-

dian disebut sebagai vektor, misalnya vektor i dan vektor j (Gambar 2.1).

Pada Gambar 2.1, bagian A menunjukkan bahwa sumbu X dan Y saling tegak lurus, yang berarti bahwa keduanya saling bebas. Vektor i dan j yang ditunjukkan pada bagian A juga saling tegak lurus, yang berarti bahwa kedua vektor tersebut saling bebas. Bagian B menunjukkan bahwa vektor i dan j tidak saling tegak lurus, yang berarti bahwa kedua vektor tersebut tidak saling bebas. Sementara itu, bagian C menunjukkan bahwa vektor i dan j saling bebas karena kedua vektor tersebut saling tegak lurus, sedangkan i dan k tidak saling bebas karena keduanya tidak saling tegak lurus. Bagian D menunjukkan dua vektor i dan j yang saling berhimpit dengan panjang vektor yang berbeda, yang mana ini menunjukkan bahwa vektor i merupakan perkalian vektor j (dengan suatu skalar). Hubungan dari dua vektor yang menunjukkan bahwa kedua vektor tidak saling bebas juga ditunjukkan pada bagian E, dimana vektor i dan j saling berkebalikan. Bagian F menunjukkan bahwa vektor i dan j merupakan perkalian antara dua vektor, sedangkan i dan k serta j dan k tidak saling bebas karena dua pasang vektor masing-masing menunjukkan arah yang berkebalikan atau berlawanan.

Sekelompok vektor itu bebas linear ketika tidak saling tegak lurus. Pada konteks hubungan antara variabel-variabel, bebas linear artinya variabel-variabel tersebut tidak saling mempengaruhi. Berikut akan dijelaskan matriks dan vektor pada analisis data multivariat.

Matriks dan vektor pada analisis data multivariat

Analisis data multivariat merupakan cabang penting dalam statistik dan ilmu data yang memungkinkan kita untuk memahami dan menganalisis hubungan kompleks antara berbagai variabel dalam suatu dataset. Dalam analisis data multivariat, matriks dan vektor merupakan konsep kunci yang mendukung pemodelan, pemahaman, dan interpretasi data yang lebih dalam.

Matriks, dalam konteks analisis data multivariat, digunakan untuk merepresentasikan data. Matriks data merupakan struktur yang mengatur informasi observasi dan variabel dalam kerangka kerja yang terstruktur, memungkinkan kita untuk menggambarkan variasi dan

hubungan antara variabel dengan cara yang sistematis. Sementara itu, vektor rerata adalah alat yang sangat berguna untuk merangkum rata-rata dari masing-masing variabel dalam dataset, memberikan pandangan yang jelas tentang bagaimana variabel-variabel tersebut berperilaku secara keseluruhan.

Matriks variansi-kovariansi adalah alat yang memungkinkan kita untuk mengukur variabilitas dan hubungan antara variabel dalam dataset multivariat. Ini membantu dalam memahami apakah variabel-variabel tersebut saling berkaitan atau tidak, dan sejauh mana hubungan antara mereka memengaruhi hasil analisis. Pada bagian ini akan dijelaskan konsep matriks dan vektor dalam analisis data multivariat. Kami juga akan menunjukkan bagaimana matriks dan vektor ini digunakan atau diterapkan, termasuk perhitungan statistik di dalamnya yang penting untuk dipahami, seperti korelasi dan kovariansi.

Matriks dan vektor dalam analisis data multivariat adalah dasar yang untuk memahami, menganalisis, dan mengambil keputusan yang berdasarkan data yang kompleks dan beragam. Dengan memahami konsep-konsep ini, kita akan memiliki alat yang kuat untuk mengeksplorasi dan memahami hubungan dalam dataset multivariat, yang kemudian dapat membantu dalam membuat keputusan yang lebih baik dan pemecahan masalah yang lebih efisien.

Matriks

Pengertian matriks dan contoh matriks

Dalam matematika, matriks sering digunakan untuk menyederhanakan penulisan dan perhitungan. Suatu matriks dapat mewakili suatu himpunan bilangan yang merupakan entri dari matriks tersebut yang disusun dalam baris dan kolom. Sebuah matriks adalah susunan segi empat siku-siku dari bilangan-bilangan. Bilangan dalam susunan tersebut dinamakan entri dalam matriks (Anton, 1997). Entri dalam matriks berupa bilangan atau elemen disusun secara mendatar (disebut baris) dan disusun secara tegak (disebut kolom). Banyaknya kolom dan baris menunjukkan ukuran dari suatu matriks. $A_{m\times n} = (a_{ij})$, $A_{m\times n}$ merupakan matriks berukuran $m \times n$ dan menunjukkan entri matriks dengan i menunjukkan nomor baris dan j menun-

jukkan nomor kolom. Representasi dari $A_{m\times n} = (a_{ij})$ yaitu sebagai berikut.

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix}$$

dimana a_{mn} menyatakan elemen atau entri dari matriks A pada baris m dan kolom n.

Dua matriks dikatakan sama jika ukurannya sama dan elemen atau entri yang bersesuaian bernilai sama. Dengan demikian, dua matriks $A_{m\times n}$ dan $B_{p\times q}$ sama, ditulis $A = B$, ketika $m = p$ dan $n = q$ dan $(a_{mn}) = (b_{pq})$ atau bahwa entri-entri dari kedua matriks tersebut yang bersesuaian nilainya sama. Bagian selanjutnya secara sekilas membahas beberapa jenis matriks.

Jenis-jenis matriks

*Matriks identitas***.** Matriks identitas ialah matriks persegi yang jumlah barisnya sama dengan jumlah kolom) dan semua elemen diagonal utamanya (dari kiri atas ke kanan bawah) sama dengan 1, sedangkan elemen-elemen di luar diagonal utamanya sama dengan 0. Menurut Suryanto (1998), matriks identitas merupakan matriks diagonal yang setiap entri atau elemen pada diagonal utamanya adalah 1. Suatu matriks identitas I berordo $n \times n$ dapat ditulis sebagai:

$$I_{n\times n} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix}$$

Berikut contoh matriks identitas berukuran 4×4.

$$I_{4\times 4} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Matriks nol. Matriks nol adalah matriks yang semua elemennya sama dengan 0. Matriks nol ini dapat dituliskan dengan 0 dengan disertai ukuran dari matriks nol tersebut. Contoh matriks nol berukuran 2×3 yaitu sebagai berikut.

$$0_{23} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

Matriks simetri (setangkap). Matriks simetri merupakan suatu matriks persegi, yaitu matriks dengan ukuran $m \times m$, yang berlaku bahwa untuk setiap i, j berlaku $a_{ij} = a_{ji}$, dimana a_{ij} menyatakan elemen atau entri dari matriks tersebut pada baris i dan kolom j. Matriks A berikut merupakan contoh suatu matriks persegi, dimana elemen-elemen matriks A saling simetris terhadap diagonal utamanya.

$$A_{3\times3} = \begin{bmatrix} 1 & 3 & 2 \\ 3 & 5 & -1 \\ 2 & -1 & -9 \end{bmatrix}$$

Matriks segitiga atas dan bawah. Matriks segitiga atas yaitu matriks persegi yang semua elemen di bawah diagonal utamanya bernilai nol. Sementara itu, matriks segitiga bawah yaitu matriks persegi yang semua elemen di atas diagonal utamanya bernilai nol. Matriks A dan B berikut secara berturut-turut merupakan contoh dari matriks segitiga atas dan matriks segitiga bawah.

$$A = \begin{bmatrix} 1 & -2 & 8 \\ 0 & 0 & 6 \\ 0 & 0 & -5 \end{bmatrix}, B = \begin{bmatrix} 2 & 0 & 0 \\ -3 & 1 & 0 \\ 9 & 6 & 0 \end{bmatrix}$$

Matriks diagonal. Matriks diagonal adalah suatu matriks persegi dengan semua entrinya yang tidak terletak pada diagonal utama adalah nol (Anton, 1997). Suatu matriks diagonal $D_{n\times n}$ dapat ditulis sebagai berikut.

$$D_{n\times n} = \begin{bmatrix} a_{11} & 0 & \dots & 0 \\ 0 & a_{22} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & a_{nn} \end{bmatrix}$$

Berikut matriks D yang merupakan contoh dari matriks diagonal.

$$D = \begin{bmatrix} -1 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

Matriks skalar. Matriks skalar adalah matriks diagonal yang semua elemen pada diagonal utamanya bernilai sama tetapi tidak nol ($c \neq 0$), dengan c adalah suatu skalar. Matriks A berikut merupakan contoh dari matriks skalar, dimana skalar yang dimaksud yaitu 4.

$$A = \begin{bmatrix} 4 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 4 \end{bmatrix} = 4 \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Transpose matriks. Jika A adalah sebarang matriks yang berukuran $m \times n$, transpose dari matriks A, yang dinyatakan dengan A^T, didefinisikan sebagai suatu matriks yang berukuran $n \times m$ yang elemen-elemen pada kolom pertamanya adalah elemen-elemen pada baris pertama dari matriks A, elemen-elemen pada kolom keduanya adalah elemen-elemen pada baris kedua dari matriks A, demikian juga elemen-elemen pada kolom ketiganya adalah elemen-elemen pada baris ketiga dari matriks A, dan seterusnya (Anton, 1997). Dari pengertian ini, apabila diberikan suatu matriks A sebagai berikut:

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \ldots & a_{mn} \end{bmatrix}$$

maka transpose dari matriks A tersebut yaitu sebagai berikut.

$$A^T = \begin{bmatrix} a_{11} & a_{21} & \cdots & a_{m1} \\ a_{12} & a_{22} & \cdots & a_{m2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1n} & a_{2n} & \ldots & a_{nm} \end{bmatrix}$$

Terdapat beberapa sifat dari transpose matriks yang menyatakan hubungan antara suatu matriks dengan transpose matriks tersebut, yaitu $(A^T)^T = A$, $(A \pm B)^T = A^T \pm B^T$, dan $(ABC)^T = A^T B^T C^T$.

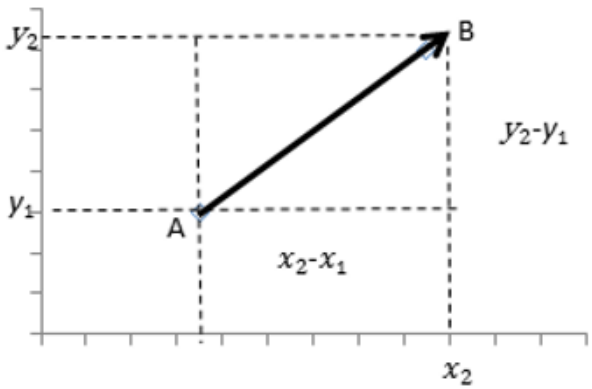
Vektor

Pengertian vektor dan contoh vektor

Suatu matriks yang terdiri dari satu baris atau satu kolom disebut vektor. Matriks yang terdiri dari satu baris disebut vektor baris, sedangkan matriks yang terdiri dari satu kolom disebut vektor kolom. Vektor merupakan sebuah besaran yang memiliki arah. Vektor digambarkan sebagai panah dengan yang menunjukkan arah vektor dan panjang garisnya disebut besar vektor. Dalam penulisannya, jika vektor berawal dari titik A dan berakhir di titik B bisa ditulis dengan sebuah huruf kecil yang di atasnya ada tanda garis atau panah seperti $\vec{v}$ atau $\bar{v}$ atau juga $\overrightarrow{AB}$.

Misalkan vektor $\bar{v}$ merupakan suatu vektor yang berawal dari titik $A(x_1, y_1)$ menuju titik $B(x_2, y_2)$. Vektor $\bar{v}$ ini dapat disajikan dalam koordinat Cartesius seperti yang ditunjukkan pada Gambar 2.2. Pada Gambar 2.2, panjang garis sejajar sumbu X sama dengan $v_1 = x_2 - x_1$ dan panjang garis sejajar sumbu Y sama dengan $v_2 = y_2 - y_1$, dimana keduanya merupakan komponen-komponen vektor $\bar{v}$. Komponen vektor $\bar{v}$ dapat ditulis untuk menyatakan vektor secara aljabar yaitu sebagai berikut.

$$\vec{v} = \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} x_2 - x_1 \\ y_2 - y_1 \end{pmatrix} \text{ atau } \vec{v} = (v_1 . v_2)$$



Gambar 2.2 Representasi dari vektor $\bar{v}$ atau $\overrightarrow{AB}$

Jenis-jenis vektor

Vektor posisi. Suatu vektor yang posisi titik awalnya di *O*(0,0) pada koordinat Cartesius dan titik ujungnya di $A(a_1, a_2)$ disebut sebagai suatu vektor posisi.

Vektor nol. Vektor nol merupakan suatu vektor yang panjangnya nol dan dinotasikan $\bar{0}$. Vektor nol tidak memiliki arah vektor yang jelas.

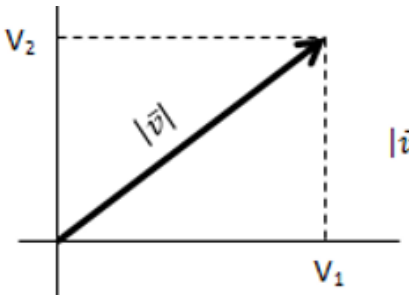
Vektor satuan. Vektor satuan merupakan suatu vektor yang panjangnya satu satuan. Vektor satuan dari $\vec{v} = \begin{pmatrix} v_1 \\ v_2 \end{pmatrix}$ yaitu sebagai berikut.

$$\bar{U}_v = \frac{\bar{v}}{|\bar{v}|} = \frac{1}{|\bar{v}|} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix}$$

Vektor basis. Vektor basis merupakan vektor satuan yang saling tegak lurus. Dalam vektor ruang dua dimensi (R^2) memiliki dua vek-

tor basis yaitu $\bar{\imath} = (1,0)$ dan $\bar{\jmath} = (0,1)$. Dalam tiga dimensi (R^3) terdapat tiga vektor basis yaitu yaitu $\bar{\imath} = (1,0,0)$, $\bar{\jmath} = (0,1,0)$ dan $\bar{k} = (0,0,1)$.

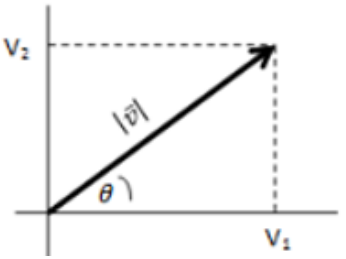
Vektor dua dimensi (R^2). Panjang segmen garis yang menyatakan vektor $\bar{v}$ (dinotasikan sebagai $|\bar{v}|$) merupakan panjang suatu vektor dalam dua dimensi (lihat Gambar 2.3).



$$|\bar{v}| = \sqrt{(v_1)^2 + (v_2)^2}$$

Gambar 2.3 Representasi dari panjang vektor $\bar{v}$

Panjang vektor tersebut dapat dikaitkan dengan sudut θ yang dibentuk oleh vektor dan sumbu X positif.



$$\cos\theta = \frac{v_1}{|\bar{v}|} \rightarrow v_1 = |\bar{v}|\cos\theta$$

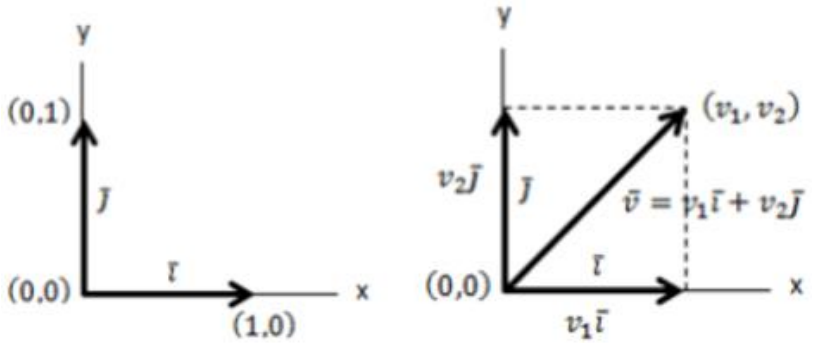
$$\sin\theta = \frac{v_2}{|\bar{v}|} \rightarrow v_2 = |\bar{v}|\sin\theta$$

$$\tan\theta = \frac{v_2}{v_1} \rightarrow \theta = \tan^{-1}\left(\frac{v_2}{v_1}\right)$$

Gambar 2.4 Hubungan ukuran sudut θ dan panjang vektor

Vektor dapat disajikan sebagai kombinasi linear dari vektor basis yaitu $\bar{\imath} = \begin{pmatrix}1\\0\end{pmatrix}$ dan $\bar{\jmath} = \begin{pmatrix}0\\1\end{pmatrix}$. Kombinasi linear untuk suatu vektor pada dua dimensi dapat dinyatakan dalam persamaan berikut dan representasinya disajikan pada Gambar 2.5.

$$\vec{v} = \begin{pmatrix}v_1\\v_2\end{pmatrix} = v_1\begin{pmatrix}1\\0\end{pmatrix} + v_2\begin{pmatrix}0\\1\end{pmatrix} = v_1\bar{\imath} + v_2\bar{\jmath}$$

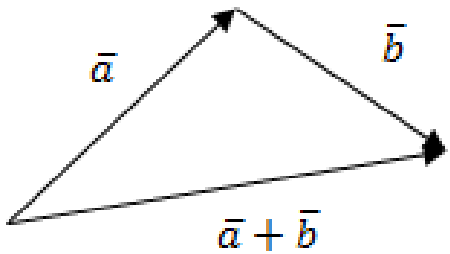


Gambar 2.5 Representasi kombinasi linear dari suatu vektor

Penjumlahan dan pengurangan vektor pada dimensi dua
Dua vektor atau lebih dapat dijumlahkan dan hasilnya disebut resultan. Penjumlahan vektor secara aljabar dapat dilakukan dengan menjumlahkan komponen yang seletak. Jika $\vec{a} = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix}$ dan $\vec{b} = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}$, maka jumlahan dari dua vektor tersebut yaitu sebagai berikut.

$$\vec{a} + \vec{b} = \begin{pmatrix} a_1 + b_1 \\ a_2 + b_2 \end{pmatrix}$$

Penjumlahan secara grafis dapat dilihat pada Gambar 2.6.



Gambar 2.6 Representasi secara grafis penjumlahan dua vektor

Pengurangan pada vektor juga serupa dengan penjumlahan yaitu sebagai berikut.

$$\vec{a} - \vec{b} = \begin{pmatrix} a_1 - b_1 \\ a_2 - b_2 \end{pmatrix}$$

Terdapat sifat-sifat dalam penjumlahan vektor yang di antaranya sebagai berikut: $\vec{a} + \vec{b} = \vec{b} + \vec{a}$ dan $\vec{a} + (\vec{b} + \vec{c}) = (\vec{a} + \vec{b}) + \vec{c}$.

Perkalian vektor pada dimensi dua dengan skalar

Suatu vektor dapat dikalikan dengan suatu skalar (bilangan real) dan akan menghasilkan suatu vektor baru. Jika $\bar{v}$ adalah suatu vektor dan k adalah skalar, maka perkalian vektor dengan skalar tersebut yaitu $k.\bar{v}$. Ada ketentuan-ketentuan mengenai arah dari vektor hasil kali antara suatu vektor dengan suatu skalar tersebut sebagai berikut.

- Jika $k > 0$, maka vektor $k.\bar{v}$ searah dengan vektor $\bar{v}$
- Jika $k < 0$, maka vektor $k.\bar{v}$ berlawanan arah dengan vektor $\bar{v}$
- Jika $k = 0$, maka vektor $k.\bar{v}$ adalah vektor identitas $\bar{0}$.

Secara grafis, perkalian ini dapat mengubah panjang dari suatu vektor (lihat Tabel 2.1). Secara aljabar perkalian suatu vektor $\bar{v}$ dengan skalar k dapat dirumuskan sebagai $k.\bar{v} = \begin{pmatrix} k.v_1 \\ k.v_2 \end{pmatrix}$.

Tabel 2.1 Representasi perkalian suatu vektor dengan suatu skalar

$\bar{v}$	$k = 0{,}5$	$k = -0{,}5$	$k = 2$	$k = -2$
$\bar{v}$	$0{,}5.\bar{v}$	$-0{,}5.\bar{v}$	$2.\bar{v}$	$-2.\bar{v}$

Perkalian skalar dua vektor di dimensi dua

Perkalian skalar dua vektor disebut juga sebagai hasil kali titik dua vektor dan ditulis sebagai:$\bar{a}.\bar{b}$ (dibaca a dot b). Perkalian skalar vektor $\bar{a}$ dan $\bar{b}$ dilakukan dengan mengalikan panjang vektor $\bar{a}$ dan panjang vektor $\bar{b}$ dengan cosinus θ. Sudut θ merupakan sudut antara vektor $\bar{a}$ dan vektor $\bar{b}$. Definisi dari perkalian skalar dua vektor tersebut yaitu $\bar{a}.\bar{b} = |\bar{a}||\bar{b}|cos\theta$. Dari definisi ini, diperoleh hasil bahwa: $\bar{a}.\bar{a} = |\bar{a}|^2$ dan $\bar{a}.(\bar{b} \pm \bar{c}) = (\bar{a}.\bar{b}) \pm (\bar{a}.\bar{c})$

Vektor di dimensi tiga

Jarak antara dua vektor di ruang dimensi tiga dapat ditentukan melalui pengembangan rumus Pythagoras. Jika titik $A(a_1, a_2, a_3)$ dan ti-

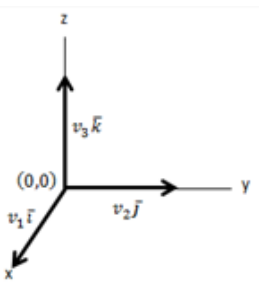
tik $B(b_1, b_2, b_3)$ maka jarak antara kedua titik tersebut, yaitu AB, adalah:

$$AB = \sqrt{(b_1 - a_1)^2 + (b_2 - a_2)^2 + (b_3 - a_3)^2}$$

Dengan menggunakan konsep serupa, jika $\vec{v} = \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix}$), maka $|\bar{v}| = \sqrt{(v_1)^2 + (v_2)^2 + (v_3)^2}$.

Vektor $\overrightarrow{AB}$ dapat dinyatakan dalam dua bentuk, yaitu dalam kolom $\overrightarrow{AB} = \begin{pmatrix} b_1 - a_1 \\ b_2 - a_2 \\ b_3 - a_3 \end{pmatrix}$ atau dalam baris, yaitu $\overrightarrow{AB} = b_1 - a_1\,, b_2 - a_2\,, b_3 - a_3$.Vektor tersebut juga dapat disajikan sebagai kombinasi linier dari vektor basis $\bar{\imath}(1,0,0)$ dan $\bar{\jmath}(0,1,0)$ dan $\bar{k}(0,0,1)$ sebagai berikut, dimana representasinya disajikan pada Gambar 2.7.

$$\bar{v} = v_1 \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} + v_2 \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} + v_3 \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} = v_1 \bar{\imath} + v_2 \bar{\jmath} + v_3\, \bar{k}$$



Gambar 2.7 Representasi kombinasi linear vektor di dimensi tiga

Operasi vektor di ruang dimensi tiga

Operasi vektor di ruang dimensi tiga secara umum memiliki konsep yang sama dengan operasi vektor di dimensi dua dalam penjumlahan, pengurangan, maupun perkalian.

- Penjumlahan dan pengurangan vektor di ruang dimensi tiga

Penjumlahan dan pengurangan vektor-vektor di ruang dimensi tiga sama dengan vektor dimensi dua yaitu sebagai berikut.

$$\vec{a} + \vec{b} = \begin{pmatrix} a_1 + b_1 \\ a_2 + b_2 \end{pmatrix}$$

$$\bar{a} + \bar{b} = \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix} + \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix} = \begin{pmatrix} a_1 + b_1 \\ a_2 + b_2 \\ a_3 + b_3 \end{pmatrix}$$

dan

$$\bar{a} - \bar{b} = \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix} - \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix} = \begin{pmatrix} a_1 - b_1 \\ a_2 - b_2 \\ a_3 - b_3 \end{pmatrix}$$

- Perkalian suatu vektor dengan suatu skalar pada ruang dimensi tiga

Jika $\bar{v}$ adalah suatu vektor dan k adalah suatu skalar, maka perkalian antara vektor dan skalar tersebut yaitu sebagai berikut.

$$k.\bar{v} = \begin{pmatrix} k.v_1 \\ k.v_2 \\ k.v_3 \end{pmatrix}$$

- Hasil kali skalar dua vektor pada ruang dimensi tiga

Selain rumus di ruang dimensi tiga, ada rumus lain dalam hasil kali skalar dua vektor di ruang dimensi tiga. Jika $\bar{a} = a_1\bar{\imath} + a_2\bar{\jmath} + a_3\bar{k}$ dan $\bar{b} = b_1\bar{\imath} + b_2\bar{\jmath} + b_3\bar{k}$, maka hasil dari $\bar{a}.\bar{b}$ sama dengan $\bar{a}.\bar{b} = (a_1b_1) + (a_2b_2) + (a_3b_3)$.

Perbedaan kesamaan matriks dan vektor

Dua buah matriks A dan B dikatakan sama, dituliskan $A = B$, apabila keduanya berorde atau berukuran sama dan semua unsur yang terkandung di dalamnya sama. Jika matriks A itu tidak sama dengan matriks B, maka ditulis $A \neq B$. Sebagai contoh, misal diberikan tiga matriks berikut.

$$A = \begin{pmatrix} 1 & 7 & 6 \\ 2 & 6 & 4 \end{pmatrix}, B = \begin{pmatrix} 1 & 7 & 6 \\ 2 & 6 & 4 \end{pmatrix}, C = \begin{pmatrix} 1 & 7 & 6 \\ 2 & -6 & 4 \end{pmatrix}$$

Dari ketiga matriks tersebut, diperoleh informasi bahwa $A = B, B \neq C$, dan $A \neq C$.

Selanjutnya, dua vektor dikatakan sama apabila keduanya sejenis, memiliki dimensi yang sama, dan semua unsur yang terkandung didalamnya sama. Sebagai contoh, misal diberikan empat vektor sebagai berikut.

$$\bar{a} = [2 \quad 4 \quad 7], \bar{b} = [2 \quad 4 \quad 7], \bar{u} = \begin{bmatrix} 2 \\ 7 \\ 4 \end{bmatrix}, \bar{v} = \begin{bmatrix} 2 \\ 2 \\ 4 \end{bmatrix}$$

Dari keempat vektor tersebut, diperoleh hubungan-hubungan bahwa $\bar{a} = \bar{b}, \bar{u} \neq \bar{v}, \bar{a} \neq \bar{u} \neq \bar{v}$, dan $\bar{b} \neq \bar{u} \neq \bar{v}$.

Rata-rata (Rerata)

Ukuran pemusatan yang paling banyak digunakan adalah rata-rata. Misalkan $x_{11}, x_{21}, \ldots, x_{n1}$ adalah n pengukuran pada variabel pertama yang tidak semuanya berbeda. Rata-rata pengukuran disebut juga rata-rata (*mean*) sampel yang ditulis dengan $\bar{x}$. Secara umum, rata-rata sampel untuk variabel ke-j dengan p variabel dan n objek diberikan sebagai berikut.

$$\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$$

dengan $j = 1, 2, \ldots, p$. Misalkan matriks acak $X = [X_i]$ berorde $p \times 1$ untuk setiap $i = 1, 2, \ldots, p$ merupakan suatu vektor acak X untuk populasi yang didefinisikan sebagai berikut.

$$E(X) = E\begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix} = \begin{bmatrix} E(X_1) \\ E(X_2) \\ \vdots \\ E(X_p) \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix} = \mu$$

Dalam analisis multivariat sering kali dihadapkan pada masalah pengamatan yang dilakukan pada suatu periode waktu untuk $p > 1$ variabel atau karakter. Akan digunakan notasi yang mendefinisikan objek ke-i pada variabel ke-j. Menurut Johnson dan Wichern (2007), secara umum sampel data multivariat dapat disajikan dalam bentuk sebagaimana yang ditunjukkan dalam Tabel 2.2.

Tabel 2.2 Penyajian sampel data multivariat

	Var-1	Var-2	...	Var-j	...	Var-p
Objek-1	x_{11}	x_{12}	...	x_{1j}	...	x_{1p}
Objek-2	x_{21}	x_{22}	...	x_{2j}	...	x_{2p}
...	...	...	...	...	...	...
Objek-i	x_{i1}	x_{i2}	...	x_{ij}	...	x_{jp}
...	...	...	...	...	...	...
Objek-n	x_{n1}	x_{n2}	...	x_{nj}	...	x_{np}

Rerata multivariat adalah rerata dari beberapa variabel atau atribut yang terkait dalam satu set data. Data multivariat biasanya terdiri dari beberapa variabel yang diukur atau diamati bersamaan. Rerata multivariat dapat memberikan pemahaman tentang rerata atau rata-rata dari masing-masing variabel tersebut secara bersamaan. Pada matriks data multivariat, masing-masing variabel bisa dihitung rata-ratanya yang dapat disajikan dalam bentuk vektor rata-rata sebagai berikut.

$$X = (X_1 \quad X_2 \quad \cdots \quad X_p)$$
$$\mu = (\mu_1 \quad \mu_2 \quad \cdots \quad \mu_p)$$
$$\boldsymbol{X} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix}$$
$$\mu_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$$
$$\mu = \frac{1}{n}1'X = \frac{1}{n}X'$$

Vektor rerata (*mean vector*) dalam konteks multivariat ialah vektor yang berisi rerata atau rata-rata dari sejumlah variabel yang diukur dalam *dataset* multivariat. Vektor rerata ini biasanya digunakan untuk menggambarkan pusat dari distribusi multivariat data. Vektor rerata menggabungkan rerata dari masing-masing variabel ke dalam bentuk vektor.

Tabel 2.3 Penggunaan simbol pada univariat dan multivariat

Perbedaan	Univariat	Multivariat
Rerata	$\bar{x}, \mu$	$\bar{\boldsymbol{x}}, \boldsymbol{\mu}$
Varians	s^2, σ^2	$\boldsymbol{\Sigma^2}$
Simpangan baku (deviasi standar)	s, σ	$\boldsymbol{\Sigma}$

Variansi

Variansi sampel (ditulis s^2) merupakan estimator dari variansi populasi σ^2. Variansi sampel variabel pertama dengan n pengamatan didefinisikan sebagai berikut.

$$s_1^2 = \frac{1}{n}\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)^2$$

Sementara itu, secara umum, variansi sampel untuk variabel ke-j diberikan sebagai berikut.

$$s_j{}^2 = s_{jj} = var(X_j) = \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2;\ j = 1, 2, \ldots, p$$

Dengan mengambil x_d sebesar vektor kolom dari matriks X_d didapat persamaan berikut.

$$s^2{}_{p\times 1} = \begin{bmatrix} s_1{}^2 \\ s_2{}^2 \\ \vdots \\ s_n{}^2 \end{bmatrix} = diag\left[\frac{1}{n-1}\right] X_{d'}{}_{p\times n} X_d{}_{n\times p}$$

Variansi populasi dinyatakan dalam σ^2 dan simpangan baku populasi adalah σ. Rumus berikut dapat digunakan untuk menentukan variansi populasi.

$$\sigma_j{}^2 = \frac{1}{N}\sum_{i=1}^{N}(x_{ij} - \mu_j)^2$$

dengan $\sigma_j{}^2$ menyatakan variansi untuk variabel X_j, x_{ij} menyatakan nilai ke-i dari variabel X_j, μ_j menyatakan rataan populasi dari variabel X_j, dan N menyatakan ukuran populasi.

Kovariansi

Kovariansi merupakan ukuran keterikatan antara dua variabel, misal X_1 dan X_2. Menurut Johnson dan Wichern (2002), kovariansi sampel untuk variabel ke-1 dan variabel ke-2 diberikan sebagai berikut.

$$s_{12} = \frac{1}{n}\sum_{i=1}^{n}(x_{11} - \bar{x}_1)(x_{12} - \bar{x}_2)$$

Secara umum, kovariansi sampel untuk variabel ke-j dan ke-k diberikan sebagai berikut.

$$s_{jk} = \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)$$

$$s_{jk} = \frac{1}{n}\left(\sum_{i=1}^{n} x_{ij}x_{ik} - n\bar{x}_j\bar{x}_k\right)$$

dengan $j, k = 1, 2, \ldots, p$, s_{jk} menyatakan kovariansi antara dua variabel yaitu variabel X_j dan X_k, x_{ij} menyatakan nilai ke-i dari variabel X_j, x_{ik} menyatakan nilai ke-i dari variabel X_k, $\bar{x}_j$ menyatakan rataan nilai variabel X_j, $\bar{x}_k$ menyatakan rataan nilai variabel X_k, n menyatakan ukuran sampel. Sehubungan dengan kovariansi, variansi sampel dapat pula diartikan sebagai kovariansi variabel ke-k dan variabel ke-j. Suatu matriks yang entrientrinya terdiri atas variansi dan kovariansi dari sekumpulan variabel disebut dengan matriks variansi-kovariansi dinotasikan dengan S yang dapat dinyatakan dalam bentuk berikut.

$$\boldsymbol{s_{p\times p}} = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ s_{p1} & s_{p2} & \ldots & s_{pp} \end{bmatrix}$$

Rumus berikut ini dapat digunakan untuk menentukan kovariansi populasi.

$$\sigma_{jk} = \frac{1}{N}\sum_{r=1}^{N}\sum_{i=1}^{N}(x_{ij} - \mu_j)(x_{rk} - \mu_k)$$

dengan σ_{jk} menyatakan kovariansi antara dua variabel yaitu X_j dan X_k, x_{ij} menyatakan nilai ke-i dari variabel X_j, x_{rk} menyatakan nilai ke-r dari variabel X_k, μ_j menyatakan rataan nilai dari variabel X_j, μ_k menyatakan rataan nilai variabel X_k, dan N menyatakan ukuran populasi. Entri dari diagonal matriks variansi-kovariansi yaitu nilai varians, sedangkan entri matriks yang bukan diagonal yaitu nilai kovariansi, dimana matriks variansi-kovariansi tersebut dapat dinyatakan sebagai berikut.

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \ldots & \sigma_{pp} \end{bmatrix}$$

Oleh karena berlaku $\sigma_{jk} = \sigma_{kj}$, berarti untuk setiap $j = 1, 2, \ldots, p$ dan $k = 1, 2, \ldots, p$ dengan $j = k$, maka akan berlaku persamaan berikut ini.

$$\boldsymbol{\Sigma} = \begin{bmatrix} {\sigma_1}^2 & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & {\sigma_2}^2 & \cdots & \sigma_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \ldots & {\sigma_p}^2 \end{bmatrix}$$

Korelasi biasa dan matriks korelasi

Koefisien korelasi digunakan untuk mengukur hubungan antara dua variabel dalam analisis korelasi dan dinotasikan dengan r. Koefisien korelasi sampel antara variabel X dan Y dinotasikan dengan r_{xy} adalah sebagai berikut (Johnson & Wichern, 2007).

$$r_{xy} = \frac{s_{xy}}{\sqrt{s_{xx}}\sqrt{s_{yy}}} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

dengan s_{xy} adalah kovariansi dari x dan y, sedangkan s_{xx} dan s_{yy} adalah simpangan baku. Nilai koefisien korelasi untuk mengukur hubungan antara dua variabel berkisar antara -1 sampai 1. Jika koefisien bertanda (+), maka kedua variabel mempunyai hubungan searah. Akan tetapi, jika koefisien bertanda (-), maka kedua variabel mempunyai hubungan tidak searah.

Koefisien korelasi untuk populasi disimbolkan ρ didefinisikan sebagai rasio kovariansi σ_{ik} dengan variansi σ_{ii} dan σ_{kk}. Dengan demikian, diperoleh korelasi populasi sebagai berikut.

$$\rho_{ik} = \frac{\sigma_{ii}}{\sigma_{kk}}$$

Koefisien korelasi populasi untuk matriks $p \times p$. Matriks korelasi populasi dinotasikan (ρ) terdiri dari koefisien korelasi dan dapat dituliskan sebagai berikut dengan $\rho_{12} = \rho_{21}$, $\rho_{13} = \rho_{31}$, ..., $\rho_{1p} = \rho_{p1}$.

$$\rho = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{21} & 1 & \cdots & \rho_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ \rho_{p1} & \rho_{p2} & \ldots & 1 \end{bmatrix}$$

Sementara itu, untuk matriks korelasi sampel dinotasikan R dan dapat dituliskan sebagai berikut dengan $r_{12} = r_{21}$, $r_{13} = r_{31}$, ..., $r_{1p} = r_{p1}$.

$$R = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{bmatrix}$$

Contoh kasus: Menentukan vektor rerata, matriks korelasi, dan matriks variansi-kovariansi

Dari data berikut, kita diminta untuk menentukan vektor rerata, matriks korelasi, dan matriks variansi-kovariansi.

Siswa	Pra	Sikap	Pengetahuan	Keterampilan
1	64	73	64	82
2	71	87	71	69
3	70	83	70	71
4	79	77	79	65
5	57	54	57	67
6	61	81	61	87
7	67	72	67	88
8	72	65	72	88
9	81	69	81	66
10	67	63	67	70
11	60	80	60	88
12	69	63	69	71
13	81	85	81	82
14	77	78	77	76
15	78	66	78	72

Pertama, kita menentukan vektor rerata dari data yang diberikan sebagai berikut. Misalkan $\bar{X}_1$, $\bar{X}_2$, $\bar{X}_3$, dan $\bar{X}_4$ secara berturut-turut menyatakan rata-rata (rerata) untuk variabel pra, sikap, pengetahuan, dan keterampilan.

$$\bar{X}_1 = \frac{1}{n}\sum_{i=1}^{n} x_{ij} = \frac{1}{15}(64 + 71 + \cdots + 78) = 70{,}27$$

$$\bar{X}_2 = \frac{1}{n}\sum_{i=1}^{n} x_{ij} = \frac{1}{15}(73 + 87 + \cdots + 66) = 73{,}07$$

$$\bar{X}_3 = \frac{1}{n}\sum_{i=1}^{n} x_{ij} = \frac{1}{15}(64 + 71 + \cdots + 78) = 70{,}27$$

$$\bar{X}_4 = \frac{1}{n}\sum_{i=1}^{n} x_{ij} = \frac{1}{15}(82 + 69 + \cdots + 72) = 76{,}13$$

Dari hasil penentuan rerata dari masing-masing variabel tersebut, diperoleh vektor rerata sebagai berikut.

$$\bar{X} = \begin{bmatrix} \bar{X}_1 \\ \bar{X}_2 \\ \bar{X}_3 \\ \bar{X}_4 \end{bmatrix} = \begin{bmatrix} 70{,}27 \\ 73{,}07 \\ 70{,}27 \\ 76{,}13 \end{bmatrix}$$

Selanjutnya, kita menentukan matriks korelasi dari data yang diberikan. Penentuan matriks korelasi ini dilakukan dengan memperhatikan bahwa matriks korelasi diberikan sebagai berikut.

$$R = \begin{bmatrix} 1 & r_{12} & r_{13} & r_{14} \\ r_{21} & 1 & r_{23} & r_{24} \\ r_{31} & r_{32} & 1 & r_{34} \\ r_{41} & r_{42} & r_{43} & 1 \end{bmatrix}$$

dengan menggunakan rumus =CORREL pada Excel diperoleh:

$$r_{12} = r_{21} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}} = 0{,}24$$

$$r_{13} = r_{31} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}} = 1$$

$$r_{14} = r_{41} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}} = -0{,}32$$

$$r_{23} = r_{32} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}} = 0{,}24$$

$$r_{24} = r_{42} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}} = 0{,}25$$

$$r_{34} = r_{43} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}} = 0{,}32$$

Dengan demikian, matriks korelasi yang terbentuk dari data yang diberikan yaitu sebagai berikut.

$$R = \begin{bmatrix} 1 & 0{,}24 & 1 & -0{,}32 \\ 0{,}24 & 1 & 0{,}24 & 0{,}25 \\ 1 & 0{,}24 & 1 & -0{,}32 \\ -0{,}32 & 0{,}25 & -0{,}32 & 1 \end{bmatrix}$$

Setelah kita menentukan vektor rerata dan matriks korelasi dari data yang diberikan, kita selanjutnya menentukan matriks variansi-kovariansi sebagai berikut dengan memperhatikan bahwa bentuk dari matriks tersebut yaitu:

$$\boldsymbol{\Sigma} = \begin{bmatrix} {\sigma_1}^2 & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \sigma_{21} & {\sigma_2}^2 & \sigma_{23} & \sigma_{24} \\ \sigma_{31} & \sigma_{32} & {\sigma_3}^2 & \sigma_{34} \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & {\sigma_4}^2 \end{bmatrix}$$

dengan $\sigma_{jk} = \frac{1}{N}\sum_{r=1}^{N}\sum_{i=1}^{N}(x_{ij} - \mu_j)(x_{rk} - \mu_k)$ dan $cov(x, y) = r_{xy}\sqrt{var(x)}\sqrt{var(y)}$. Dengan memperhatikan hal ini, dapat diperoleh hasil sebagai berikut yang merupakan bagian variansi.

$${\sigma_1}^2 = r_{11}\sqrt{var(x_1)}\sqrt{var(x_1)} = 1(7{,}77)(7{,}77) = 60{,}35$$

$${\sigma_2}^2 = r_{22}\sqrt{var(x_2)}\sqrt{var(x_2)} = 1(9{,}58)(9{,}58) = 91{,}78$$

$${\sigma_3}^2 = r_{33}\sqrt{var(x_3)}\sqrt{var(x_3)} = 1(7{,}77)(7{,}77) = 60{,}35$$

$${\sigma_4}^2 = r_{44}\sqrt{var(x_4)}\sqrt{var(x_4)} = 1(8{,}77)(8{,}77) = 76{,}98$$

Adapun hasil untuk bagian kovariansi yaitu sebagai berikut.

$$\sigma_{12} = \sigma_{21} = r_{12}\sqrt{var(x_1)}\sqrt{var(x_2)} = 0{,}24(7{,}77)(9{,}58) = 17{,}77$$

$\sigma_{13} = \sigma_{31} = r_{13}\sqrt{var(x_1)}\sqrt{var(x_3)} = 1(7{,}77)(7{,}77) = 60{,}35$
$\sigma_{14} = \sigma_{41} = r_{14}\sqrt{var(x_1)}\sqrt{var(x_4)} = -0{,}32(7{,}77)(8{,}77) = -22{,}11$
$\sigma_{23} = \sigma_{32} = r_{23}\sqrt{var(x_2)}\sqrt{var(x_3)} = 0{,}24(9{,}58)(7{,}77) = 17{,}77$
$\sigma_{24} = \sigma_{42} = r_{24}\sqrt{var(x_2)}\sqrt{var(x_4)} = 0{,}25(7{,}77)(8{,}77) = 20{,}92$
$\sigma_{34} = \sigma_{43} = r_{34}\sqrt{var(x_3)}\sqrt{var(x_4)} = -0{,}32(7{,}77)(8{,}77) = -22{,}98$

Berdasarkan hasil pada bagian variansi dan kovariansi yang telah diperoleh, dapat diperoleh bahwa matriks variansi-kovariansi pada data yang diberikan yaitu sebagai berikut.

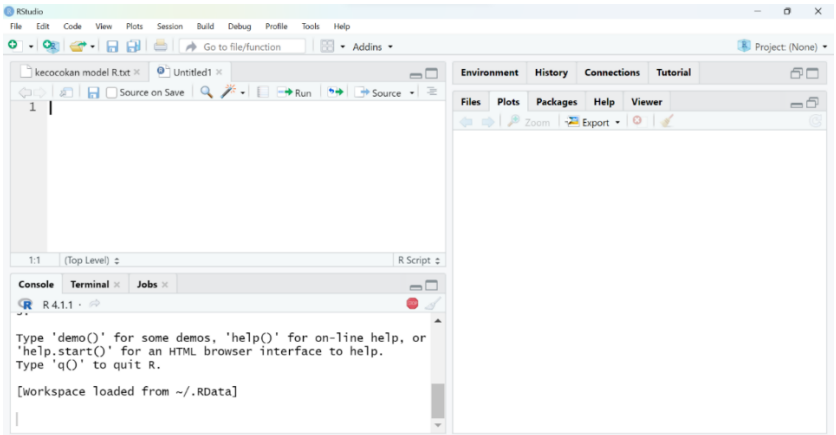
$$\boldsymbol{\Sigma} = \begin{bmatrix} 60{,}35 & 17{,}77 & 60{,}35 & -22{,}11 \\ 17{,}77 & 91{,}78 & 17{,}77 & 20{,}91 \\ 60{,}35 & 17{,}77 & 60{,}35 & -22{,}11 \\ -22{,}11 & 20{,}91 & -22{,}11 & 76{,}98 \end{bmatrix}$$

Program R: Instalasi dan contoh aplikasinya

R merupakan suatu bahasa pemrograman yang digunakan untuk menangani data. Penanganan ini mulai dari menyajikan, melakukan pengodean (*coding*), menganalisis data, melakukan pemodelan, sampai dengan menyajikan hasil analisis dan visualisasinya. Program R ini dapat diakses dan digunakan secara gratis oleh siapa pun melalui beberapa IDE yang tersedia seperti RStudio. Selain tersedia layanan bebas biaya, program R di bawah lingkungan IDE RStudio juga menawarkan layanan berbayar yang memungkinkan pengguna untuk menganalisis data yang berukuran besar dalam waktu yang sangat cepat dan berbagai layanan unggulan lainnya. Namun demikian, layanan yang bersifat gratis pun sudah memiliki banyak fitur yang banyak yang sudah dapat dikatakan cukup untuk melakukan berbagai analisis data multivariat yang akan dibahas dalam buku ini.

Keunggulan lain dari RStudio yaitu adanya independensi pada pengembangan berbagai fitur atau paket (*package*) yang memudahkan pengguna untuk bekerja dengan data atau analisis data. Banyak paket yang dikembangkan secara bebas oleh pengguna dan diverifikasi oleh pengguna lainnya yang kemudian dapat dimanfaatkan secara luas oleh pengguna lain. Paket-paket yang tersedia tersebut kemudian juga selalu diperbaharui oleh pengembang paket tersebut sesuai kebutuhan dan perkembangan keilmuan.

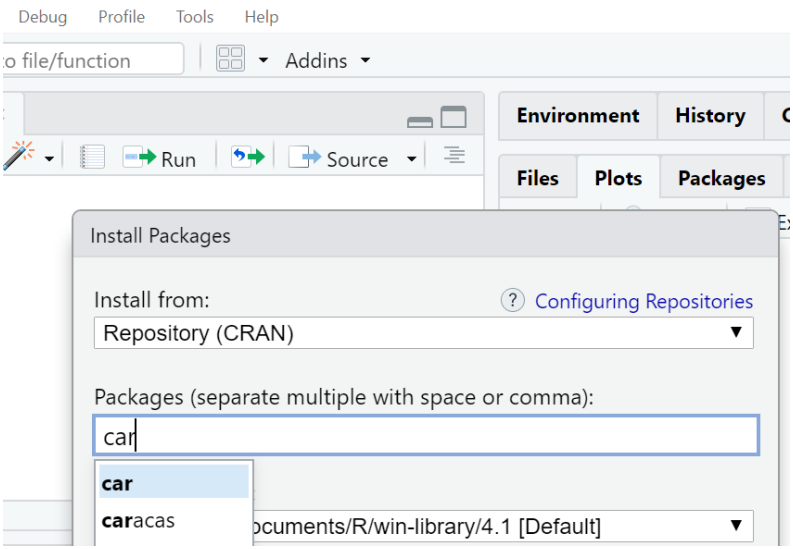
Untuk memasang program R dan RStudio dalam perangkat komputer atau laptop, calon pengguna terlebih dahulu perlu mengunduh file instalasi untuk kedua program tersebut yang secara berturut-turut dapat didapatkan atau diunduh pada laman the R Project for Statistical Computing dan the Comprehensive R Archive Network melalui https://cran.r-project.org/mirrors.html dan https://cloud.r-project.org dan tautan laman Posit https://posit.co/products/open-source/rstudio/. Petunjuk untuk memasang program R dan RStudio juga tersedia pada Hands-On Programming with R yang dapat diakses melalui tautan https://rstudio-education.github.io/hopr/starting.html. Setelah kedua program tersebut berhasil diunduh, calon pengguna selanjutnya dapat melakukan instalasi pada perangkat komputer atau laptop yang mereka miliki. Gambar 2.8 menunjukkan tampilan awal dari RStudio.



Gambar 2.8 Jendela utama RStudio

Pengguna dapat menggunakan RStudio secara langsung tanpa memerlukan suatu paket untuk melakukan operasi-operasi dasar matematika atau statistika. Seperti yang telah disebutkan sebelumnya, bahwa ketika bekerja dengan suatu data dengan ukuran besar dan memerlukan analisis yang lebih kompleks, pengguna dapat memasang atau melakukan instalasi suatu atau beberapa paket (*package*) dan menggunakan paket-paket tersebut. Instalasi paket tersebut dapat dilakukan dengan cara memilih menu Tools kemudian memilih

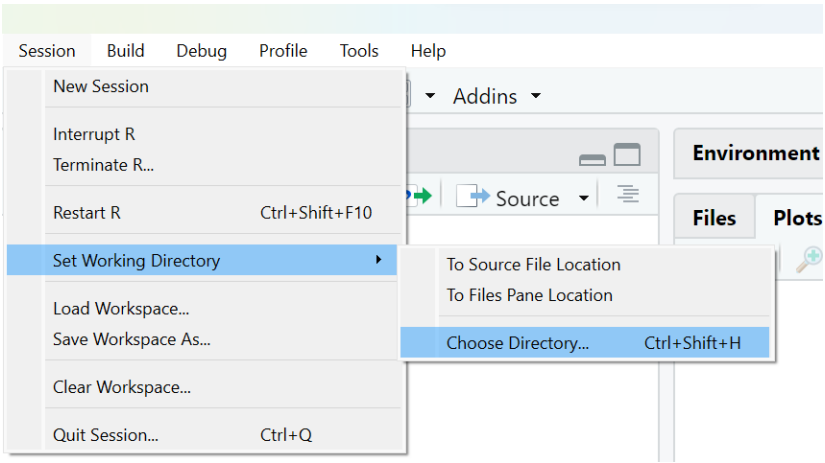
Install Packages. Setelah itu akan muncul jendela Install Packages yang memerlukan kita untuk mengetikkan nama paket yang kita hendak pasang pada kotak teks Packages. Sebagai contoh, Gambar 2.9 menunjukkan bahwa paket yang akan dipasang untuk digunakan yaitu 'car'. Instalasi atau pemasangan paket juga dapat dilakukan dengan cara mengetikkan perintah `install.packages("car")` pada panel Console.



Gambar 2.9 Instalasi paket 'car' pada RStudio

Pada bagian pendahuluan mengenai penggunaan RStudio ini, kami akan menyajikan beberapa fungsi dasar yang digunakan untuk menganalisis data, khususnya pada analisis data multivariat. Sebelum melakukan analisis data lebih lanjut, hal pertama yang perlu dilakukan terlebih dahulu yaitu menentukan di mana sumber data yang akan dianalisis dan tempat menyimpan semua hasil analisis, yang mana itu dikenal dengan direktori (*directory)* atau folder. Kita dapat menentukan direktori ini salah satunya dengan cara memilih menu Session, kemudian memilih Set Working Directory dan memilih Choose Directory (Ctrl + Shift + H) (lihat Gambar 2.10). Setelah itu akan muncul tampilan jendela yang mengarahkan kita untuk menentukan direktori atau folder yang dimaksud.

Selain menentukan direktori, hal yang penting juga diketahui yaitu memanggil atau mengaktifkan paket-paket yang sudah terpasang dan siap digunakan untuk berbagai keperluan. Hal ini dilakukan dengan menggunakan perintah `library(package)`. Sebagai contoh, misalkan kita akan menggunakan paket 'car', untuk mengaktifkan paket tersebut, perintah yang kita gunakan yaitu `library(car)`. Selain itu, perintah-perintah yang diketikkan pada panel R Script dapat dijalankan dengan menekan tombol Run pada RStudio atau dengan menekan tombol Ctrl + Shift + Enter.



Gambar 2.10 Menentukan direktori pada RStudio

Sebagai latihan dalam memanfaatkan program R dan RStudio, kami menyediakan suatu data untuk dianalisis, dimana data tersebut yang tersaji pada suatu tabel di halaman 31. Kolom "siswa" merupakan nomor urut siswa, variabel "pra" merupakan hasil tes awal pembelajaran, variabel "sikap" merupakan hasil penilaian sikap siswa, variabel "pengetahuan" merupakan hasil tes pengetahuan siswa terhadap suatu mata pelajaran, dan variabel "keterampilan" merupakan hasil penilaian keterampilan siswa. Kita dapat membaca data tersebut di RStudio dengan mengimpor dari format teks langsung (`*.txt`) atau format Excel (`*.xls` atau `*.xlsx`) atau format koma atau titik sebagai pemisah (`*.csv`). Pada RStudio, kalimat dengan tanda pagar (`#`) di bagian awal hanya untuk memberi keterangan saja.

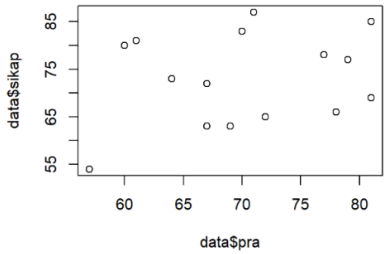
Siswa	Pra	Sikap	Pengetahuan	Keterampilan
1	64	73	64	82
2	71	87	71	69
3	70	83	70	71
4	79	77	79	65
5	57	54	57	67
6	61	81	61	87
7	67	72	67	88
8	72	65	72	88
9	81	69	81	66
10	67	63	67	70
11	60	80	60	88
12	69	63	69	71
13	81	85	81	82
14	77	78	77	76
15	78	66	78	72

Berikut adalah perintah yang dapat dijalankan untuk membaca data yang ada tersebut di RStudio.

```
#Membaca data
data <- read.csv2("Latihan_4var.csv", sep = ",")
```

Analisis awal yang kita dapat lakukan pada data yang ada tersebut yaitu membuat diagram pencar (*scatter plot*) antar variabel-variabel yang ada, misalnya antara variabel "pra" dan "sikap" dengan menggunakan perintah berikut dan hasilnya disajikan pada Gambar 2.11.

```
plot(data$pra,data$sikap)
```



Gambar 2.11 Diagram pencar antara variabel "pra" dan "sikap

Selain itu, kita juga dapat membentuk vektor rerata, matriks korelasi, dan matriks variansi-kovariansi berdasarkan data yang diberikan yang secara berturut-turut dilakukan menggunakan perintah-perintah berikut. Vektor rerata, matriks korelasi, dan matriks variansi-kovariansi dari data yang ada tersebut secara berturut-turut adalah sebagai berikut. Pada bagian selanjutnya, kami akan menyediakan penerapan lebih lanjut paket-paket yang tersedia dalam RStudio untuk melakukan analisis data multivariat.

```
> sapply(data[2:5],mean)
         pra         sikap  pengetahuan keterampilan
    70.26667     73.06667     70.26667     76.13333
> cor(data[2:5])
                    pra     sikap pengetahuan keterampilan
pra           1.0000000 0.2387165   1.0000000   -0.3243699
sikap         0.2387165 1.0000000   0.2387165    0.2488708
pengetahuan   1.0000000 0.2387165   1.0000000   -0.3243699
keterampilan -0.3243699 0.2488708  -0.3243699    1.0000000
> cov(data[2:5])
                   pra    sikap pengetahuan keterampilan
pra           60.35238 17.76667    60.35238    -22.10952
sikap         17.76667 91.78095    17.76667     20.91905
pengetahuan   60.35238 17.76667    60.35238    -22.10952
keterampilan -22.10952 20.91905   -22.10952     76.98095
```

Bab 3
MANOVA dan MANCOVA

Multivariate analysis of variance (MANOVA) merupakan suatu teknik analisis data untuk menguji perbandingan rata-rata (*means*) antara dua atau lebih kelompok data. Teknik ini merupakan bentuk multivariat (menggunakan dua atau lebih variabel terikat) dari teknik analisis data *analysis of variance* (ANOVA). Sebagai analisis multivariat, MANOVA mampu menganalisis perbedaan dua variabel dependen atau lebih, sedangkan ANOVA hanya terbatas pada satu variabel dependen saja.

Selain MANOVA ada juga *multivariate analysis of covariance* (MANCOVA). Sama seperti sebelumnya, MANCOVA adalah bentuk multivariat dari *analysis of covariance* (ANCOVA). MANCOVA merupakan gabungan antara MANOVA dan regresi multivariat. MANCOVA menggunakan setidaknya dua variabel dependen yang dianggap simultan. MANCOVA memiliki kemiripan dengan MANOVA, namun perbedaannya adalah terdapat tambahan variabel independen yang secara statistik berada di level *scale* (interval atau rasio) yang ditambahkan sebagai kovariat. Kovariat sendiri adalah variabel yang pengaruh atau efeknya akan kita hilangkan dalam melihat perbedaan variabel dependen berdasarkan kelompok atau aspek. Jumlah kovariat yang dapat kita masukkan tergantung pada ukuran sampel. Kovariat ditambahkan sehingga dapat mengurangi *error term* dan agar analisis menghilangkan efek kovariat pada hubungan antara variabel pengelompokan independen dan variabel dependen kontinu.

Sebagai contoh, pada suatu eksperimen semu pada pendidikan, ada variabel terikat prestasi belajar yang berupa sikap, pengetahuan, dan keterampilan pada kelompok dengan pembelajaran dengan pembelajaran berbasis masalah dan pembelajaran berbasis proyek. Pene-

liti dapat menggunakan variabel kecerdasan intelektual (IQ) sebagai kovariat. Jika peneliti akan membandingkan sikap, pengetahuan, dan keterampilan secara bersama-sama (simultan), peneliti dapat melakukan analisis tersebut dengan MANOVA. Jika pengaruh kecerdasan intelektual (IQ) terhadap hasil belajar diabaikan dahulu, sehingga yang dilihat pengaruhnya hanya dari pembelajaran, maka peneliti dapat menggunakan MANCOVA.

Perbedaan utama antara MANOVA dan MANCOVA adalah adanya huruf "C" yang berarti *covariance* atau *covariate* (kovariat). ANOVA, ANCOVA, MANOVA, dan MANCOVA sering digunakan dalam penelitian eksperimen. Teknik analisis data ini juga bisa diterapkan dalam jenis penelitian survei atau penggunaan data sekunder, sehingga penting untuk mengetahui MANOVA dan MANCOVA.

Beberapa contoh penelitian yang menggunakan MANOVA yaitu pengaruh strategi peta konsep terhadap motivasi belajar dan pemahaman konsep dalam pembelajaran IPA di SMPN 1 Yogyakarta dan pengaruh model PjBL terhadap kreativitas dan berpikir kritis siswa SMAN 1 Bantul. Adapun contoh penelitian pada MANCOVA yaitu pengaruh pekerjaan orang tua terhadap nilai ujian Matematika dan Fisika siswa kelas A SMA Tamansiswa Padang dengan dikontrol IQ dan perbandingan pengaruh metode pembelajaran dikontrol motivasi terhadap nilai praktik Kimia dan Biologi siswa kelas XII IPA SMAN 2 Yogyakarta.

Konsep dasar pada MANOVA dan MANCOVA

MANOVA merupakan pengembangan dari ANOVA, yaitu sebagai metode statistik suatu teknik statistik yang digunakan untuk menghitung pengujian signifikansi perbedaan rata-rata secara bersamaan antara kelompok untuk dua atau lebih variabel dependen atau terikat (Rencher, 1998). Berdasarkan jumlah variabel dependen dan independen, analisis statistik yang digunakan dapat dipahami pada Tabel 3.1.

Uji-*t* (*t*-test) berfungsi untuk menguji hipotesis penelitian mengenai pengaruh dari masing-masing variabel bebas secara parsial terhadap variabel terikat. Pada *t*-test, terdapat dua sampel bebas atau inde-

penden dengan hanya memiliki satu variabel dependen. Aplikasinya misalnya untuk melakukan pemeriksaan pada perlakuan sebelum (sampel 1) dan sesudah (sampel 2) atau melihat perbedaan performa pada kelompok yang pembaginya hanya dua kelompok (misalnya jenis kelamin, pria dan wanita).

Tabel 3.1 Metode statistik berdasarkan variabel bebas dan terikat

Aspek	Banyaknya variabel terikat	
Banyaknya variabel bebas	Satu	Dua atau lebih
Dua kelompok (kasus khusus)	*t*-test	Hotelling's T^2
Dua atau lebih kelompok (kasus umum)	Analysis of variance (ANOVA)	Multivariate analysis of variance (MANOVA)

Uji Hotelling's T^2 (Hotelling, 1931) berfungsi untuk melihat perbedaan antara dua kelompok percobaan yang masing-masing kelompok terdiri dari dua variat atau lebih, dan akan dilakukan analisis statistik pada variat tersebut secara serentak. Uji Hotelling's T^2 atau statistik T^2 pada dua sampel bebas adalah salah satu teknik analisis statistik komparasional multivariat yang digunakan untuk membandingkan dua kelompok sampel yang diteliti. Uji Hotelling's T^2 merupakan statistik multivariat yang menjadi pengembangan uji-*t* dua sampel bebas (perbedaan rerata dua kelompok yang bersifat independen). Perbedaannya antara keduanya terletak pada jumlah variabel dependen. Pada uji-*t* dua sampel bebas hanya memiliki satu variabel dependen, sedangkan uji Hotelling's T^2 memiliki lebih dari satu variabel dependen.

ANOVA berfungsi untuk menghitung pengujian signifikansi perbedaan rerata secara bersamaan antar kelompok untuk satu atau lebih variabel terikat. Sementara itu, MANOVA memiliki fungsi untuk menghitung pengujian signifikansi perbedaan rata-rata secara bersamaan antara kelompok untuk dua atau lebih variabel terikat. Selanjutnya, MANCOVA merupakan gabungan antara MANOVA dan regresi multivariat. Analisis MANCOVA merupakan analisis

yang melibatkan dua variabel terikat atau dependen yang dianggap simultan. MANCOVA memiliki kemiripan dengan MANOVA, namun terdapat selang independen yang ditambahkan sebagai kovariat (Gudono, 2017).

Selanjutnya, mengapa kita menggunakan MANOVA dan MANCOVA? Karena kedua uji tersebut memiliki kelebihan dalam penerapannya, yaitu sebagai berikut.

- Analisis multivariat dapat menghitung dan menganalisis lebih dari dua variabel secara bersamaan
- Mengetahui indikator pembentuk suatu variabel, menyediakan bukti validitas dan reliabilitas suatu instrumen, mengkonfirmasi ketepatan model dan menguji pengaruh suatu variabel terhadap variabel lain.
- Memungkinkan peneliti untuk menyelidiki hubungan antara kategori variabel
- Mengidentifikasi kelompok-kelompok variabel yang anggotanya memiliki kesamaan
- Membuat ringkasan informasi yang meringkas jumlah variabel yang banyak menjadi sejumlah faktor yang lebih sedikit (reduksi data).

Setiap uji tentu memiliki keterbatasan. Hal ini juga terjadi pada MANOVA dan MANCOVA. Adapun keterbatasan yang dimaksud yaitu membutuhkan ukuran sampel yang besar. Selain itu, kurangnya praktisi yang terlibat dalam penulisan artikel menyebabkan kurangnya perhatian analisis multivariat pada bidang manufaktur.

Asumsi pada MANOVA dan MANCOVA

Sebagaimana semua uji atau teknik analisis data, uji asumsi diperlukan untuk memeriksa apakah data yang akan kita analisis ini sesuai atau tidak dengan teknik analisis tersebut. Asumsinya yaitu jika datanya lolos uji asumsi, berarti datanya bisa kita asumsikan layak untuk dianalisis. Namun demikian, apabila data tersebut tidak lolos uji asumsi, berarti kita mengasumsikan bahwa datanya tidak layak untuk dianalisis, dan apabila dianalisis, hasil yang diperoleh tidak akan memenuhi standar. Inilah mengapa ujinya dinamakan uji asumsi.

Ada beberapa asumsi yang harus dipenuhi sebelum menganalisis menggunakan MANOVA atau MANCOVA. Adapun asumsi yang harus dipenuhi pada MANOVA di antaranya yaitu sebagai berikut (Huberty et al., 2006).

- Sampel dalam kelompok (*within group*) berdistribusi normal
- Observasi dalam (*within*) dan antar (*between*) sampel bersifat independen
- Variansi observasi dalam sel data adalah sama (*homogeneity of variances*)
- Homogenitas variansi (kovariansi antar sel homogen)
- Hubungan antar variabel-variabel dependen, hubungan antar kovariat (jika ada kovariat), dan hubungan variabel dependen dengan kovariat adalah linier.

Semua Asumsi MANCOVA sama dengan MANOVA, namun terdapat asumsi tambahan terkait dengan kovariat. Beberapa asumsi pengujian MANCOVA yang harus dipenuhi adalah sebagai berikut (Gudono, 2017):

- Normalitas

Distribusi normal multivariat merupakan perluasan dari distribusi normal univariat. Pada analisis multivariat asumsi multivariat normal perlu diperiksa untuk memastikan data pengamatannya mengikuti distribusi normal agar statistik inferensi dapat digunakan saat menganalisis data. Variabel-variabel $y_1, y_2, \dots, y_p$ dikatakan berdistribusi normal multivariat dengan parameter μ dan Σ jika mempunyai densitas peluang sebagai berikut.

$$f\left(y_1, y_2, \dots y_p\right) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(y-\pi)' \Sigma^{-1}(y-\pi)}$$

dengan y_i merupakan variabel yang diamati ($i = 1, 2, \dots p$), μ sebagai rata-rata sampel, dan Σ matriks variansi-kovariansi. Jika data $y_1, y_2, \dots, y_p$ berdistribusi normal multivariat, $(y-\pi)'\Sigma^{-1}(y-\pi)$ berdistribusi ${x_p}^2$. Berdasarkan sifat ini, pemeriksaan distribusi normal dilakukan dengan membuat *plot chi-square* (χ^2) guna menghitung jarak Mahalanobis dari setiap observasi terhadap *centroid*.

- Homogenitas

Untuk bisa melakukan analisis variansi (ANOVA, ANCOVA, MANOVA, MANCOVA), kita wajib mengasumsikan bahwa apabila

variabel-variabel yang ada kita interaksikan, maka variansi dan kovariansi mereka tidak terlalu jauh berbeda satu sama lainnya. Apabila data-datanya terlalu jauh berbeda satu sama lainnya, berarti analisis variansi tidak dapat dilakukan. Uji ini menggunakan Box's M dengan langkah sebagai berikut.

Hipotesis:

H_0: $\Sigma_1 = \Sigma_2 = ... = \Sigma_k$ (matriks variansi-kovariansi homogen)

H_1: $\exists \Sigma_i \neq \Sigma_j, i \neq j$ (terdapat matriks variansi-kovariansi yang tidak homogen)

Taraf signifikansi: $\alpha = 0.05$

Statistik uji:

$$M = \sum_{i=j}^{k} (n_i - 1) \ln|S| - \sum_{i=j}^{k} (n_i - 1) \ln|S_i|$$

$$C^{-1} = 1 - \left\{\frac{2p^2 + 3p - 1}{6(p+1)(k-1)}\right\}\left\{\sum_{i=j}^{k} \frac{1}{n_1 - 1} - \frac{1}{\sum_{i=j}^{k}(n_1 - 1)}\right\}$$

dengan $S = \frac{\sum_{i=j}^{k}(n_1 - 1)S_i}{\sum_{i=j}^{k}(n_1 - 1)}$, S adalah matriks kovariansi gabungan penduga bagi adalah matriks kovarians untuk $i = 1, 2, ..., k$, p adalah banyaknya respons yang diamati, dan $n_i$adalah ukuran sampel ke-i, selanjutnya menghitung MC^{-1}. Adapun kriteria keputusan pada uji tersebut yaitu bahwa H_0 ditolak jika $MC^{-1} > \chi^2_{v=\frac{1}{2}(k-1)(p)(p+1);(\alpha)}$.

- Multikolinearitas

Multikolinearitas berarti bahwa antara variabel dependen yang kita teliti, terdapat interaksi satu sama lain. Hal ini menandakan bahwa perubahan di variabel yang ingin kita amati, ternyata tidak murni akibat dari kategorinya atau perbedaan di variabel independennya.

Hipotesis:

H_0: $B = 0$ (variabel X tidak mempengaruhi variabel Y)

H_0: $B \neq 0$ (variabel X mempengaruhi variabel Y)

Taraf signifikansi: $\alpha = 0.05$ dengan statistik uji menggunakan formula Wilks' lambda sebagai berikut.

$$\Lambda = \frac{|E_{yx}|}{|E_{yx} + H_z|} = \frac{|E_{yy} - E_{yz}.E_{xx.}^{-1}E_{xy}|}{|E_{yy}|}$$

Statistik Wilks' lambda dapat ditransformasikan ke statistik F. Dengan demikian dapat dilakukan perbandingan dengan tabel distribusi F. Adapun kriteria keputusannya pada distribusi F, dengan H_0 ditolak jika $F_{\text{hitung}} > F_{\text{tabel}}$ atau sehingga dapat diartikan bahwa variabel konkomitan mempengaruhi terhadap variabel dependen.

- Linearitas

Linearitas artinya hubungan antara variabel independen dengan dependennya bersifat linier atau tidak jauh berbeda satu sama lainnya. Uji linearitas dan normalitas umum ditemui pada uji asumsi untuk teknik analisis data apa pun.

- Sifat variabel

Untuk uji asumsi yang satu ini, untungnya tidak diperlukan analisis statistika apa pun. Kita hanya perlu melihat apakah variabel independen (X) yang dilibatkan, bersifat kualitatif (berada di level nominal atau ordinal), sedangkan variabel dependen (Y) bersifat kuantitatif (berada di level interval atau rasio).

- Kesamaan kemiringan antar perlakuan (*homogeneity of regression slopes).*

Pada model MANCOVA, harus memenuhi asumsi bahwa terdapat hubungan variabel dependen dengan variabel konkomitan homogen antar perlakuan. Untuk menguji hipotesis ini, dilakukan perhitungan matriks jumlah kuadrat dan hasil kali silang galat tiap kelompok. Misalkan merupakan matriks jumlah kuadrat, maka hasil kali silang galat tiap kelompok adalah sebagai berikut:

$$E_{lk} = \begin{bmatrix} E_{xxlk} & E_{xylk} \\ E_{yxlk} & E_{yylk} \end{bmatrix}$$

Matriks untuk regresi dihitung secara terpisah pada masing-masing kelompok dan hasilnya dijumlahkan.

Hipotesis:

H_0: $B_1 = B_2 = B_3$ (koefisien regresi homogen antar perlakuan)

H_1: $\exists B_i \neq B_j$ untuk $i \neq j$, $i, j = 1,2,3$ (koefisien regresi tidak homogen antar perlakuan)

Taraf signifikansi: $\alpha = 0.05$ dengan statistik uji sebagai berikut.

$$H_{lk} = \sum_{l=1}^{g} \sum_{k=1}^{b} E_{yxlk} . E_{xxlk}^{-1} . E_{xylk} - E_{yx} . E_{xx}^{-1} . E_{xy}$$

Matriks jumlah kuadrat dalam model penuh dinyatakan melalui persamaan berikut.

$$E = E_{yy} - \sum_{l=1}^{g}\sum_{k=1}^{b} E_{yxlk}.E_{xxlk}^{-1}.E_{xylk}$$

Jika menggunakan statistik uji Wilks' lambda (Λ), statistik uji tersebut disajikan oleh persamaan berikut.

$$\Lambda = \frac{|E|}{|E + H_{lk}|} = \frac{E_{yy} - \sum_{l=1}^{g}\sum_{k=1}^{b} E_{yxlk}.E_{xxlk}^{-1}.E_{xylk}}{E_{yy} - E_{yx}.E_{xx}^{-1}.E_{xy}}$$

Kriteria keputusan dari uji tersebut yaitu bahwa H_0 ditolak jika nilai Wilks' lambda (Λ) sangat kecil (mendekati nol) atau nilai signifikansi dari statistik tersebut kurang dari taraf signifikansi. Harapan dari uji ini yaitu bahwa H_0 diterima, yang berarti terdapat kesamaan kemiringan pada kelompok perlakuan. Setelah asumsi MANCOVA terpenuhi, MANCOVA dapat dilakukan melalui uji Wilks' lambda dengan hipotesis tertentu tergantung pada tujuan dari masing-masing penelitian.

Contoh kasus dan analisis MANOVA dan MANCOVA menggunakan program R dan RStudio

Contoh kasus

Universitas Negeri Jawa Utara (UNJawara) sedang melakukan program pelatihan kepada para dosen, untuk meningkatkan performa akademik dan performa penelitian mereka. UNJawara ingin melihat apakah ada perbedaan antara yang tidak diberikan pelatihan atau tidak, dan model pelatihan apakah yang paling efektif dalam meningkatkan performa akademik dan penelitian tersebut. UNJawara paham bahwa para dosen memiliki IQ yang berbeda-beda satu sama lainnya (Data hasil pelatihan disajikan pada Tabel 3.2). Mereka ingin melihat pengaruh pelatihan murni tanpa melibatkan faktor IQ. Dalam kasus tersebut, kita menguji hipotesis berikut.

H_0: Pelatihan tidak berdampak pada performa akademik dan penelitian para dosen

H_1: Pelatihan berdampak pada performa akademik dan penelitian para dosen

Tabel 3.2 Data hasil pelatihan

No.	Jenis kelamin	Pangkat	IQ	Metode pelatihan	Performa akademik	Performa penelitian
1	1	2	89	1	80	98
2	1	1	75	1	72	95
3	2	2	96	2	79	100
4	1	1	101	3	90	98
5	1	2	122	3	88	92
6	1	3	144	3	100	93
7	1	3	102	1	68	92
8	1	2	116	3	93	84
9	1	2	117	2	74	90
10	1	2	119	3	92	89
11	1	3	101	2	78	87
12	2	3	105	2	84	85
13	1	4	131	3	90	87
14	2	3	98	1	65	87
15	2	1	122	2	83	91
16	2	2	97	1	79	87
17	2	2	127	2	82	85
18	1	3	99	1	80	79
19	2	3	90	1	75	90
20	2	2	108	2	75	84
21	1	1	90	1	75	84
22	1	1	113	2	72	87
23	1	1	112	3	89	92
24	1	2	100	1	78	84
25	2	1	102	3	87	85
26	1	4	111	1	80	83
27	1	2	100	1	74	79
28	1	2	107	3	82	90
29	1	2	91	1	72	80
30	1	2	121	2	80	84
31	2	3	83	3	90	79
32	1	1	84	2	68	83

No.	Jenis kelamin	Pangkat	IQ	Metode pelatihan	Performa akademik	Performa penelitian
33	1	2	105	3	90	79
34	1	2	107	3	90	74
35	1	3	86	2	78	83
36	2	2	93	3	90	79
37	2	2	101	1	76	90
38	2	2	105	1	71	80
39	2	1	129	2	83	84
40	2	1	128	2	82	79
41	1	3	116	3	99	83
42	2	3	97	2	89	79
43	1	1	127	1	82	74
44	1	1	112	2	75	74
45	1	2	100	3	82	74
46	2	2	102	3	96	74
47	2	2	84	1	72	74
48	2	4	86	2	80	73
49	2	1	93	2	80	74
50	1	2	129	1	66	70

Catatan. Jenis kelamin (data nominal): 1 = Pria dan 2 = Wanita; Pangkat (data ordinal): 1 = Penata, 2 = Asisten Ahli, 3 = Lektor, dan 4 = Lektor Kepala; Metode pelatihan (data nominal): 1 = Tidak diberikan pelatihan, 2 = *Teamwork building*, dan 3 = *Mentoring*. IQ merupakan data interval serta performa akademik dan penelitian merupakan data rasio.

Prosedur analisis

Analisis dimulai dengan membuat data Excel dengan format file .csv (*comma-separated value*) sebagai input untuk dianalisis menggunakan program R di bawah lingkungan kerja RStudio. Masukkan data yang tersedia pada Tabel 3.2 pada suatu Spreadsheet (misal Excel) dan simpan file dalam format .csv. Silakan simpan file yang memuat data tersebut dengan nama yang mudah diingat atau dengan nama yang singkat/pendek. Dalam kasus ini, kami memberi nama file tersebut dengan nama "dmm" untuk mengindikasikan bahwa file tersebut memuat data MANOVA dan MANCOVA. Atur folder yang memuat file tersebut sebagai direktori kerja pada RStudio dengan ca-

ra menekan menu `Session` > `Set Working Directory` > `Choose Directory` kemudian pilih folder di mana file tersebut disimpan. Tulis perintah berikut pada jendela atau panel Source, di mana di jendela atau panel Source tersebut kita dapat menuliskan perintah-perintah (*script*) yang akan dijalankan untuk analisis secara keseluruhan.

```
#Membuka data di R
data <- read.csv('dmm.csv', header= T, sep = ",")
data
```

Perintah tersebut kemudian dijalankan dengan blok perintah tersebut dan kemudian tekan tombol Run atau Ctrl + Enter. Jika sudah muncul data yang dimaksud pada jendela atau panel Console, maka data tersebut sudah siap untuk dianalisis lebih lanjut. Analisis lebih lanjut tersebut diawali dengan uji asumsi dengan menggunakan beberapa paket, yaitu 'MVN', 'biotools', dan 'jmv'. Jika paket-paket tersebut belum terpasang di RStudio, maka pasang terlebih dahulu paket-paket tersebut. Jika paket-paket tersebut telah terpasang, maka jalankan paket-paket tersebut dengan perintah berikut.

```
#Load library
library(MVN)
library(biotools)
library(jmv)
```

Setelah paket-paket yang diperlukan untuk uji asumsi telah dijalankan, langkah selanjutnya yaitu melakukan uji asumsi. Uji asumsi yang pertama kali dilakukan yaitu menguji sebaran data yang dimiliki, apakah mengikuti sebaran normalitas multivariat atau tidak. Ini dapat dilakukan dengan perintah berikut.

```
#Uji normalitas multivariat
> mvn(data[,c(4,6,7)], multivariatePlot = 'qq')
```

Hasil yang diharapkan melalui penggunaan perintah tersebut terkait uji asumsi normalitas multivariat yaitu sebagai berikut. Dengan mencermati hasil yang diperoleh, khususnya terkait hasil Henze-Zir-

kler, diperoleh temuan bahwa semua data yang ada, yaitu performa akademik, performa penelitian, dan IQ memenuhi asumsi normalitas multivariat. Lebih lanjut, analisis secara univariat untuk masing-masing variabel juga telah menunjukkan bahwa tiga variabel independen yang menjadi fokus dalam studi tersebut juga mengikuti suatu sebaran normal ($p > \alpha = 0{,}05$).

```
# Multivariat Normal
> mvn(data[,c(4,6,7)], multivariatePlot = 'qq')
$multivariateNormality
           Test        HZ                 p value               MVN
1 Henze-Zirkler       0.5637126             0.7348243             YES

$univariateNormality
Test                 Variable Statistic    p          value      Normality
1 Anderson-Darling    IQ                   0.4038     0.3430     YES
2 Anderson-Darling    Akademik             0.4473     0.2693     YES
3 Anderson-Darling    Penelitian           0.4879     0.2139     YES

$Descriptives
            N Mean   Std.Dev     Median Min Max  25th   75th      Skew       Kurtosis
IQ         50 105.46 15.043081   102    75  144  96.25  116.00    0.3250027  -0.5162051
Akademik   50 81.10  8.386700    80     65  100  75.00  88.75     0.2216334  -0.6137504
Penelitian 50 84.00  7.194102    84     70  100  79.00  89.75     0.1379416  -0.6672508
```

Asumsi selanjutnya yang perlu diuji yaitu asumsi homogenitas. Uji ini dapat dilakukan dengan menggunakan perintah berikut. Dengan menggunakan perintah tersebut dapat diperoleh hasil bahwa ketiga variabel kuantitatif yang dilibatkan dalam studi tersebut, yaitu IQ, performa akademik, dan performa penelitian homogen ($\chi^2(12) = 13{,}432$, $p = 0{,}3384 > \alpha = 0{,}05$).

```
# Uji homogenitas
boxM(data = data[,c(4,6,7)], grouping = data[,5])
       Box's M-test for Homogeneity of Covariance Matrices

data:  data[, c(4, 6, 7)]
Chi-Sq (approx.) = 13.432, df = 12, p-value = 0.3384
```

Setelah asumsi homogenitas terpenuhi, asumsi berikutnya yang perlu disediakan yaitu asumsi multikolinearitas. Dengan menggunakan perintah berikut diperoleh hasil bahwa tidak terdapat hubungan yang bermakna pada antara dua variabel dependen. Hal ini ditunjukkan oleh koefisien korelasi Pearson yang sangat kecil atau mendekati

nol, yaitu sekitar 0,04. Dengan demikian, asumsi-asumsi yang diperlukan untuk analisis lebih jauh, yaitu uji hipotesis, telah terpenuhi.

```
# Uji multikolinearitas
cor.test(x=data$Akademik,y=data$Penelitian,method='pearson'
)$estimate
       cor
0.04667833
```

Setelah semua asumsi telah terpenuhi, langkah selanjutnya yaitu melakukan uji hipotesis menggunakan MANOVA dan MANCOVA. Uji hipotesis menggunakan dua analisis tersebut disajikan secara terpisah. *Pertama*, kami menyajikan uji hipotesis berdasarkan MANOVA sebagai berikut. Dalam uji hipotesis menggunakan MANOVA ini, tiga variabel akan dilibatkan, yaitu pelatihan (X), performa akademik (Y1), dan performa penelitian (Y2). Perintah yang digunakan dan hasil yang diperoleh dari perintah tersebut yaitu sebagai berikut.

```
uji1 <- manova(cbind(data$Akademik, data$Penelitian)~ data$Pelatihan, data =
data)
summary(uji1)

jmv::anovaOneW(
    formula = Akademik + Penelitian ~ Pelatihan,
    data = data,
    phMethod = "tukey",
    phTest = TRUE,
    phFlag = TRUE)

summary(uji1)
                Df  Pillai  approx F num  Df den Df   Pr(>F)
data$Pelatihan   1   0.61364 37.323        2      47   1.97e-10 ***
Residuals        48
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Hasil MANOVA yang diperoleh di atas menunjukkan bahwa metode pelatihan yang dipakai secara simultan signifikan menentukan performa akademik dan performa penelitian para dosen ($p < 0{,}01$) pada taraf signifikansi 0,05. Dengan kata lain, perbedaan jenis metode pelatihan yang dipakai akan mempengaruhi perbedaan performa akademik dan performa penelitian para dosen.

POST HOC TESTS

Tukey Post-Hoc Test – Akademik

		1	2	3
1	Mean difference	–	-4.529412	-16.08824
	t-value	–	-2.621014	-9.167563
	df	–	47.00000	47.00000
	p-value	–	0.0310421	< .0000001
2	Mean difference		–	-11.55882
	t-value		–	-6.586567
	df		–	47.00000
	p-value		–	0.0000001
3	Mean difference			–
	t-value			–
	df			–
	p-value			–

Note. * p < .05, ** p < .01, *** p < .001

Tukey Post-Hoc Test – Penelitian

		1	2	3
1	Mean difference	–	0.2352941	-0.6176471
	t-value	–	0.09350600	-0.2417057
	df	–	47.00000	47.00000
	p-value	–	0.9951917	0.9683290
2	Mean difference		–	-0.8529412
	t-value		–	-0.3337840
	df		–	47.00000
	p-value		–	0.9405134
3	Mean difference			–
	t-value			–
	df			–
	p-value			–

Note. * p < .05, ** p < .01, *** p < .001

Melalui uji lanjutan (*post-hoc*) kita menemukan bahwa metode pelatihan yang paling efektif dalam meningkatkan performa akademik dan performa penelitian yaitu metode *mentoring*. Adapun dengan melihat hasil pelatihan dari sisi performa akademik maupun penelitian, kita menemukan bahwa perbedaan skor performa akademik jauh lebih terlihat dibandingkan dengan perbedaan skor performa penelitian. Pada perbandingan antara kelompok kontrol yang tidak diberikan pelatihan (1), kelompok yang diberikan pelatihan dengan menggunakan metode *teamwork building* (2), dan kelompok yang

diberikan pelatihan berupa metode *mentoring* (3), kita menemukan bahwa secara signifikan, pelatihan melalui metode *teamwork building* memberikan dampak performa akademik dosen yang lebih baik dibanding tidak diberikan pelatihan (*mean difference* = 4,529412, *p* = 0,0310421 < 0,05). Begitu juga dengan pelatihan dengan metode *mentoring* yang menunjukkan dampak pada performa akademik dosen yang lebih baik secara signifikan dibandingkan dengan tanpa pelatihan (*mean difference* = 16,08824, *p* < 0,05). Secara umum, dengan demikian, pelatihan dengan metode *mentoring* secara signifikan menjadi yang paling baik dari segi dampaknya terhadap performa akademik dosen. Pada *post-hoc* performa penelitian, kita tidak menemukan adanya perbedaan signifikan antar ketiga kelompok berdasarkan mtode pelatihan. Ini mengindikasikan bahwa pelatihan tidak mempengaruhi performa penelitian dari dosen.

Kita selanjutnya melakukan MANCOVA untuk mengontrol IQ. Perintah yang digunakan untuk melakukan analisis tersebut dan hasil yang diperoleh melalui perintah tersebut yaitu sebagai berikut.

MANCOVA

Multivariate Tests

		value	F	df1	df2	p
Pelatihan	Pillai's Trace	0.66783704	11.530310	4	92	0.0000001
	Wilks' Lambda	0.3326321	16.512201	4	90	< .0000001
	Hotelling's Trace	2.00491433	22.054058	4	88	< .0000001
	Roy's Largest Root	2.00421066	46.096845	2	46	< .0000001
IQ	Pillai's Trace	0.05798749	1.385033	2	45	0.2607805
	Wilks' Lambda	0.9420125	1.385033	2	45	0.2607805
	Hotelling's Trace	0.06155703	1.385033	2	45	0.2607805
	Roy's Largest Root	0.06155703	1.385033	2	45	0.2607805

Univariate Tests

	Dependent Variable	Sum of Squares	df	Mean Square	F	p
Pelatihan	Akademik	2253.441176	2	1126.720588	46.08519380	< .0000001
	Penelitian	6.352941	2	3.176471	0.05781509	0.9438929
IQ	Akademik	68.421108	1	68.421108	2.79856432	0.1011375
	Penelitian	2.319981	1	2.319981	0.04222609	0.8380961
Residuals	Akademik	1124.637715	46	24.448646		
	Penelitian	2527.327078	46	54.941893		

ASSUMPTION CHECKS

Box's Homogeneity of Covariance Matrices Test

χ^2	df	p
0.5827412	6	0.9966812

Shapiro-Wilk Multivariate Normality Test

W	p
0.9824936	0.6607890

Adapun ketika kita melanjutkan dengan menggunakan MANCOVA untuk mengontrol IQ, kita menemukan hasil bahwa pelatihan sebetulnya lebih disebabkan oleh faktor IQ, bukan oleh metode pelatihan yang digunakan. Ini ditunjukkan dengan signifikansi antara MANOVA yang melibatkan metode pelatihan (X1), IQ (X2), performa akademik (Y1), dan performa penelitian (Y2), yang mana kita menemukan hasil yang signifikan ($p < 0{,}001$). Akan tetapi, ketika kita mengontrol IQ sebagai kovariat, pengaruh pelatihan menjadi tidak signifikan. Ini mengindikasikan bahwa IQ lebih berpengaruh terhadap performa akademik dan penelitian dosen, dibandingkan metode pelatihan yang dilakukan kepada para dosen tersebut.

Bab 4
Analisis Diskriminan

Dalam kehidupan sehari-hari, kita sering kali dipertemukan dengan adanya sebuah kondisi di mana pengelompokan-pengelompokan terhadap hal-hal tertentu itu penting. Dalam memandang baik atau buruknya sesuatu hal, tanpa kita sadari kita sering kali menjadi terbiasa dengan adanya proses pengelompokan-pengelompokan dalam hidup yang kita jalani. Misalnya saja pada beberapa peristiwa berikut.

Peristiwa 1. Dalam suatu universitas X, Indeks prestasi kumulatif (IPK) atau *grade point average* (GPA) digunakan dalam mengelompokkan mahasiswa yang mendaftar di universitas tersebut ke dalam salah satu kategori di dalam program studi pascasarjana yang ada pada sekolah tersebut. Jika di masa yang akan datang terdapat calon mahasiswa dengan IPK/GPA tertentu mendaftar di universitas tersebut, maka mahasiswa tersebut akan digolongkan ke dalam salah satu kategori yang sudah ditentukan universitas X tersebut.

Peristiwa 2. Dalam suatu Sekolah Menengah Atas (SMA), nilai Ujian Akhir Sekolah (UAS) dari berbagai mata pelajaran dijadikan sebagai pertimbangan pihak sekolah untuk menentukan siswa masuk ke dalam kelas penjurusan, misalnya jurusan IPA dan IPS, ataupun IPA, IPS, dan Bahasa. Jika di waktu yang akan datang terdapat calon siswa baru dengan nilai UAS tertentu memilih jurusan di sekolah tersebut, maka siswa tersebut akan dikelompokkan ke dalam salah satu jurusan yang sudah ditentukan SMA tersebut.

Peristiwa 3. Seorang analis keuangan ingin mengetahui variabel-variabel atau faktor-faktor yang membedakan atau memisahkan antara perusahaan yang sehat dengan perusahaan yang mengalami kebangkrutan.

Peristiwa 4. Seorang manajer keuangan di suatu perusahaan ingin mengidentifikasi faktor-faktor yang dapat membedakan antara kon-

sumen yang ingin membeli produk merek tertentu. Dengan informasi inilah dapat diprediksi penjualan produk tersebut.

Beberapa peristiwa yang digambarkan di atas sering dikenal sebagai analisis diskriminan (*discriminant analysis*). Analisis diskriminan adalah teknik analisis multivariat yang digunakan untuk mengelompokkan observasi-observasi ke dalam salah satu kategori (dalam hal ini kelompok atau populasi) berdasarkan pada variabel-variabel tertentu. Analisis diskriminan bertujuan mengklasifikasikan suatu objek ke dalam kelompok yang saling lepas (*mutually exclusive* atau *disjoint*) dan menyeluruh (*exhaustive*) berdasarkan faktor penjelas.

Analisis diskriminan telah banyak digunakan dalam penelitian bidang pendidikan, seperti penelitian yang dilakukan oleh Erimafa et al. (2009) yang berjudul "*Application of discriminant analysis to predict the class of degree for graduating students in a university system*", penelitian yang dilakukan Akintunde dan Mattew (2019) yang berjudul "*Discriminant analysis of psycho-social predictors of mathematics achievement of gifted students in Nigeria*", dan penelitian yang dilakukan Koybasi (2020) yang berjudul "*Examining teachers' disposition towards sustainable education through discriminant analysis*".

Teori dasar pada analisis diskriminan

Konsep dasar pada analisis diskriminan

Analisis diskriminan adalah teknik analisis multivariat yang digunakan untuk mengelompokkan observasi-observasi ke dalam salah satu kategori (dalam hal ini kelompok atau populasi) berdasarkan pada variabel-variabel tertentu. Analisis diskriminan merupakan perluasan dari analisis regresi. Analisis diskriminan menggunakan kombinasi linear dari dua atau lebih variabel independen untuk membentuk suatu fungsi diskriminan yang digunakan dalam membedakan satu kelompok dengan kelompok lainnya. Persamaan kombinasi linear analisis diskriminan yaitu $D = b_1X_1 + b_2X_2 + b_3X_3 + \cdots + b_kX_k$, dengan D merupakan skor diskriminan, b adalah koefisien diskriminan, dan X merupakan variabel independen atau prediktor.

Analisis diskriminan akan menghasilkan suatu fungsi yang dapat membedakan antara dua kelompok atau lebih. Terbentuknya fungsi tersebut dikarenakan adanya pengaruh antara beberapa variabel bebas yang dapat membedakan dua atau lebih kelompok populasi yang ada terhadap variabel terikat. Pada dasarnya, fungsi diskriminan dapat digunakan untuk mendeskripsikan variabel-variabel bebas suatu observasi yang dapat membedakan dari kelompok populasi yang ada. Dengan kata lain, analisis diskriminan merupakan suatu metode yang dapat digunakan sebagai kriteria pengelompokan yang dilakukan berdasarkan perhitungan statistik terhadap kelompok populasi. Perhitungan fungsi diskriminan akan menghasilkan suatu nilai. Nilai-nilai yang dihasilkan dari fungsi diskriminan dikenal dengan skor diskriminan.

Analisis diskriminan, seperti yang telah disebutkan sebelumnya, bertujuan mengklasifikasikan suatu objek ke dalam kelompok-kelompok yang saling lepas dan menyeluruh berdasarkan faktor penjelas. Menurut Rencher (1995), terdapat dua tujuan utama pemisahan kelompok dalam analisis diskriminan, yaitu: (1) *aspek deskriptif*, yaitu menggambarkan pemisahan kelompok, di mana fungsi diskriminan digunakan untuk mendeskripsikan atau menjelaskan perbedaan antara dua atau beberapa kelompok dan (2) *aspek prediksi*, yaitu mengelompokkan pengamatan ke dalam kelompok, di mana beberapa variabel digunakan untuk menentukan satu sampel individu atau objek ke dalam salah satu dari beberapa kelompok.

Analisis diskriminan secara lebih spesifik memiliki tujuan sebagai berikut.

- Mengidentifikasi variabel-variabel yang dapat membedakan atau memisahkan antar kelompok atau populasi
- Menggunakan variabel-variabel yang telah teridentifikasi tersebut untuk menyusun suatu fungsi yang dapat menjelaskan perbedaan atau pemisahan antar kelompok atau populasi.
- Menggunakan variabel-variabel yang telah teridentifikasi tersebut untuk menyusun suatu cara atau aturan pengelompokan observasi di masa datang ke dalam salah satu kelompok atau populasi.

Jenis-jenis variable pada analisis diskriminan

Analisis diskriminan merupakan salah satu dari analisis statistika multivariat. Analisis statistika multivariat merupakan analisis statistika yang digunakan pada data yang memiliki lebih dari dua variabel secara bersamaan, dengan menggunakan teknik analisis multivariat maka dapat menganalisis pengaruh beberapa variabel terhadap variabel-variabel lainnya dalam waktu yang bersamaan. Teknik analisis multivariat berdasarkan karakteristiknya dapat dibagi menjadi dua, yaitu teknik dependensi dan teknik interdependensi. Teknik dependensi ini merupakan teknik yang digunakan untuk melihat pengaruh atau memprediksi variabel dependen berdasarkan beberapa variabel independen yang mempengaruhi. Analisis multivariat yang termasuk teknik dependensi yaitu analisis regresi berganda, analisis diskriminan, analisis konjoin, MANOVA, MANCOVA, ANOVA, ANCOVA, dan korelasi kanonis. Teknik interdependensi merupakan teknik yang digunakan untuk mengelompokkan atau mereduksi beberapa variabel menjadi variabel baru yang lebih sedikit, tetapi tidak mengurangi informasi yang terkandung oleh variabel asli. Analisis multivariat yang termasuk teknik interdependensi yaitu analisis klaster, penskalaan multidimensi, analisis korelasi kanonis, dan analisis faktor.

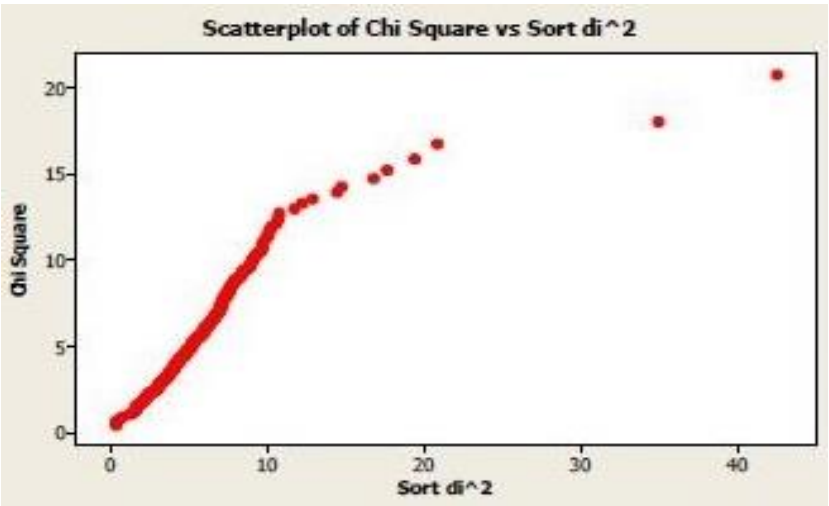
Analisis diskriminan merupakan salah satu teknik analisis multivariat yang termasuk teknik dependensi, yaitu melihat pengaruh variabel dependen berdasarkan beberapa variabel independen. Analisis diskriminan digunakan pada kasus dimana variabel bebas berupa data metrik (yaitu interval atau rasio) dan variabel terikat berupa data non-metrik (yaitu nominal atau ordinal). Analisis diskriminan ditandai dengan ciri khusus, yaitu data variabel dependen yang harus berupa data kategori, sedangkan data independen berupa data non-kategori. Jika ditinjau pada persamaan $Y_1 = X_1 + X_2 + X_3 + \cdots + X_n$, maka jelas bahwa data untuk variabel Y_1 berupa data merik, sedangkan data untuk variabel X_1, X_2, ..., X_n merupakan data non-metrik.

Asumsi-asumsi pada analisis diskriminan

Berikut ini adalah asumsi-asumsi yang harus dipenuhi agar analisis diskriminan dapat digunakan atau hasil yang diperoleh dapat bermakna atau menghasilkan kesalahan yang kecil.

- Normalitas multivariat

Variabel independen harusnya berdistribusi normal. Jika tidak berdistribusi normal, hal ini akan menyebabkan masalah pada ketepatan fungsi (model diskriminan). Untuk memeriksa data apakah berdistribusi normal multivariat, Q-Q *plot* antara *square distance* (d_j^2) atau jarak Mahalanobis dengan nilai kuantil dari distribusi $\chi^2\left(\frac{j-0,5}{n}\right)$ dapat digunakan. Jika hasil plot menunjukkan suatu garis lurus, maka data tersebut dapat dikatakan berdistribusi normal multivariat. Hipotesis nol yang digunakan dalam uji normalitas multivariat ini yaitu data berdistribusi normal multivariat, sedangkan hipotesis alternatifnya yaitu data tidak berdistribusi normal multivariat. Hipotesis alternatif ditolak yang berarti data berdistribusi normal multivariat ketika *scatter plot* atau Q-Q *plot* yang terbentuk mengikuti pola suatu garis lurus (lihat Gambar 4.1).



Gambar 4.1 Plot d_j^2 terhadap sebaran khi-kuadrat (Q-Q *plot* normalitas multivariat)

- Homoskedastisitas

Matriks kovarians dari semua variabel independen seharusnya sama atau homogen Matriks kovarians dua kelompok relatif sama atau dengan kata lain kondisi homoskedastisitas terpenuhi. Pelanggaran terhadap asumsi ini akan mempengaruhi ketepatan klasifikasi dan hasil uji signifikansi. Pengujian dapat dilakukan dengan uji Box's M. Untuk menguji kesamaan matriks variansi-kovariansi kelompok I (S_1) dan kelompok II (S_2) digunakan dalam hipotesis.

- Kesamaan vektor nilai rataan

Uji kesamaan nilai vektor rataan dari kelompok-kelompok (*test of equality of group means*) dapat dilakukan sebagai berikut. Statistik uji yang digunakan dalam pengujian hipotesis tersebut adalah statistik Wilks' lambda (Λ) yang nilainya berkisar antara 0 sampai 1 dengan ketentuan: jika nilai Wilks' lambda mendekati 0, maka data tiap kelompok cenderung berbeda. Akan tetapi, jika nilai Wilks' lambda mendekati 1, berarti data tiap kelompok cenderung sama (tidak berbeda). Selain menggunakan Wilks' lambda, uji tersebut dapat dilakukan melalui uji F, di mana ketika nilai signifikansi yang diperoleh dari uji tersebut kurang dari taraf signifikansi, maka hipotesis nol pada uji tersebut ditolak. Hasil yang diharapkan dari uji ini yaitu bahwa kita gagal menolak hipotesis nol.

Jenis-jenis analisis diskriminan

Analisis diskriminan adalah teknik multivariat untuk memisahkan objek-objek dalam kelompok yang berbeda dan mengelompokkan objek baru ke dalam kelompok-kelompok tersebut. Tujuan utamanya yaitu mengetahui perbedaan antarkelompok. Pada umumnya peubah respons atau variabel dependen terdiri dari dua klasifikasi, namun pada beberapa kasus peubah respons atau variabel dependen tersebut memiliki lebih dari dua klasifikasi. Jika hanya terdapat dua klasifikasi, maka analisis diskriminan tersebut disebut analisis diskriminan dua kelompok *two-groups discriminant analysis*. Akan tetapi, jika terdapat lebih dari dua klasifikasi, maka analisis diskriminan tersebut disebut *multiple discriminant analysis* (MDA).

Jenis uji analisis diskriminan mana yang sebaiknya digunakan dapat ditentukan berdasarkan terpenuhinya asumsi-asumsi pada anali-

sis diskriminan. Terpenuhi atau tidaknya asumsi dalam analisis diskriminan ini menentukan uji yang digunakan. Apabila asumsi normal multivariat dan homoskedastisitas terpenuhi, analisis diskriminan yang sebaiknya digunakan yaitu analisis diskriminan linear (*linear discriminant analysis*, LDA). Apabila asumsi normal multivariat terpenuhi, sedangkan asumsi homoskedastisitas tidak terpenuhi, analisis diskriminan yang sebaiknya digunakan yaitu analisis diskriminan kuadratik (*quadratic discriminant analysis*, QDA). Selanjutnya, apabila asumsi normal multivariat tidak terpenuhi, sedangkan asumsi homoskedastisitas terpenuhi, maka analisis diskriminan yang cocok untuk digunakan yaitu analisis diskriminan dengan metode Fisher. Terakhir, apabila kedua asumsi tersebut tidak terpenuhi, maka analisis sebaiknya dilakukan menggunakan regresi logistik. Penjelasan pada tiga jenis analisis diskriminan tersebut yaitu sebagai berikut.

- Analisis diskriminan linear (*linear discriminant analysis*, LDA)

Analisis diskriminan linear merupakan metode analisis diskriminan yang digunakan apabila terdapat kondisi data berdistribusi normal multivariat dan asumsi keidentikan matriks variansi-kovariansi antarkelompok terpenuhi. Fungsi diskriminan merupakan kombinasi linear asal yang akan menghasilkan cara terbaik dalam pemisahan kelompok. Banyaknya fungsi diskriminan yang terbentuk tergantung dari g kelompok dan p banyaknya variabel bebas. Fungsi diskriminan linear yang terbentuk mempunyai bentuk umum berupa persamaan linier yaitu sebagai berikut.

$$\boldsymbol{y_1 = I_{i1}X_1 + I_{i2}X_2 + \cdots + I_{ip}X_p}$$

dengan $i = 1, 2, \ldots, g$, atau dapat ditulis sebagai berikut.

$$L(x) = ax = \bar{x}_{i^t} S^{-1}_{pooled} x$$

di mana L merupakan skor diskriminan linear, a merupakan vektor koefisien pembobot fungsi diskriminan, $\bar{x}$ merupakan vektor nilai rataan kelompok ke I, dan S_{pooled} merupakan vektor nilai rataan kelompok ke-i dengan formula sebagai berikut.

$$S_{pooled} = \frac{(n_1 - 1)\mathrm{S}_1 + (n_1 - 1)\mathrm{S}_2 + \cdots + (n_g - 1)\mathrm{S}_g}{(n_1 + n_1 + \cdots + n_g - g)}$$

Berdasarkan fungsi diskriminan linear, dapat diperoleh skor diskriminan linier yang digunakan untuk mengalokasikan x ke dalam kelompok k, jika

$$L_k(x) = maks\ (L_1(x), L_2(x), \dots, L_g(x))$$

Oleh karena itu, pengklasifikasian x ke dalam kelompok k dapat menggunakan perbandingan skor diskriminan linier maksimum dengan titik tengah dari *optimum cutting score* (m) yang didefinisikan sebagai berikut.

$$m = \frac{1}{2}(\bar{x}_1 - \bar{x}_2)^t S_{pooled}^{-1}(\bar{x}_1 - \bar{x}_2)$$

dengan aturan pengelompokan yaitu jika $L > m$, maka objek pengamatan akan diklasifikasikan ke dalam kelompok 1. Apabila $L \leq m$, maka objek pengamatan akan diklasifikasikan ke dalam kelompok 2.

- Analisis diskriminan kuadratik (*quadratic discriminant analysis*, QDA)

Analisis diskriminan kuadratik merupakan metode analisis diskriminan yang digunakan apabila terdapat kondisi data berdistribusi normal multivariat dan asumsi keidentikan matriks variansi kovariansi antar kelompok tidak terpenuhi. Ada dua jenis fungsi diskriminan kuadratik, yaitu analisis diskriminan kuadratik dua kelompok dan analisis diskriminan kuadratik g kelompok. Fungsi diskriminan kuadratik dua kelompok dibentuk berdasarkan pada asumsi bahwa kedua kelompok menyebar normal multivariat dan matriks variansi kovariansi dari dua kelompok tidak sama. Sementara itu, sama halnya dengan diskriminan kuadratik dua kelompok, diskriminan kuadratik g kelompok memiliki distribusi normal multivariat dan matriks kovariansi dari g kelompok berbeda.

- Analisis diskriminan dengan metode Fisher

Prinsip utama dari fungsi diskriminan Fisher adalah pemisahan sebuah populasi. Fungsi diskriminan yang terbentuk dapat digunakan untuk pengelompokan suatu observasi berdasarkan kelompok-kelompok tertentu. Metode Fisher ini tidak mengasumsikan data harus berdistribusi normal, tapi dalam perhitungan salah satu syarat yang harus diperhatikan adalah data yang digunakan harus memiliki matriks kovariansi yang sama untuk setiap kelompok populasi yang diberi-

kan. Analisis diskriminan dengan metode fisher ini didasarkan atas fungsi diskriminan yang mempunyai bentuk umum sebagai berikut.

$$Y_1 = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon_1$$

Analisis diskriman dengan metode Fisher merupakan salah satu metode dalam mendapatkan fungsi diskriminan. Metode linear Fisher sebenarnya berasal dari statistik klasifikasi linear untuk dua populasi normal. Pada metode ini pengamatan multivariat X ditransformasikan kepengamatan univariat Y, dimana Y berasal dari populasi pertama dan kedua untuk dipisahkan sebanyak mungkin untuk pengamatan lainnya. Fisher menyarankan untuk mengambil kombinasi linear dari X untuk menghasilkan Y yang merupakan suatu fungsi yang cukup sederhana untuk pemasangan dari X.

Langkah-langkah analisis diskriminan

Langkah-langkah yang ditempuh dalam melakukan analisis diskriminan diberikan sebagai berikut.

1. Memilah variabel-variabel yang menjadi variabel terikat (dependen) dan variabel bebas (independen).
2. Melakukan uji normalitas multivariat.
3. Melakukan uji homoskedastisitas.
4. Melakukan uji korelasi (multikolinearitas).
5. Melakukan uji MANOVA/kesamaan vektor nilai rataan.
6. Menentukan metode untuk membuat fungsi diskriminan, yaitu metode *simultaneous estimation*, dimana semua variabel dimasukkan secara bersama-sama lalu dilakukan proses diskriminan dan metode *stepwise estimation*, dimana variabel dimasukkan satu per satu ke dalam model diskriminan. Rencher (1998) menyebutkan bahwa apabila suatu pengamatan menggunakan banyak variabel, maka untuk mengefisiensi pemilihan variabel yang akan digunakan dalam pembentukan fungsi diskriminan yaitu melalui analisis diskriminan bertahap atau langkah demi langkah (*stepwise discriminant analysis*). Terdapat tiga model yang dapat digunakan dalam analisis diskriminan bertahap, yaitu sebagai berikut.
   - *Forward selection*, yaitu model yang memasukkan variabel masing-masing *step* dengan memilih nilai statistik *F* maksi-

mum. Proporsi dari statistik F maksimum yang melebihi F_α adalah lebih besar dari α.

- *Backward elimination*, yaitu operasi yang mirip dengan menghapus variabel yang kurang berkontribusi pada masing-masing tahapan, indikasi yang digunakan adalah memilih F.
- *Stepwise selection*, merupakan kombinasi dari *forward selection* dan *backward elimination*. Variabel yang dipilih masing-masing *step* adalah variabel yang diuji kembali, apakah variabel yang dimasukkan awal memiliki statistik F yang besar di antara variabel yang gagal untuk dimasukkan. Prosedur *stepwise selection* ini sudah terkenal pelaksanaannya.

Ada tiga kriteria agar suatu variabel dapat dimasukkan dalam pembentukan fungsi diskriminan, yaitu memiliki nilai *F* terbesar, memiliki nilai Wilk's lambda terkecil, dan variabel yang memiliki signifikansi kurang dari taraf signifikansi yang digunakan.

7. Menguji signifikansi fungsi diskriminan yang terbentuk. Wilks' lambda, Pillai's, dan uji F merupakan tiga di antara beberapa uji yang dapat digunakan untuk menguji signifikansi dari fungsi diskriminan yang terbentuk.
8. Menguji ketepatan klasifikasi dari fungsi diskriminan. Ketepatan klasifikasi dari fungsi diskriminan dapat diidentifikasi melalui *apparent error rate* (APER). Menurut Johnson (2007), APER menyatakan nilai yang menunjukkan perbandingan keanggotaan kelompok aktual dengan keanggotaan kelompok prediksi seperti yang ditunjukkan dalam Tabel 4.1:

Tabel 4.1 Kesalahan klasifikasi

Hasil pengamatan	Hasil prediksi	
	Kelompok 1	Kelompok 2
Kelompok 1	n_{11}	n_{12}
Kelompok 2	n_{21}	n_{22}

APER dihitung dengan menggunakan persamaan $APER = \frac{n_{12}+n_{21}}{n_1+n_2}$. Secara umum, APER dapat dihitung melalui persamaan $APER = \frac{N-\sum_{i=1}^{g} n_{ii}}{N}$, dengan N menyatakan banyaknya penga-

matan pada semua kelompok, n_{ii} menyatakan banyaknya pengamatan yang tepat diklasifikasikan dari kelompok aktual ke-i pada kelompok prediksi I, dan g adalah banyaknya kelompok.

9. Melakukan interpretasi fungsi diskriminan.
10. Menyediakan bukti validitas fungsi diskriminan. Pada analisis diskriminan akan dibuat sebuah model seperti regresi, yaitu antara satu variabel terikat (dependen) dan banyak variabel bebas (independen). Prinsip analisis diskriminan yaitu ingin membuat model yang dapat secara jelas menunjukkan perbedaan (diskriminasi) antar isi variabel dependen. Penyediaan bukti validitas fungsi diskriminan membutuhkan data *training* dan *testing*.

    Data *training* digunakan algoritma klasifikasi untuk membentuk sebuah model *classifier.* Model ini merupakan representasi pengetahuan yang akan digunakan untuk memprediksi kelompok dari data baru yang belum pernah ada. Data *testing* digunakan untuk mengukur sejauh mana *classifier* berhasil melakukan klasifikasi dengan benar. Data yang ada pada data *testing* seharusnya tidak boleh ada pada data *training* sehingga dapat diketahui apakah model *classifier* sudah tepat atau belum dalam melakukan klasifikasi. Ukuran data *training* yaitu hasil kali antara proporsi data *training* dengan ukuran data keseluruhan. Sementara itu, ukuran data *testing* yaitu selisih antara data keseluruhan dengan ukuran data *training.*

Contoh kasus dan analisis diskriminan menggunakan program R dan RStudio

Contoh kasus

Suatu sekolah menengah atas menyediakan empat jurusan, yaitu Matematika dan Ilmu Pengetahuan Alam (MIPA) (*a*), Ilmu Pengetahuan Sosial (IPS) (*b*), Bahasa (*c*), dan Agama (*d*). Sekolah mengelompokkan peserta didik yang masuk ke masing-masing jurusan tersebut berdasarkan lima variabel, yaitu nilai rata-rata mata pelajaran Matematika di semester 1 dan 2 (P1), nilai rata-rata mata pelajaran IPA di semester 1 dan 2 (P2), nilai rata-rata mata pelajaran Ekonomi dan Sosiologi di semester 1 dan 2 (P3), nilai rata-rata mata pelajaran

Bahasa Indonesia dan Bahasa Inggris di semester 1 dan 2 (P4), dan nilai rata-rata mata pelajaran Al-Qur'an Hadits dan Akidah-Akhlak di semester 1 dan 2 (P5). Hasil penentuan jurusan peserta didik pada sekolah tersebut disajikan pada Tabel 4.2.

Tabel 4.2 Hasil penjurusan peserta didik

Jurusan	P1	P2	P3	P4	P5
a	82	85	86,5	85	80
a	85	84	84	85	76
a	84	87	89	86	80
a	84	84	85	83	83
a	82	87	87	85	83
a	80	86	86	89	79,5
⋮	⋮	⋮	⋮	⋮	⋮
b	79	79	85	82	80
b	78	82,5	84	83	82
b	75	81	85	83	79
b	76,5	80	82	80	84
b	75	86	82	83	80
b	79	78	86	87	83
b	79	80,5	80	81	84
⋮	⋮	⋮	⋮	⋮	⋮
c	81,5	80	77,5	75	77
c	77	79	81	78	76
c	78	79	77,5	77	83
c	78	79,5	82	77	81
c	77	81	79,5	77	81
⋮	⋮	⋮	⋮	⋮	⋮
d	79	73,5	76,5	77	84
d	76	79,5	81	78	80
d	75	74,5	78,5	77	79
d	77	73	81	80,5	81,5
d	78	75	75	77,5	79
d	79	73	78	76	77
⋮	⋮	⋮	⋮	⋮	⋮

Dari kasus dan data yang disajikan, akan diselidiki jawaban atas empat pertanyaan berikut.

1. Apakah pengelompokan peserta didik ke dalam jurusan yang dilakukan oleh sekolah telah sesuai?
2. Berapa persen ketepatan sekolah dalam melakukan pengelompokan peserta didik ke dalam masing-masing jurusan?
3. Berapa kesalahan klasifikasi sekolah dalam melakukan penjurusan peserta didik?
4. Bagaimana persamaan model yang dapat digunakan oleh sekolah untuk melakukan penjurusan peserta didik?

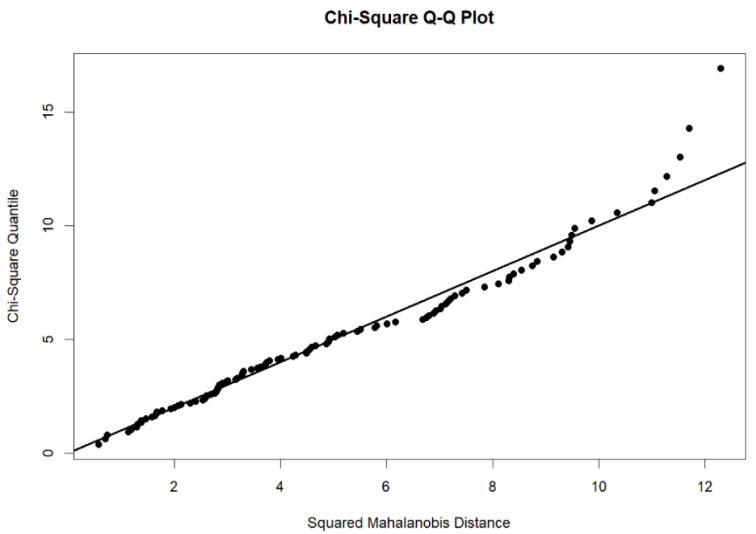
Prosedur analisis

Penyelesaian dari kasus yang ada di sekolah tersebut dapat dilakukan dengan menggunakan analisis diskriminan. Analisis diskriminan digunakan untuk memprediksi pengelompokan peserta didik ke dalam masing-masing jurusan, menentukan persentase ketepatan prediksi pengelompokan peserta didik ke dalam masing-masing jurusan, menentukan kesalahan prediksi atau klasifikasi oleh sekolah dalam melakukan penjurusan peserta didik, dan menentukan persamaan model diskriminan yang dapat digunakan oleh sekolah untuk melakukan penjurusan peserta didik. Variabel yang ada dalam kasus ini ada dua, yaitu variabel dependen (kategori) berupa jurusan dan variabel independen (prediktor) berupa nilai rata-rata mata pelajaran Matematika dan IPA di semester 1 dan 2, nilai rata-rata mata pelajaran Ekonomi dan Sosiologi di semester 1 dan 2, nilai rata-rata mata pelajaran Bahasa Indonesia dan Bahasa Inggris di semester 1 dan 2, dan nilai rata-rata mata pelajaran Al-Qur'an Hadits dan Akidah-Akhlak di semester 1 dan 2.

Ada tiga uji asumsi yang perlu dilakukan dalam analisis diskriminan, yaitu asumsi distribusi normal multivariat, homoskedastisitas, dan kesamaan vektor nilai rataan. Asumsi yang pertama diselidiki, terutama ketika hendak menggunakan analisis diskriminan jenis LDA, yaitu apakah variabel prediktor yang digunakan dalam analisis diskriminan berdistribusi normal multivariat. Untuk menyelidiki terpenuhinya asumsi ini dengan menggunakan program R dan RStudio, diperlukan paket 'MVN', sehingga paket ini perlu terlebih dahulu di-

pasang untuk selanjutnya dapat digunakan. Perintah yang digunakan untuk keperluan ini dan hasil yang diperoleh yaitu sebagai berikut.

```
install.packages("MVN")
library(MVN)
mvn(data.jur[,2:6], multivariatePlot = 'qq')
$multivariateNormality
           Test        HZ   p value MVN
1 Henze-Zirkler 0.9554043 0.1206278 YES
```



Gambar 4.2 Q-Q *plot* distribusi normal multivariat

Berdasarkan hasil pengujian normal multivariat, variabel prediktor berdistribusi normal secara multivariat (p = 0,121 > α = 0,05). Q-Q *plot* yang tersaji pada Gambar 4.2 juga menunjukkan titik-titik menyebar merata sepanjang garis linear dan tidak ditemukan penyimpangan yang jauh. Hal ini mengindikasikan bahwa variabel prediktor berdistribusi normal secara multivariat.

Asumsi kedua yang akan dibuktikan yaitu asumsi homoskedastisitas. Uji asumsi ini digunakan untuk melihat apakah matriks variansi dan kovariansi antar kelompok bersifat homogen atau tidak. Penggunaan LDA mensyaratkan bahwa uji asumsi ini terpenuhi. Paket 'biotools' dapat digunakan untuk memeriksa terpenuhi atau tidaknya

asumsi ini. Melalui paket tersebut, kita dapat melakukan uji Box's M yang merupakan uji yang digunakan untuk memeriksa asumsi homoskedastisitas. Hasil uji Box's M menunjukkan bahwa matriks variansi dan kovariansi antar kelompok bersifat homogen, dan dengan demikian asumsi homoskedastisitas terpenuhi ($\chi^2(45) = 55{,}133$, $p = 0{,}1432$).

```
install.packages("biotools")
library(biotools)
boxM(data = data.jur[,2:6],grouping = data.jur[,1])

*** Hasil Uji Homoskedastisitas

        Box's M-test for Homogeneity of Covariance Matrices

data:  data.jur[, 2:6]
Chi-Sq (approx.) = 55.133, df = 45, p-value = 0.1432
```

Asumsi berikutnya yang perlu terpenuhi yaitu asumsi kesamaan vektor nilai rataan. Uji asumsi ini digunakan untuk melihat kesamaan vektor nilai rataan yaitu apakah terdapat perbedaan antar variabel prediktor ditinjau dari variabel kategori. Berdasarkan hasil uji kesamaan vektor nilai rataan, terdapat perbedaan rataan yang signifikan antar variabel prediktor ($p < 0{,}05$). Setelah dilakukan pengujian seluruh asumsi untuk analisis diskriminan, kita dapat mengetahui bahwa seluruh asumsi dapat terpenuhi, sehingga dapat dilanjutkan dengan pengujian LDA.

```
mnv <- manova(cbind(P1,P2,P3,P4,P5)~JUR,data=data.jur)
summary.manova(mnv,test = "Wilks")

*** Hasil Uji Kesamaan Vektor Nilai Rataan
           Df   Wilks approx F num Df den Df    Pr(>F)
JUR         3 0.11365   22.088     15 276.46 < 2.2e-16 ***
Residuals 104
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Setelah asumsi-asumsi diperlukan sudah terpenuhi, langkah selanjutnya yaitu menentukan variabel prediktor yang signifikan. Analisis diskriminan bertahap dapat dilakukan untuk menentukan varia-

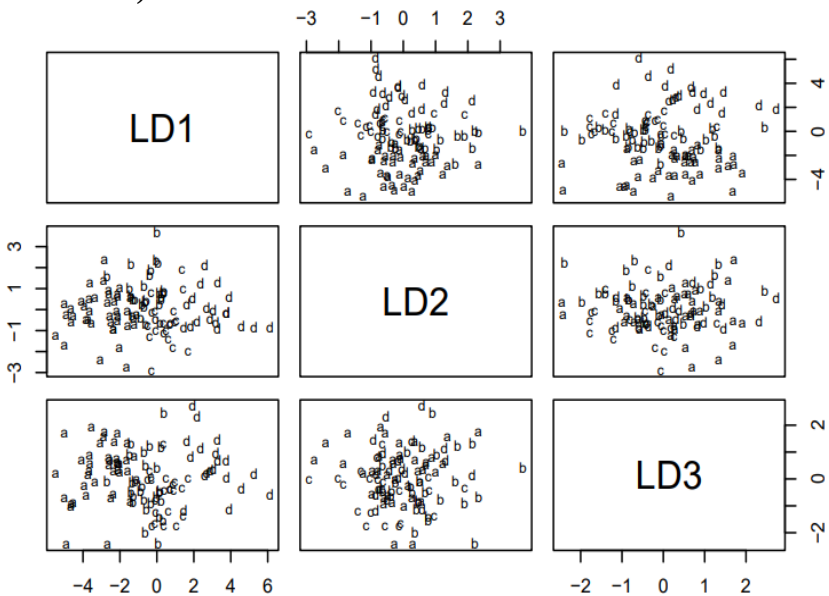
bel prediktor yang signifikan dalam pembentukan fungsi diskriminan. Analisis tersebut digunakan untuk menentukan variabel prediktor yang dominan dalam pembentukan fungsi diskriminan. Berdasarkan hasil uji signifikansi variabel prediktor yang digunakan seperti yang ditunjukkan di bawah ini, semua variabel prediktor yang digunakan bersifat signifikan (p < 0,05) dengan kontribusi tertinggi pada variabel P2 dan terendah pada variabel P4.

```
Pil<-greedy.wilks(JUR~P1+P2+P3+P4+P5,data=data.jur,nivaeu=1)
Values calculated in each step of the selection procedure:
  vars Wilks.lambda F.statistics.overall p.value.overall F.statistics.diff p.value.diff
1   P2    0.3420362             66.68714    3.949386e-24         66.687139 3.949386e-24
2   P1    0.2302517             37.21747    2.140973e-30         16.668416 6.478601e-09
3   P3    0.1600686             30.99286    5.355757e-36         14.907528 3.943452e-08
4   P5    0.1336676             25.40459    1.156058e-37          6.649581 3.800339e-04
5   P4    0.1136509             22.08782    5.388398e-39          5.870818 9.777748e-04
```

Langkah selanjutnya yaitu membentuk fungsi atau persamaan model diskriminan. Fungsi atau persamaan diskriminan ini menunjukkan suatu kombinasi linear dari berbagai variabel prediktornya. Pembentukan fungsi atau persamaan model diskriminan ini dilakukan melalui LDA dengan menggunakan paket ‘MASS’.

```
install.packages(“MASS”)
library(MASS)
fit <- lda(JUR~P1+P2+P3+P4+P5,data.jur,na.action="na.omit")
Prior probabilities of groups:
        a         b         c         d
0.3703704 0.2500000 0.2037037 0.1759259

Group means:
        P1       P2       P3       P4       P5
a 81.51250 83.73750 85.42500 84.47500 80.08750
b 79.40926 80.50000 81.62963 81.55556 81.70370
c 78.13636 80.18182 81.22727 79.18182 79.75000
d 77.10526 74.57895 78.10526 77.15789 78.78947

Coefficients of linear discriminants:
          LD1         LD2         LD3
P1 -0.3241236 -0.01587025  0.20503078
P2 -0.2561956 -0.09213567 -0.27737803
P3 -0.1962062 -0.38342751 -0.01960151
P4 -0.1078330  0.43762132  0.24233182
P5 -0.1037539  0.29907025 -0.20755799
Proportion of trace:
   LD1    LD2    LD3
0.9213 0.0484 0.0303
```

Prior probabilities of groups yang ditunjukkan pada analisis di atas menunjukkan probabilitas keanggotaan setiap kelompok atau kategori. Hasil analisis telah menunjukkan bahwa keanggotaan pada kelompok jurusan MIPA menjadi yang tertinggi di antara kelompok jurusan lainnya dan keanggotaan pada kelompok jurusan Agama menjadi yang terendah. Koefisien diskriminasi linear yang diperoleh pada analisis di atas digunakan untuk membentuk persamaan fungsi diskriminan. Berdasarkan hasil yang didapatkan, terdapat tiga fungsi diskriminan linear yang terbentuk, yaitu LD1, LD2, dan LD3 (lihat Gambar 4.3).



Gambar 4.3 *Plot* diskriminan linear (*linear discriminant*, LD)

Persamaan fungsi diskriminan dapat digunakan untuk mengklasifikasikan objek baru ke dalam kategori yang ada. Berdasarkan hasil *proportion of trace*, fungsi diskriminan 1 (LD1) merupakan fungsi diskriminan terbaik sehingga dapat dikatakan bahwa fungsi diskriminan tersebut dapat memberikan pemisahan yang maksimal yaitu sebesar 92,13%. Oleh karena itu, fungsi diskriminan yang selanjutnya digunakan yaitu LD1. Persamaan fungsi diskriminan yang terbentuk yaitu $D = -0{,}32414236X_1 - 0{,}2561956X_2 - 0{,}1962062X_3 -$

$0{,}1078330X_4 - 0{,}10375392X_5$, dengan D merupakan skor diskriminan dan X merupakan variabel prediktor atau variabel independen.

Langkah selanjutnya yaitu menentukan ketepatan klasifikasi. Menentukan ketepatan klasifikasi berarti menentukan seberapa tepat klasifikasi yang dilakukan oleh sekolah pada data pengamatan berdasarkan fungsi diskriminan yang terbentuk. Selanjutnya, kesalahan klasifikasi merujuk pada tingkat kesalahan klasifikasi yang dilakukan pada data pengamatan. Ketepatan klasifikasi dapat ditentukan dengan menggunakan paket 'klaR'.

```
install.packages("klaR")
library(klaR)
j <- lda(JUR~P1+P2+P3+P4+P5,data.jur,na.action = "na.omit",CV=T)
conf <- table(list(observed=data.jur$JUR,predicted=j$class))
conf
*** Menentukan tabel klasifikasi hasil observasi dan prediksi
        predicted
observed  a  b  c  d
       a 38  2  0  0
       b  1 23  3  0
       c  0  3 19  0
       d  0  0  1 18
*** Menentukan persentase ketepatan hasil klasifikasi
        a         b         c         d
0.9500000 0.8518519 0.8636364 0.9473684
[1] 0.9074074
```

Berdasarkan hasil yang didapatkan di atas, ketepatan hasil klasifikasi yang dilakukan pada data pengamatan berdasarkan fungsi diskriminan yang terbentuk yaitu sebesar 90,74%. Ini berarti kesalahan klasifikasi yang dilakukan sebesar 9,26%. Ketepatan klasifikasi tertinggi yaitu pada jurusan MIPA (95%) diikuti oleh jurusan Agama (94,74%), dan yang terendah pada jurusan IPS (85,19%). Ketepatan hasil klasifikasi juga dapat diidentifikasi berdasarkan tabel klasifikasi (hasil observasi dan prediksi). Hasil analisis yang menunjukkan tabel klasifikasi dan prediksi menunjukkan bahwa ada sebanyak 38 peserta didik yang menurut data observasi dan fungsi diskriminan masuk di jurusan IPA dan dua peserta didik yang menurut fungsi diskriminan masuk di jurusan IPS. Ini berarti ada dua peserta didik yang dikelompokkan secara keliru.

Langkah berikutnya yaitu mengklasifikasikan objek baru. Analisis diskriminan tidak hanya digunakan untuk menentukan seberapa besar ketepatan klasifikasi yang dilakukan, tetapi juga dapat digunakan untuk memprediksi objek baru yang hendak diklasifikasikan ke dalam kelompok tertentu. Penentuan prediksi objek baru ini dilakukan dengan memasukkan skor pada variabel prediktor ke dalam persamaan fungsi diskriminan untuk kemudian dibandingkan dengan nilai *centroid* yang didapatkan. Misalkan ada peserta didik baru yang hendak dikelompokkan ke dalam jurusan di sekolah tersebut. Nilai rata-rata setiap mata pelajaran yang didapatkan yaitu P1 = 81, P2 = 79, P3 = 84, P4 = 78, dan P5 = 80. Nilai yang didapatkan tersebut selanjutnya disubstitusikan ke dalam persamaan fungsi diskriminan yang telah terbentuk sebagai berikut. Hasil skor diskriminan yang didapatkan kemudian dibandingkan dengan nilai *centroid*.

$$D = (-0{,}32414236 \times 81) - (0{,}2561956 \times 79) - (0{,}1962062 \times 84) - (0{,}1078330 \times 78) - (0{,}10375392 \times 80)$$

Klasifikasi objek baru dapat juga dilakukan berdasarkan fungsi linear masing-masing kategori dalam analisis diskriminan. Masing-masing kategori didapatkan *intercept* dan koefisien yang digunakan untuk membentuk fungsi diskriminan.

```
#Intercept pada kelompok jurusan a
a <- data.jur %>%filter(data.jur$JUR == 'a')
a <- data.jur%>%filter(JUR=='a')
a_mean <- sapply(a[, -1], mean)
a_mean
          [,1]
[1,] -1941.545
#Koefisien
a_cov <- cov(a[, -1])
a_cov
          P1       P2      P3       P4       P5
[1,] 31.1959 16.28191 18.4095 7.338922 20.81888
```

```
#Intercept pada kelompok jurusan b
b <- data.jur %>% filter(JUR == 'b')
b_mean <- sapply(b[, -1], mean)
b_mean
          [,1]
[1,] -1855.311
#Koefisien
b_cov <- cov(b[, -1])
b_cov
           P1     P2       P3       P4       P5
[1,] 30.29842 15.714 17.56725 7.392626 20.97058
```

```
#Intercept pada kelompok jurusan c
c <- data.jur %>%filter(JUR == 'c')
c_mean <- sapply(c[, -1], mean)
c_mean
          [,1]
[1,] -1801.532
#Koefisien
c_cov <- cov(c[, -1])
c_cov
           P1       P2       P3       P4       P5
[1,] 29.91794 15.67325 17.91558 6.579182 20.50851
```

```
#Intercept pada kelompok jurusan d
d <- data.jur %>%filter(JUR == 'd')
d_mean <- sapply(d[, -1], mean)
d_mean
          [,1]
[1,] -1699.936
#Koefisien
d_cov <- cov(d[, -1])
d_cov
          P1       P2       P3       P4       P5
[1,] 29.2621 14.62154 17.15055 6.801304 20.16388
```

Persamaan fungsi linier diskriminan yang terbentuk pada masing-masing kategori jurusan disajikan sebagai berikut: Masing-masing skor dari tiap variabel prediktor dimasukkan ke dalam masing-masing fungsi diskriminan untuk kemudian bandingkan. Objek diklasifikasikan sesuai skor diskriminan (D) yang tertinggi.

- Jurusan a

$$D = -1941{,}545 + 31{,}1959P1 + 16{,}28191P2 + 18{,}4095P3 + 7{,}338922P4 + 20{,}81888P5$$

- Jurusan b

$$D = -1855{,}311 + 30{,}29842P1 + 15{,}714P2 + 17{,}56725P3 + 7{,}392626P4 + 20{,}97058P5$$

- Jurusan c

$$D = -1801{,}532 + 29{,}91794P1 + 15{,}67325P2 + 17{,}91558P3 + 6{,}579182P4 + 20{,}50851P5$$

- Jurusan d

$$D = -1699{,}936 + 29{,}2621P1 + 14{,}62154P2 + 17{,}15055P3 + 6{,}801304P4 + 20{,}16388P5$$

Bab 5
Regresi Logistik

Cara yang biasanya dipilih untuk mengembangkan model penelitian yaitu menggunakan model regresi berganda. Namun, ketika variabel dependen berupa data kategoris, menggunakan model regresi berganda bukan pilihan yang tepat. Sebagai contoh, ketika seorang peneliti di bidang kesehatan ingin meneliti pengaruh faktor-faktor pada tingkat atau status kesembuhan pasien COVID-19, yaitu cepat atau lambat sembuh. Oleh karena itu dibutuhkan alternatif pemodelan lain, dimana regresi logistik merupakan alternatif pilihan yang tepat.

Analisis regresi logistik telah banyak digunakan dalam penelitian di bidang pendidikan, seperti penelitian yang dilakukan oleh Prette et al. (2012) yang berjudul "*Role of social performance in predicting learning problems: Prediction of risk using logistic regression analysis*", penelitian Wahyuni et al. (2018) yang berjudul "Analisis regresi logistik terhadap keputusan penerimaan beasiswa PPA di FMIPA UNNES menggunakan *software* Minitab", penelitian oleh Şirin dan Şahin (2020) yang berjudul "*Investigation of factors affecting the achievement of university students with logistic regression analysis: school of physical education and sport example*", dan penelitian Fuente-Mella et al. (2021) yang berjudul "*Multinomial logistic regression to estimate the financial education and financial knowledge of university students in Chile*".

Teori dasar pada regresi logistik

Konsep dasar pada regresi logistik

Regresi logistik merupakan suatu metode dalam analisis statistik untuk menggambarkan hubungan antara variabel respons (variabel dependen) yang bersifat kategoris dengan satu atau lebih variabel inde-

penden dalam skala kontinu dan/atau kategoris. Karena variabel independen bisa bersifat kontinu dan atau kategoris, sedangkan variabel dependennya harus kategoris, hubungan antara variabel independen dan dependennya tidak linear. Karena hubungan antara variabel tersebut tidak linear, metode *ordinary least square* (OLS) tidak dapat digunakan sebagaimana yang dilakukan di regresi linear. Apabila dipaksakan menggunakan OLS, akan terjadi pelanggaran asumsi, yakni *error* tidak terdistribusi normal, ragam (variansi) dari *error* tidak homogen (terjadi heteroskedastisitas pada ragam *error*), dan nilai duga Y (*fitted value*) melebihi rentang antara nol dan satu.

Analisis regresi logistik menjadi alternatif dari analisis diskriminan yang digunakan ketika variabel respons berupa kategori. Kelebihan analisis regresi logistik yaitu tidak memerlukan asumsi normal ganda dan informasi ragam antar kelompok/kelas respons homogen. Selain itu, dalam hal interpretasi, regresi logistik lebih mudah dilakukan sebagaimana regresi linear. Akan tetapi, untuk variabel respons yang kelasnya lebih dari dua, analisis diskriminan lebih baik untuk digunakan.

Analisis regresi logistik dikategorikan ke dalam tiga jenis, yaitu analisis regresi logistik biner, analisis regresi multinomial, dan analisis regresi logistik ordinal.

- *Analisis regresi logistik biner*. Hanya terdapat dua kelas kategori variabel respons atau dependen pada analisis regresi logistik biner. Misalnya, keberhasilan studi mahasiswa yang dinyatakan dengan keterangan lulus atau gagal, kondisi seorang pasien apakah mengidap kanker atau tidak, nasabah bank lancar bayar atau gagal membayar tagihan, dan sebagainya.
- *Analisis regresi multinomial*. Terdapat lebih dari dua kelas kategori variabel respons pada analisis regresi multinomial. Sebagai contoh yaitu faktor-faktor yang mempengaruhi pilihan perguruan tinggi pada siswa SMA dan SMK yang dibedakan menjadi empat kategori, yaitu SMA negeri dalam kota, SMA negeri luar kota, SMA swasta dalam kota, dan SMA swasta luar kota; faktor-faktor yang mempengaruhi perbedaan status pendidikan yang dibedakan menjadi tiga kategori, yaitu berstatus tidak/belum sekolah, masih bersekolah, dan tidak sekolah lagi; dan faktor-faktor yang

mempengaruhi pelanggaran lalu lintas yang dibedakan menjadi empat kategori, yaitu jenis kelamin pengendara, usia pengendara, jenis kendaraan, dan waktu kejadian pelanggaran.

- *Analisis regresi logistik ordinal*. Terdapat lebih dari dua kelas kategori variabel respons dengan skala data ordinal pada analisis regresi ordinal. Contohnya, kesehatan mental guru selama pandemi COVID-19 yang dinyatakan dengan keterangan gejala gangguan stress rendah, sedang atau tinggi; tingkat pendidikan seseorang yang dinyatakan dengan keterangan SD, SMP, SMA; dan status akreditasi sekolah yang dinyatakan dengan cukup, baik, unggul.

Bab ini secara khusus membahas tentang analisis regresi logistik biner, yaitu jenis regresi logistik yang variabel dependennya hanya terdiri dari dua kelompok, yaitu 0 dan 1, positif atau negatif, sukses atau gagal, berhasil atau tidak berhasil, lulus atau tidak lulus, dan sebagainya.

Salah satu tujuan analisis regresi logistik adalah untuk mengidentifikasi hubungan linear variabel prediktor/independen dengan variabel respons/dependen. Identifikasi hubungan tersebut dilakukan untuk mengetahui seberapa penting variabel prediktor yang digunakan. Oleh karena itu, ukuran *statistical significance* (pengaruh nyata) dari variabel independen menjadi penting digunakan. Tujuan lain analisis regresi logistik adalah klasifikasi, yaitu teknik multivariat yang berfokus untuk memprediksi atau membedakan kelompok/kelas dari sekumpulan pengamatan yang diamati. Dalam hal ini, model regresi logistik disusun untuk mendapatkan aturan pemisahan pengamatan (pengklasifikasian). Aturan tersebut digunakan untuk memprediksi pengamatan baru yang belum diketahui kelas responsnya. Untuk tujuan ini, ukuran yang digunakan yaitu keakuratan model regresi logistik dalam memprediksi kelas respons dari pengamatan atau sekumpulan pengamatan baru.

Asumsi-asumsi pada regresi logistik

Asumsi-asumsi yang perlu dipenuhi pada analisis regresi logistik yaitu sebagai berikut (Gudono, 2017). *Pertama*, variabel dependen harus bersifat kategoris (biasanya dikotomus). Data kategoris yaitu data yang menjelaskan karakteristik dari data tersebut. Sebagai contoh

jenis kelamin, bahasa, dan kewarganegaraan. Variabel dependen pada regresi logistik biasanya menggunakan skala dikotomus. Skala dikotomus yang dimaksud yaitu skala data nominal dengan dua kategori, misalnya ya dan tidak, baik dan buruk, atau tinggi dan rendah. *Kedua*, tidak ada korelasi yang signifikan antar variabel independen. *Ketiga*, linearitas dalam format logit. Hubungan antara logit dependen variabel dengan variabel independennya harus linier. Jika jumlah variabel independen lebih dari satu, maka salah satu cara sederhana yang dapat digunakan yaitu melihat koefisien variabel interaksi antar variabel independen tersebut. Jika interaksi tersebut signifikan, maka kemungkinan besar terdapat hubungan tidak linier. *Keempat*, jumlah observasi untuk setiap variabel harus memadai dan jumlah sampel secara keseluruhan cukup besar.

Hosmer dan Lemesbow (1980) menyatakan bahwa ukuran sampel setidaknya 400 unit untuk bisa mendapatkan *goodness of fit* yang baik. Uji *goodness of fit* dilakukan untuk menentukan apakah model yang dibentuk sudah tepat atau tidak. Itu dikatakan tepat apabila tidak ada perbedaan signifikan antara model dengan nilai observasinya (Lemeshow & Hosmer, 1982). Peducci et al. (Gudono, 2017) menguraikan bahwa ukuran sampel untuk regresi logistik dapat dihitung dengan rumus $N = 10k/p$ dengan k menyatakan jumlah variabel independen (kovariat) dan p menyatakan proporsi terkecil dari kategori dalam variabel dependen. Long (1997) menyatakan bahwa jika hasil rumus tersebut kurang dari 100, maka sebaiknya sampel ditambah agar menjadi minimum 100.

Model persamaan pada regresi logistik

Bab ini menjelaskan regresi logistik biner, yaitu variabel dependen yang hanya terdiri dari dua kelompok, sukses atau gagal, ada juga yang menggunakan istilah positif atau negatif, yang dituliskan sebagai 1 dan 0. Atas dasar tersebut, peluang sukses (positif) dinotasikan dengan $P(Y = 1)$. Pada kondisi variabel prediktor tunggal, itu dinotasikan dengan $\pi(x)$ untuk menekankan bahwa nilai dari $P(Y = 1)$ bergantung pada nilai dari variabel dependen x.

Model regresi logistik, diperkenalkan oleh Joseph Berkson pada tahun 1944, merupakan bentuk linear dari logit *odds ratio*, yaitu

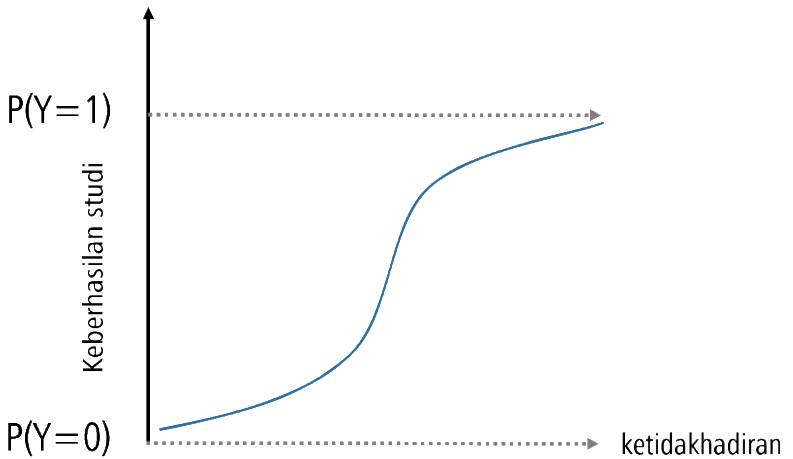
ukuran perbandingan kejadian sukses $(Y = 1)$ dan gagal $(Y = 0)$. Logit *odds ratio* dinotasikan dengan $logit[\pi(x)]$. Fungsi regresi logistik merupakan logit peluang sukses yang didefinisikan sebagai logaritma natural dari *odds* sebagai berikut (Agresti, 2013).

$$\begin{aligned} logit[\pi(x)] &= logit(Y = 1) \\ &= \ln(odds) \\ &= \ln\left(\frac{\pi(x)}{1-\pi(x)}\right) \\ &= \beta_0 + \beta_1 x_1 \end{aligned}$$

Persamaan tersebut dapat dinyatakan dalam bentuk eksponensial sebagai berikut.

$$P(Y = 1) = \frac{\exp(\beta_0 + \beta_1 x_1)}{1 + \exp(\beta_0 + \beta_1 x_1)}$$

Persamaan tersebut menggambarkan hubungan antara variabel independen dengan variabel dependen yang membentuk kurva logistik atau *sigmoid* yang berbentuk seperti huruf S. Misalkan dibuat grafik hubungan jumlah ketidakhadiran mahasiswa dengan keberhasilan studi yang dinyatakan dengan lulus atau tidak, yang diberi label 1 dan 0 secara berurut. Grafik hubungan kedua variabel tersebut akan berbentuk kurva *sigmoid* seperti yang tersaji pada Gambar 5.1.



Gambar 5.1 Kurva sigmoid hubungan antara ketidakhadiran mahasiswa dan keberhasilan studi mereka

Kemudian, jika terdapat p prediktor, maka fungsi regresi logistik akan menjadi

$$\ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

atau

$$P(Y=1) = \frac{\exp\left(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p\right)}{1 + \exp\left(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p\right)} = \frac{\exp(\boldsymbol{\beta X})}{1 + \exp(\boldsymbol{\beta X})}$$

Pendugaan parameter dan pengujian hipotesis pada regresi logistik

Pendugaan parameter pada regresi logistik menggunakan pendugaan *maximum likelihood*. Untuk ukuran sampel n dan p variabel independen, *likelihood* untuk regresi logistik biner diberikan sebagai berikut.

$$l(\boldsymbol{\beta}) = \prod_{i=1}^{n} \pi(\boldsymbol{x}_i)^{y_i}(1-\pi(\boldsymbol{x}_i))^{1-y_i}$$

$$L(\boldsymbol{\beta}) = \frac{\prod_{i=1}^{n} \exp\left(y_i(b_0 + b_1 x_{i1} + \cdots + b_p x_{ip})\right)}{\prod_{1=1}^{n}(1 + \exp(b_0 + b_1 x_{i1} + \cdots + b_p x_{ip}))}$$

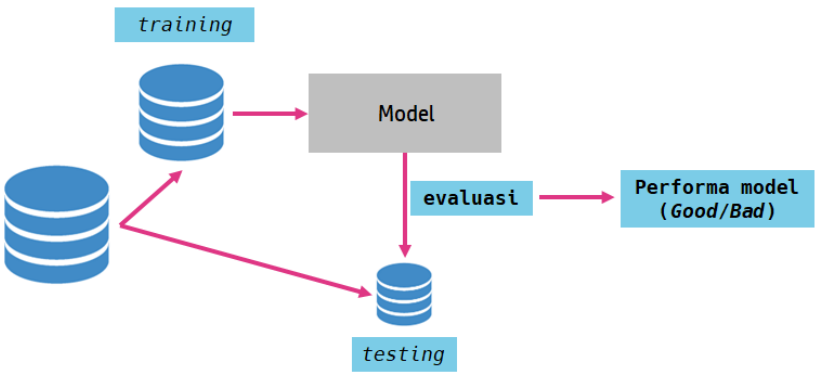
Likelihood ratio test (LRT) digunakan untuk menguji secara simultan variabel prediktor. Hipotesis nol pada uji ini menyatakan bahwa semua koefisien regresi bernilai 0, yaitu $H_0: \beta_p = 0$. *Maximum likelihood* dari hipotesis nol dinotasikan dengan L_0 yang merepresentasikan model tanpa variabel independen/prediktor. Kemudian, hipotesis alternatif menyatakan bahwa setidaknya ada sebuah koefisien regresi yang tidak bernilai nol. *Maximum likelihood* dari hipotesis alternatif dinotasikan dengan L_1. Rasio dari L_0 dan L_1 menghasilkan nilai LRT yang dikenal sebagai statistik G yang didefinisikan sebagai berikut.

$$G^2 = -2\ln\left(\frac{L_0}{L_1} = \frac{likelihood\ tanpa\ prediktor}{likelihood\ dengan\ prediktor}\right)$$

Statistik G tersebut mengikuti sebaran *chi-square* dengan derajat bebas 1 sesuai unsur koefisien regresi di H_0, yaitu β_0.

Tahap berikutnya yaitu pengujian parsial untuk mengidentifikasi variabel prediktor yang berpengaruh secara signifikan. Hal ini dilakukan dengan menggunakan uji Wald (dinotasikan dengan Z). Hipotesis nol dan alternatif secara berturut-turut dinyatakan dengan $H_0: \beta_p = 0$ dan $H_1: \beta_p \neq 0$. Uji Wald dilakukan dengan menggunakan persamaan $Z = \frac{b_p}{SE(b_p)}$. Selanjutnya, kita dapat memperoleh selang kepercayaan untuk setiap koefisien regresi di bawah taraf signifikansi 5% yang diberikan sebagai $b_p \pm 1.96\, SE(b_p)$.

Klasifikasi merupakan teknik multivariat yang memprediksi atau membedakan kelompok/kelas dari sekumpulan pengamatan/objek serta memprediksi objek baru menggunakan hasil klasifikasi yang diperoleh. Klasifikasi adalah tujuan kedua dalam penyusunan model regresi logistik. Karena tujuan ini sehingga regresi logistik dikenal sebagai sebuah model prediktif.



Gambar 5.2 Proses penyusunan model prediktif (regresi logistik)

Penyusunan model prediktif dimulai dengan membagi suatu *data set* menjadi dua bagian. Bagian pertama dikenal dengan data *training* dan bagian kedua dikenal dengan data *testing*. Proporsi data *training* dan *testing* bisa 70% : 30, 75% : 25% atau 80% : 20%, tidak ada aturan yang baku, tergantung dari peneliti. Data *training* digunakan untuk menyusun model regresi logistik. Setelah model telah disusun, langkah selanjutnya adalah memvalidasi atau menguji kemampuan prediksi model tersebut menggunakan data *testing*. Infor-

masi yang diperoleh dari validasi model adalah prediksi kelas variabel target dari data *testing*. Prediksi kelas variabel target inilah yang akan digunakan dalam tahap selanjutnya, yaitu evaluasi model. Hal yang perlu diperhatikan lebih lanjut yaitu bahwa jika terdapat perlakuan pada data *training*, maka perlakuan tersebut juga diterapkan pada data *testing*, misalnya transformasi variabel kategoris menjadi sebuah variabel *dummy*. Proses penyusunan model prediktif secara lebih jelas disajikan pada Gambar 5.2.

Telah disebutkan sebelumnya bahwa kejadian sukses dan gagal sering juga dikenal dengan istilah positif dan negatif. Kelas positif merupakan kelas yang menjadi pusat perhatian, dinotasikan dengan $P(Y = 1)$. Sebutan ini popular digunakan terutama pada tahap evaluasi model klasifikasi. Ide dasar dari evaluasi model prediktif secara umum yaitu memadankan prediksi kelas variabel target data *testing* dengan kelas variabel target data *testing* yang dikenal dengan kelas aktual. Secara intuitif, model yang baik yaitu model yang memiliki prediksi kelas target yang sama-sama atau mendekati kelas aktual, sehingga semakin banyak pengamatan yang sepadan dengan kelas aktualnya maka dapat dikatakan bahwa kualitas model yang disusun semakin baik.

Proses memadankan prediksi kelas dengan kelas aktual menghasilkan empat kondisi. Prediksi kelas positif bernilai sama dengan kelas aktualnya yang juga positif dikenal dengan *true positive* (TP) dan prediksi kelas negatif bernilai sama dengan kelas aktualnya dikenal dengan *true negative* (TN). Dua kondisi lainnya adalah kesalahan prediksi, *misclassified*, yaitu, prediksi kelas positif sedangkan kelas aktualnya negatif disebut *false positive* (FP) dan prediksi kelas negatif sedangkan kelas aktualnya positif disebut *false negative* (FN).
Evaluasi model regresi logistik dilakukan dengan tabulasi hasil pemadanan hasil prediksi dengan kelas aktual yang dikenal dengan tabel klasifikasi atau *confusion matrix*. Tabel klasifikasi terdiri dari empat bagian sebagaimana yang disajikan pada Gambar 5.3. Dari tabel klasifikasi tersebut digunakan tiga ukuran dasar yang pada gilirannya akan digunakan dalam evaluasi model, yaitu (1) sensitivitas, ukuran yang menyatakan ketepatan prediksi pada kelas positif, dimana semakin tinggi ketepatan prediksi pengamatan yang sesung-

guhnya positif, maka semakin tinggi sensitivitas yang diperoleh; (2) spesifisitas (*specificity*), ukuran ketepatan prediksi pada kelas negatif, dimana semakin tinggi ketepatan prediksi pengamatan yang sesungguhnya negatif, maka semakin tinggi spesifisitas yang diperoleh; dan(3) akurasi, ukuran tingkat ketepatan prediksi secara keseluruhan, baik pada kelas positif maupun negatif. Ketiga ukuran tersebut didefinisikan sebagai berikut.

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

$$spesifisitas = \frac{TN}{TN + FP}$$

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

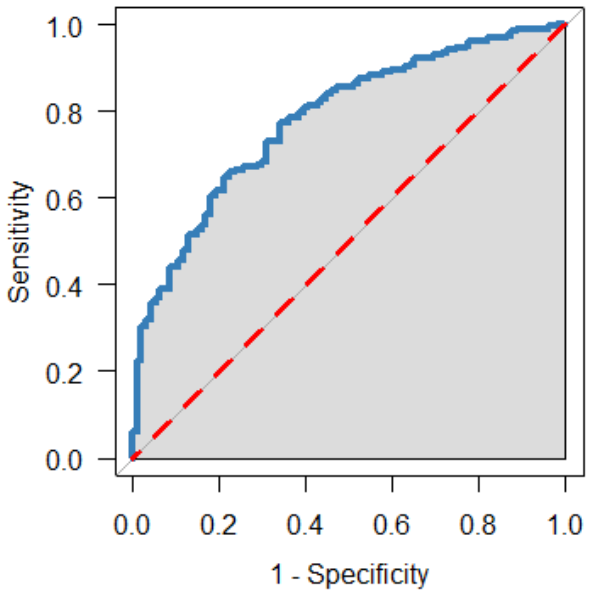
		Prediksi kelas	
		Positif	Negatif
Kelas sebenarnya	Positif	TP	FN
	Negatif	FP	TN

Gambar 5.3 Tabel klasifikasi

Teknik lain yang populer digunakan untuk mengevaluasi model prediktif yaitu dengan menggunakan kurva *receiver operating characteristics* (ROC). Kurva ini mengilustrasikan kemampuan model dalam membedakan kelas positif dan negatif. Kurva dari ROC (lihat Gambar 5.4) dibentuk menggunakan dua unsur, yaitu *true positive rate* (TPR) dan *false positive rate* (FPR) yang diformulasikan sebagai berikut.

$$TPR = \frac{TP}{TP + FN} = sensitivitas$$

$$FPR = \frac{FP}{FP + TN} = 1 - spesifisitas$$



Gambar 5.4 Ilustrasi kurva ROC

Garis diagonal berwarna merah pada Gambar 5.4 adalah *threshold* yang membagi dua secara tepat sementara dan garis biru adalah garis yang membentuk kurva ROC. Semakin dekat jarak kurva ROC dengan *threshold*, kemampuan model semakin buruk karena ini sama saja (proporsi TPR dan FPR adalah 50:50). Sementara itu, semakin jauh kurva ROC dengan *threshold*, yaitu menuju pojok kiri atas, semakin baik kemampuan prediksi dari model yang disusun. Hubungan nilai ROC dengan kemampuan model dikategorikan menurut performanya sebagaimana yang disajikan dalam Tabel 5.1.

Tabel 5.1 Kategori performa model prediktif berdasarkan ROC

ROC	Performa
0.5 – 0.6	*Poor*
0.6 – 0.7	*Average*
0.7 – 0.8	*Good*
0.8 – 0.9	*Very good*
0.9 – 1.0	*Excellent*

Contoh kasus dan analisis regresi logistik menggunakan program R dan RStudio

Pada bagian ini disajikan suatu kasus untuk mendemonstrasikan penerapan analisis logistik dan proses analisis regresi logistik yang dilakukan dengan menggunakan program R dan RStudio. Kasus yang digunakan yaitu sebagai berikut. Suatu penelitian dilakukan untuk memprediksi keberhasilan studi mahasiswa, apakah lulus atau tidak lulus, berdasarkan kriteria penilaian. Penelitian ini melibatkan 395 mahasiswa sebagai sampel dan menggunakan 15 variabel independen (prediktor) berupa informasi latar belakang mahasiswa sebagaimana yang disajikan dalam Tabel 5.2. Variabel respons/dependen berupa data kategoris (lulus/tidak lulus) dan diberi label 1 = lulus dan 0 = tidak lulus. *Data set* yang lengkap digunakan dalam kasus ini dapat diakses secara terbuka di UC Irvine Machine Learning Repository (https://archive.ics.uci.edu/dataset/320/student+performance).

Tabel 5.2 Variabel independen/prediktor dalam kasus penentuan status keberhasilan studi mahasiswa

Kode	Variabel independen	Keterangan
P1	Variabel Independen 1	Umur
P2	Variabel Independen 2	Jenis kelamin
P3	Variabel Independen 3	Daerah tempat tinggal
P4	Variabel Independen 4	Keluarga
P5	Variabel Independen 5	Pendidikan ayah
P6	Variabel Independen 6	Pendidikan ibu
P7	Variabel Independen 7	Jam belajar
P8	Variabel Independen 8	Beasiswa
P9	Variabel Independen 9	Kegiatan ekstrakurikuler
P10	Variabel Independen 10	Akses internet
P11	Variabel Independen 11	Hubungan romantis
P12	Variabel Independen 12	*Hangout*
P13	Variabel Independen 13	Studi lanjut
P14	Variabel Independen 14	Kesehatan
P15	Variabel Independen 15	Ketidakhadiran

Dari kasus yang tersaji, kami akan mendemonstrasikan bagaimana analisis regresi logistik dilakukan pada data untuk kasus tersebut untuk mengetahui variabel independen mana saja yang berpengaruh signifikan terhadap variabel dependen, yaitu lulus atau tidak lulus studi seorang mahasiswa. Selain itu, pada bagian ini didemonstrasikan pengujian pada model regresi logistik yang dibuat, apakah model tersebut memiliki kemampuan klasifikasi dan prediksi yang baik.

Ada empat paket program R yang diperlukan yang digunakan untuk melakukan analisis regresi logistik di RStudio. Keempat paket tersebut yaitu 'car', 'caret', 'performance', dan 'pROC'. Apabila paket tersebut belum terpasang di RStudio, maka langkah awal sebelum melakukan analisis pada data yang dimiliki yaitu memasang paket tersebut dengan menggunakan perintah `install.packages("nama paket")`. Setelah paket-paket tersebut berhasil terpasang di RStudio, analisis regresi logistik pada data yang ada siap untuk dilakukan dengan menggunakan paket tersebut dengan cara menggunakan fungsi `library(nama paket)`. Setelah memanggil paket-paket tersebut, tahap awal yang perlu dilakukan pada data yang akan dianalisis yaitu memanggil data tersebut dan memastikan bahwa data tersebut telah sesuai dengan yang diinginkan termasuk juga memastikan kesesuaian karakteristik atau struktur dari data yang ada tersebut.

```
# Memasang paket-paket yang diperlukan untuk analisis regresi
logistik
> install.packages("car")
> install.packages("caret")
> install.packages("performance")
> install.packages(pROC)
# Memanggil paket-paket untuk digunakan dalam analisis regresi
logistik
> library(car)
> library(caret)
> library(performance)
> library(pROC)

# Impor data
data_set <- read.csv("07-RegLog.csv", stringsAsFactors = T, sep =
",") # nama file dan sep (separator) disesuaikan dengan data_set

#Melihat struktur data
str(data_set)
```

```
> str(data_set)
'data.frame':    395 obs. of 16 variables:
$ kelulusan      : int  0 0 1 1 1 1 1 0 1 1 ...
$ umur           : int  18 17 15 15 16 16 16 17 15 15 ...
$ gender         : Factor w/ 2 levels "F","M": 1 1 1 1 1 2 2 1 2 2 ...
$ daerah         : Factor w/ 2 levels "R","U": 2 2 2 2 2 2 2 2 2 2 ...
$ keluarga       : Factor w/ 2 levels "A","T": 1 2 2 2 2 2 2 1 1 2 ...
$ pend_ayah      : int 4 1 1 2 3 3 2 4 2 4 ...
$ pend_ibu       : int 4 1 1 4 3 4 2 4 3 3 ...
$ jam_belajar    : int 2 2 2 3 2 2 2 2 2 2 ...
$ beasiswa       : Factor w/ 2 levels "no","yes": 1 2 1 2 2 2 1 2 2 2 ...
$ ekskul         : Factor w/ 2 levels "no","yes": 1 1 1 2 1 2 1 1 1 2 ...
$ internet       : Factor w/ 2 levels "no","yes": 1 2 2 2 1 2 2 1 2 2 ...
$ hub_romantis   : Factor w/ 2 levels "no","yes": 1 1 1 2 1 1 1 1 1 1 ...
$ hangout        : int  4 3 2 2 2 2 4 4 2 1 ...
$ studi_lanjut   : Factor w/ 2 levels "no","yes": 2 2 2 2 2 2 2 2 2 2 ...
$ kesehatan      : int  3 3 3 5 5 5 3 1 1 5 ...
$ ketidakhadiran : int  6 4 10 2 4 10 0 6 0 0 ...
```

```
# Mengubah atau menyesuaikan struktur data. Karena struktur data beberapa
variabel independen belum sesuai  dengan yang dikehendaki, misalnya variabel
‘pendidikan ayah’ masih berupa int (integer) seharusnya factor, penyesuaian
struktur data perlu dilakukan.

data_set$pend_ayah <- as.factor(data_set$pend_ayah)
data_set$pend_ibu <- as.factor(data_set$pend_ibu)
data_set$jam_belajar <- as.factor(data_set$jam_belajar)
data_set$hangout <- as.factor(data_set$hangout)
data_set$kesehatan <- as.factor(data_set$kesehatan)


# Mengubah struktur data yang baru
str(data_set)
> str(data_set)
'data.frame': 395 obs. of 16 variables:
$ kelulusan      : int 0 0 1 1 1 1 1 0 1 1 ...
$ umur           : int 18 17 15 15 16 16 16 17 15 15 ...
$ gender         : Factor w/ 2 levels "F","M": 1 1 1 1 1 2 2 1 2 2 ...
$ daerah         : Factor w/ 2 levels "R","U": 2 2 2 2 2 2 2 2 2 2 ...
$ keluarga       : Factor w/ 2 levels "A","T": 1 2 2 2 2 2 2 1 1 2 ...
$ pend_ayah      : Factor w/ 5 levels "0","1","2","3",..: 5 2 2 3 4 4 ...
$ pend_ibu       : Factor w/ 5 levels "0","1","2","3",..: 5 2 2 5 4 5...
$ jam_belajar    : Factor w/ 4 levels "1","2","3","4": 2 2 2 3 2 2  ...
$ beasiswa       : Factor w/ 2 levels "no","yes": 1 2 1 2 2 2 1 2 2 2 ...
$ ekskul         : Factor w/ 2 levels "no","yes": 1 1 1 2 1 2 1 1 1 2 ...
$ internet       : Factor w/ 2 levels "no","yes": 1 2 2 2 1 2 2 1 2 2 ...
$ hub_romantis   : Factor w/ 2 levels "no","yes": 1 1 1 2 1 1 1 1 1 1 ...
$ hangout        : Factor w/ 5 levels "1","2","3",..: 4 3 2 2 4 2 1 ...
$ studi_lanjut   : Factor w/ 2 levels "no","yes": 2 2 2 2 2 2 2 2 2 ...
$ kesehatan      : Factor w/ 5 levels "1","2","3",..: 3 3 5 3 1 1 5 ...
$ ketidakhadiran : int  6 4 10 2 4 10 0 6 0 0 ...
```

```
# Pemisahan data training dan testing
Data set dibagi menjadi data training dan data testing dengan proporsi
masing-masing 70% dan 30%.
#Melihat dimensi data
dim(data_set)
```

```
# Mengunci pengacakan
set.seed (1001)
#Membuat data training dan data testing (70:30)
acak <- sample(1:nrow(data_set), 0.70*nrow(data_set))
train <- data.frame(data_set)[acak,]
data.frame(colnames(train))
test <- data.frame(data_set)[-acak,]
```

Pemisahan data menghasilkan 276 data amatan untuk *training* dan 119 data amatan untuk *testing*.

```
# Membuat model regresi logistik
*** Semua variabel independen
reglog <- glm(kelulusan ~., train, family = "binomial"(link = logit))
summary(reglog)
reglogmodif <- glm(kelulusan ~., train[,], family = "binomial"(link =
logit))
*** Semua variabel independen
Call:
glm(formula = kelulusan ~ ., family = binomial(link = logit),
    data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5645  -0.9760   0.4955   0.8312   2.0571

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)       31.79635 1710.76447   0.019  0.98517
umur              -0.12379    0.13505  -0.917  0.35935
genderM            0.83087    0.35270   2.356  0.01849 *
daerahU            0.26137    0.37317   0.700  0.48367
keluargaT         -0.48360    0.52903  -0.914  0.36065
pend_ayah1       -14.07987  899.18078  -0.016  0.98751
pend_ayah2       -13.46441  899.18082  -0.015  0.98805
pend_ayah3       -13.38515  899.18082  -0.015  0.98812
pend_ayah4       -12.49812  899.18091  -0.014  0.98891
pend_ibu1        -13.71236 1455.39770  -0.009  0.99248
pend_ibu2        -13.89914 1455.39766  -0.010  0.99238
pend_ibu3        -14.03290 1455.39766  -0.010  0.99231
pend_ibu4        -13.65550 1455.39768  -0.009  0.99251
jam_belajar2       0.53977    0.38790   1.391  0.16407
jam_belajar3       1.12978    0.53435   2.114  0.03449 *
jam_belajar4       1.87405    0.78274   2.394  0.01666 *
beasiswayes       -0.43177    0.31236  -1.382  0.16689
ekskulyes         -0.27287    0.30959  -0.881  0.37810
internetyes        0.09315    0.39638   0.235  0.81421
hub_romantisyes   -0.72211    0.32946  -2.192  0.02839 *
hangout2          -1.37687    0.87867  -1.567  0.11712
```

```
hangout3          -1.63987    0.88097  -1.861  0.06268 .
hangout4          -2.46553    0.90783  -2.716  0.00661 **
hangout5          -2.39711    0.91454  -2.621  0.00876 **
studi_lanjutyes    0.57556    0.64738   0.889  0.37397
kesehatan2        -0.40368    0.62519  -0.646  0.51848
kesehatan3        -0.36169    0.56780  -0.637  0.52412
kesehatan4        -0.77937    0.58191  -1.339  0.18046
kesehatan5        -0.68571    0.53735  -1.276  0.20192
ketidakhadiran    -0.03810    0.01913  -1.992  0.04640 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 354.06  on 275  degrees of freedom
Residual deviance: 290.60  on 246  degrees of freedom
AIC: 350.6

Number of Fisher Scoring iterations: 14

*** Tanpa variabel independen
reglog1 <-glm(kelulusan~1 , train, family = "binomial"(link = logit))
summary(reglog1)
*** Tanpa Variabel Independen
Call:
glm(formula = kelulusan ~ 1, family = binomial(link = logit),
    data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4677  -1.4677   0.9126   0.9126   0.9126

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.6607     0.1270   5.202 1.97e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 354.06  on 275  degrees of freedom
Residual deviance: 354.06  on 275  degrees of freedom
AIC: 356.06
Number of Fisher Scoring iterations: 4
```

Dari hasil kedua model tersebut, dengan dan tanpa variabel independen, terlihat bahwa nilai *residual deviance* model tanpa variabel independen lebih tinggi daripada model dengan variabel independen. Dengan demikian, model regresi logistik dengan variabel independen masih lebih baik daripada model tanpa variabel independen.

```
# Uji simultan
anova(reglog1, reglog, test = "LRT")
Analysis of Deviance Table
Model 1: kelulusan ~ 1
Model 2: kelulusan ~ umur + gender + daerah + keluarga + pend_ayah
+ pend_ibu +
    jam_belajar + beasiswa + ekskul + internet + hub_romantis +
    hangout + studi_lanjut + kesehatan + ketidakhadiran
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)

1       275     354.06
2       246     290.60 29   63.468 0.0002244 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Analysis of Deviance Table (Type II tests)
```

```
# Uji parsial
library(car)
Anova(reglog, type = 'II', test = 'Wald')
Response: kelulusan
               Df   Chisq Pr(>Chisq)
umur            1  0.8402    0.35935
gender          1  5.5494    0.01849 *
daerah          1  0.4906    0.48367
keluarga        1  0.8356    0.36065
pend_ayah       4  7.6405    0.10567
pend_ibu        4  0.9720    0.91401
jam_belajar     3  7.7648    0.05113 .
beasiswa        1  1.9107    0.16689
ekskul          1  0.7769    0.37810
internet        1  0.0552    0.81421
hub_romantis    1  4.8041    0.02839 *
hangout         4 13.0273    0.01114 *
studi_lanjut    1  0.7904    0.37397
kesehatan       4  2.5082    0.64317
ketidakhadiran  1  3.9668    0.04640 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Hasil uji simultan dan parsial di atas menunjukkan bahwa variabel gender, hubungan romantis, *hangout*, dan ketidakhadiran signifikan pada taraf signifikasi 5%.

```
# Uji kecocokan model (goodness of fit), hipotesis nol model fit
library(performance)
performance_hosmer(reglog, n_bins = 10)
```

```
***Hosmer-Lemeshow Goodness-of-Fit Test

  Chi-squared: 4.813
           df: 8
      p-value: 0.777
Summary: model seems to fit well.
```

```
# Klasifikasi
# Nilai cut-off yang digunakan adalah 0,50.
pred_1 = predict(reglog, subset(test, select= c(2:16)), type =
'response')
kelulusan = data.frame(y_aktual = test$kelulusan, peluang = pred_1,
y_prediksi = ifelse(pred_1 >= .50, '1', '0'))
cm = table(kelulusan$y_aktual, kelulusan$y_prediksi)
library(caret)
confusionMatrix(cm, positive = '1')
head(kelulusan)
   y_aktual   peluang y_prediksi
1         0 0.8501052      lulus
9         1 0.9495149      lulus
11        0 0.9086130      lulus
12        1 0.6279629      lulus
20        1 0.7861004      lulus
21        1 0.9911283      lulus

Confusion matrix and statistics
        0    1
  0   11   25
  1   22   61

               Accuracy : 0.605
                 95% CI : (0.5113, 0.6934)
    No Information Rate : 0.7227
    P-Value [Acc > NIR] : 0.9980
                  Kappa : 0.0415
 Mcnemar's Test P-Value : 0.7705

            Sensitivity : 0.7093
            Specificity : 0.3333
         Pos Pred Value : 0.7349
         Neg Pred Value : 0.3056
             Prevalence : 0.7227
         Detection Rate : 0.5126
   Detection Prevalence : 0.6975
      Balanced Accuracy : 0.5213

       'Positive' Class : 1
```

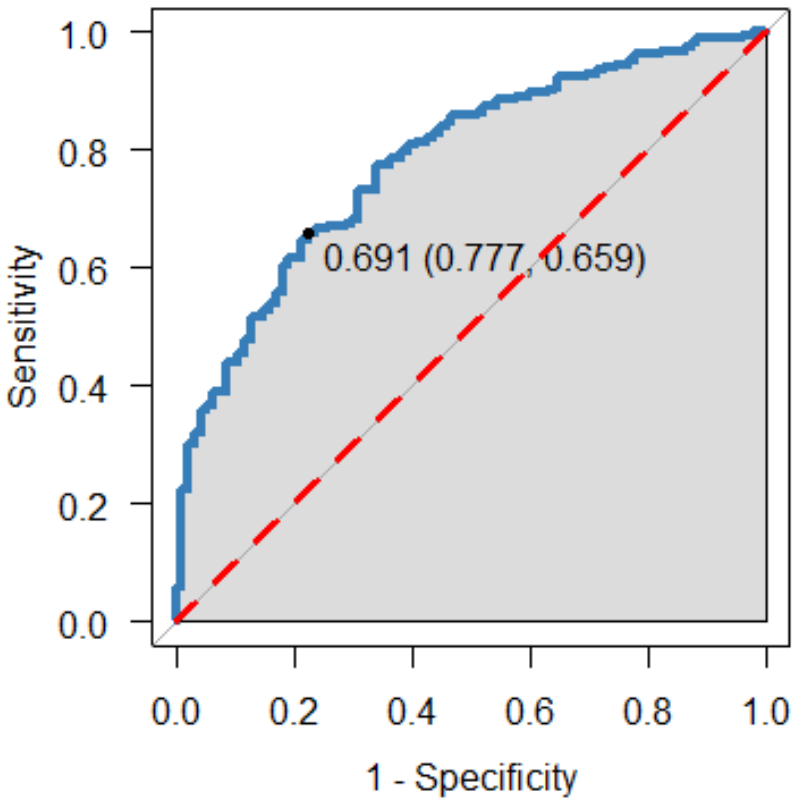
Hasil klasifikasi kelas prediksi dan kelas aktual berdasarkan *confusion matrix* di atas disajikan dalam Tabel 5.4. Tabel 5.4 menunjukkan jumlah mahasiswa yang awalnya diprediksi tidak lulus pada kenyataannya lulus ada sebanyak 22 orang, sebaliknya jumlah mahasiswa yang awalnya diprediksi lulus namun pada kenyataannya tidak lulus ada sebanyak 25 orang. Sementara itu, jumlah mahasiswa yang diprediksi tidak lulus dan sesuai kenyataan adalah sebanyak 11 orang. Sisanya, sebanyak 61 orang mahasiswa diprediksi lulus dan kenyataannya memang lulus sesuai dengan prediksi tersebut.

Tabel 5.4 Tabel klasifikasi status keberhasilan studi mahasiswa

		Kelas prediksi	
		1	0
Kelas aktual	1	61	22
	0	25	11

```
# ROC. Kurva ROC disajikan pada Gambar 5.5.
library(pROC)
roc <- roc(kelulusan~fitted(reglog), data = train)
par(pty = "s") #m = maximum
plot(roc, print.thres="best", bty = 'l', las = 1, col =
"#377eb8",
     lwd = 4, legacy.axes = T, auc.polygon = T)
plot.roc(kelulusan~fitted(reglog1), data = train, add = T, col =
"red",
         lwd = 4,)
auc(roc)
roc.info <- roc(kelulusan~fitted(reglog), data = train)
roc.df <- data.frame(
  tpp = roc.info$sensitivities*100,
  fpp = (1-roc.info$specificities)*100,
  thresholds = roc.info$thresholds)
head(roc.df)
Area under the curve: 0.7804
        tpp       fpp thresholds
1 100.00000 100.00000       -Inf
2 100.00000  98.93617  0.1072279
3 100.00000  97.87234  0.1164497
4  99.45055  97.87234  0.1230960
5  99.45055  96.80851  0.1311481
6  99.45055  95.74468  0.1534965
```

Hasil analisis pada bagian klasifikasi juga menunjukkan nilai akurasi, sensitivitas, dan spesifisitas model secara berturut-turut yaitu sebesar 0,6050; 0.7039; dan 0,333. Hal ini menunjukkan bahwa kemampuan model regresi logistik dalam memprediksi secara tepat pengamatan yang kelas aktualnya positif dan negatif adalah 60,50%, prediksi secara tepat pengamatan yang kelas aktualnya positif sebesar 70,39%, dan prediksi secara tepat pengamatan yang kelas aktualnya positif sebesar 33,33%.



Gambar 5.5 Kurva ROC

Berdasarkan kurva ROC yang disajikan pada Gambar 5, nilai dari ROC yang diperoleh yaitu sebesar 0,691. Selain itu, berdasarkan klasifikasi kemampuan prediksi model berdasarkan nilai ROC pada Ta-

bel 5.1, kemampuan model regresi logistik yang disusun berada pada kategori *average*.

Bab 6
Analisis Klaster

Suatu teknik yang baru dikembangkan untuk menyelesaikan masalah analisis data adalah teknik analisis klaster. Teknik ini akan mencari kategori-kategori atau pola sampel data (*dataset*) berdasarkan proses pembentukan grup-grup data homogen yang disebut klaster. Analisis klaster merupakan suatu kelas teknik untuk mengklasifikasi objek atau kasus (responden) ke dalam kelompok yang relatif homogen, yang disebut klaster (*cluster*). Objek atau kasus dalam setiap kelompok cenderung mirip satu sama lain dan berbeda jauh (tidak sama) dengan objek dari klaster lainnya. Analisis klaster juga disebut analisis klasifikasi atau taksonomi numerik (*numerical taxonomy*). Analisis klaster mengklasifikasi objek sehingga setiap objek yang paling dekat kesamaannya dengan objek lain berada dalam klaster yang sama. Secara logika, klaster-klaster yang baik adalah klaster yang terbentuk memiliki homogenitas internal yang tinggi antar anggota dalam satu klaster (*within-cluster*) dan heterogenitas eksternal yang tinggi antar anggota dalam satu klaster (*between-cluster*).

Berbeda dengan teknik multivariat lainnya, analisis klaster ini tidak mengestimasi himpunan variabel secara empiris karena himpunan variabel itu ditentukan oleh peneliti itu sendiri. Fokus dari analisis klaster adalah membandingkan objek berdasarkan himpunan variabel. Hal inilah yang menyebabkan para ahli mendefinisikan himpunan variabel sebagai tahap yang penting di dalam analisis klaster. Kebanyakan metode pengklasteran merupakan prosedur yang relatif sederhana yang tidak didukung oleh suatu penalaran yang bersifat ekstensif. Sebagian besar metode pengklasteran di dasarkan pada algoritma. Jadi, analisis klaster sangat kontras bila dibandingkan dengan analisis variansi, regresi berganda, analisis diskriminan, dan analisis faktor yang didasarkan pada penalaran statistik yang ekstensif.

Walaupun banyak metode pengklasteran yang mempunyai ciri atau sifat statistik yang penting, ada keperluan untuk memahami aspek kesederhanaan dari metode tersebut. Solusi analisis klaster bersifat tidak unik karena anggota klaster untuk setiap penyelesaian/solusi tergantung pada beberapa elemen prosedur dan beberapa solusi yang berbeda pendapat dengan mengubah satu elemen atau lebih. Solusi klaster secara keseluruhan bergantung pada variabel-variabel yang digunakan sebagai dasar untuk menilai kesamaan. Penambahan atau pengurangan variabel-variabel yang relevan dapat mempengaruhi substansi hasil dari analisis klaster. Secara umum, analisis klaster bisa dikatakan sebagai proses menganalisis baik atau tidaknya suatu proses pembentukan klaster. Analisis klaster bisa diperoleh dari kepadatan klaster yang dibentuk. Kepadatan suatu klaster bisa ditentukan dengan variansi pada *within-cluster* dan pada *between-cluster*.

Penelitian tentang analisis klaster ini telah banyak dilakukan dalam berbagai bidang, salah satunya yang dilakukan oleh Nafkiyah et al. (2022) tentang analisis klaster dalam pengelompokan kabupaten atau kota di Provinsi Jawa Timur berdasarkan indikator pendidikan pada penelitian ini menggunakan metode klaster hierarki (*hierarchical clustering*) yaitu metode *complete linkage*. Contoh lain yaitu penelitian yang dilakukan oleh Wibowo dan Habanabakize (2022) dengan menggunakan *k-means clustering* untuk klasifikasi standar kualifikasi pendidikan dan pengalaman kerja guru SMK di Indonesia. Hasil penelitiannya tersebut menunjukkan bahwa sebuah pembanding dalam pengelolaan guru SMK di Indonesia sangat penting. Klasterisasi memberikan wawasan tentang pengelompokan provinsi-provinsi yang memiliki kinerja yang baik untuk dapat diteladani. Penelitian lain (Matahari et al., 2015) berfokus pada pengelompokan sekolah dasar di Riau berdasarkan Indikator Mutu Sekolah dengan menggunakan analisis klaster hierarki dan non-hierarki. Hasil penelitian mereka menunjukkan terbentuknya tiga kelompok atau klaster.

Konsep dasar dan prosedur pada analisis klaster

Analisis klaster merupakan suatu teknik analisis multivariat yang bertujuan untuk membuat klaster data observasi ataupun variabel-variabel ke dalam klaster sehingga masing-masing klaster bersifat ho-

mogen sesuai dengan faktor yang digunakan untuk melakukan pengklasteran. Karena yang diinginkan yakni mendapatkan klaster yang homogen, kesamaan skor nilai yang dianalisis digunakan sebagai dasar untuk membentuk klaster. Data mengenai ukuran kesamaan tersebut dapat dianalisis dengan menggunakan analisis klaster sehingga dapat ditentukan anggota klaster (Gudono, 2011).

Selanjutnya, berikut dijelaskan mengenai langkah-langkah atau prosedur analisis klaster.

1. *Merumuskan masalah*. Merumuskan masalah merupakan hal paling penting dalam analisis klaster karena menentukan pemilihan variabel-variabel yang akan dipergunakan untuk pembentukan klaster. Memasukkan satu atau dua variabel yang tidak relevan dengan masalah pengklasteran akan mendistorsi hasil pengklasteran (Supranto, 2004).
2. *Memilih ukuran jarak*. Tujuan analisis klaster adalah mengelompokkan objek yang mirip ke dalam klaster yang sama. Oleh karena itu memerlukan beberapa ukuran untuk mengetahui seberapa mirip atau berbeda objek-objek tersebut. Pendekatan yang biasa digunakan yaitu mengukur kemiripan yang dinyatakan dalam jarak antara pasangan objek. Pada analisis klaster, ada tiga ukuran untuk mengukur kesamaan antar objek, yaitu ukuran asosiasi, ukuran korelasi, dan ukuran kedekatan.
   - *Ukuran asosiasi*. Ukuran asosiasi biasanya dipakai untuk mengukur data berskala non-metrik (yaitu nominal atau ordinal) dengan cara mengambil bentuk-bentuk dari koefisien korelasi pada tiap objek dengan menentukan nilai mutlak dari korelasi-korelasi yang bernilai negatif (Simamora, 2005).
   - *Ukuran korelasi*. Ukuran korelasi biasanya dipakai untuk mengukur data skala matriks. Akan tetapi ukuran ini jarang digunakan karena titik beratnya pada nilai suatu pola tertentu, padahal titik berat analisis klaster terletak pada besarnya objek. Kesamaan antar objek dapat diketahui dari koefisien korelasi antar pasangan objek yang diukur dengan menggunakan beberapa variabel.
   - *Ukuran kedekatan*. Ada empat ukuran kedekatan. *Pertama*, jarak Euclides, di mana ini mengukur jumlah kuadrat perbedaan

nilai pada masing-masing variabel. Jarak Euclides diberikan sebagai $d_{ij} = \sqrt{\sum_{k=1}^{p}\left(X_{ik} - X_{jk}\right)^2}$ dengan d_{ij} merupakan jarak antara objek ke-i dan obyek ke-j, p jumlah variabel klaster, X_{ik} data dari subjek ke-i pada variabel ke-k, dan X_{jk} data dari subjek ke-j pada variabel ke-k. *Kedua*, kuadrat jarak Euclides yang merupakan variasi dari jarak Euclides yang diberikan sebagai $d_{ij} = \sum_{k=1}^{p}\left(X_{ik} - X_{jk}\right)^2$. *Ketiga*, jarak City-Block atau jarak Manhattan yang merepresentasikan jumlah perbedaan mutlak di dalam nilai untuk setiap variabel yang diberikan sebagai berikut dengan N adalah banyaknya objek, N_j jumlah objek dalam klaster j, N_{jkl} jumlah objek di klaster j untuk variabel kategori ke-k dengan kategori ke-l, $\hat{\sigma}_k^2$ adalah ragam dugaan untuk variabel kontinu ke-k untuk keseluruhan objek, $\hat{\sigma}_{jk}^2$ adalah ragam dugaan untuk variabel kontinu ke-k untuk keseluruhan objek dengan klaster j, K^A adalah banyaknya variabel kontinu, K^B adalah banyaknya variabel kategori, dan L_K adalah banyaknya kategori untuk variabel kategori ke-k.

$$\xi_j = -N\left(\sum_{K=1}^{K^A}\frac{1}{2}log\left(\hat{\sigma}_k^2 + \hat{\sigma}_{jk}^2\right) - \sum_{k=1}^{K^B}\sum_{l=1}^{L_K}\frac{N_{jkl}}{N_j}log\left(\frac{N_{jkl}}{N_j}\right)\right)$$

$$\xi_s = -N\left(\sum_{K=1}^{K^A}\frac{1}{2}log\left(\hat{\sigma}_k^2 + \hat{\sigma}_{sk}^2\right) - \sum_{k=1}^{K^B}\sum_{l=1}^{L_K}\frac{N_{skl}}{N_j}log\left(\frac{N_{skl}}{N_j}\right)\right)$$

$$\xi_{(jS)} = -N\left(\sum_{K=1}^{K^A}\frac{1}{2}log\left(\hat{\sigma}_k^2 + \hat{\sigma}_{(js)k}^2\right) - \sum_{k=1}^{K^B}\sum_{l=1}^{L_K}\frac{N_{(js)kl}}{N_j}log\left(\frac{N_{(js)kl}}{N_j}\right)\right)$$

Asumsi yang ada pada jarak Manhattan yaitu variabel kontinu normal, variabel kategori menyebar multinomial dan antar variabel bersifat saling bebas. Metode klaster dua tahap cukup tegar terhadap pelanggaran asumsi tersebut sehingga metode ini masih dapat digunakan ketika terjadi pelanggaran asumsi. Jarak Euclides dan Manhattan dapat digunakan jika antar variabel memiliki satuan yang sama dan korelasi antar variabel tidak nyata. Namun demikian, jika satuan antar variabel tidak

sama, jarak Euclides dan Manhattan dapat digunakan dengan ditransformasi ke dalam bentuk baku. Jika ada korelasi antar variabel yang nyata, jarak yang dapat digunakan yaitu jarak Mahalanobis. Apabila hendak menggunakan jarak Euclides, maka variabel asal ditransformasi menggunakan analisis komponen utama. *Keempat*, jarak Chebyshev yang merepresentasikan nilai maksimum perbedaan mutlak di setiap variabel.

3. *Standardisasi data*. Proses standardisasi dilakukan apabila di antara variabel-variabel yang diteliti terdapat perbedaan ukuran satuan yang besar. Perbedaan satuan yang mencolok dapat mengakibatkan perhitungan pada analisis klaster menjadi tidak valid. Oleh karena itu, proses standarisasi perlu dilakukan dengan cara transformasi (standarisasi pada data asli sebelum dianalisis lebih lanjut). Transformasi dilakukan terhadap variabel yang relevan ke dalam bentuk *z-score*.
4. *Memilih prosedur pengklasteran*. Proses pembentukan klaster dapat dilakukan dengan dua cara, yaitu dengan hierarki dan non-hierarki. Cara hierarki terdiri atas metode *agglomerative* (penggabungan) dan metode *divisive* (pemecahan). Metode *agglomerative* dapat dilakukan melalui teknik *linkage*, *variance*, dan *centroid*. Teknik *linkage* terdiri atas *single linkage*, *complete linkage*, dan *average linkage*. Sementara itu, teknik *variance* terdiri atas *Ward*. Selanjutnya, metode non-hierarki terdiri atas tiga metode, yaitu *sequential thereshold*, *parallel*, dan *optimizing partitioning* (Gudono, 2011). Penjelasan dari masing-masing metode-metode atau teknik-teknik tersebut disajikan pada bagian setelah penjelasan semua langkah dalam analisis klaster.
5. *Menentukan banyaknya klaster*. Masalah utama dalam analisis klaster ialah menentukan berapa banyaknya klaster. Pada dasarnya tidak ada aturan yang baku untuk menentukan berapa banyaknya klaster. Meskipun demikian, beberapa petunjuk bisa digunakan, seperti hasil kajian teoretis, kajian secara konseptual, atau berdasarkan praktik yang ada sebelumnya. Sebagai contoh, apabila tujuan dari pembentukan klaster adalah untuk mengidentifikasi segmen pasar, manajemen mungkin menghendaki klaster da-

lam jumlah tertentu, misal tiga, empat, atau lima; dan jumlah klaster tentu seharusnya berguna (Supranto, 2004).

6. *Menginterpretasikan profil klaster*. Tahap menginterpretasikan profil klaster ini meliputi pengujian pada masing-masing klaster yang terbentuk untuk memberikan nama atau keterangan secara tepat sebagai gambaran sifat dari klaster tersebut dan menjelaskan bagaimana mereka bisa berbeda secara relevan pada tiap dimensi. Rata-rata (*centroid*) setiap klaster pada setiap variabel digunakan ketika memulai proses interpretasi.
7. *Melakukan validasi dan pembuatan profil* (*profiling*). Validasi ini bertujuan untuk menjamin bahwa solusi yang dihasilkan dari analisis klaster dapat mewakili populasi dan dapat diperumum untuk objek lain. Validasi ini dilakukan dengan membandingkan solusi klaster dan melalui korespondensi hasil. Sementara itu, pembuatan profil digunakan untuk menjelaskan karakteristik setiap klaster berdasarkan profil tertentu. Titik beratnya pada karakteristik yang secara signifikan berbeda antar klaster dan memprediksi anggota dalam suatu klaster. Hasil analisis klaster dapat digunakan untuk berbagai kepentingan sesuai dengan materi yang dianalisis.

Metode hierarki dan non-hierarki pada analisis klaster

Metode hierarki pada analisis klaster

Metode hierarki (*hierarchical method*) adalah suatu metode pada analisis klaster yang membentuk tingkatan tertentu karena proses pembentukan klaster dilakukan secara bertahap atau bertingkat. Hasil pembentukan klaster dengan metode hierarki dapat disajikan dalam bentuk dendrogram. Dendrogram adalah representasi visual dari langkah-langkah dalam analisis klaster yang menunjukkan bagaimana klaster terbentuk dan nilai koefisien jarak pada setiap langkah. Angka di sebelah kanan pada dendrogram merepresentasikan objek penelitian. Objek-objek tersebut dihubungkan oleh garis dengan objek yang lain sehingga pada akhirnya akan membentuk suatu klaster (Simamora, 2005).

Penggunaan metode hierarki dalam analisis klaster membawa keuntungan dalam hal mempercepat pengolahan dan menghemat wak-

tu karena data yang dimasukkan akan membentuk hierarki atau membentuk tingkatan sendiri. Ini akan mempermudah dalam penafsiran. Namun demikian, metode ini memiliki kelemahan berupa seringnya timbul kesalahan pada data pencilan (*outlier*), adanya perbedaan ukuran jarak yang digunakan, dan terdapatnya variabel yang tidak relevan. Metode hierarki ini dapat dilakukan melalui *agglomerative* dan *divisive* yang dijelaskan sebagai berikut.

- Metode *agglomerative*

Metode *agglomerative* menganggap setiap objek merupakan suatu klaster. Dua objek dengan jarak terdekat kemudian digabung menjadi satu klaster. Setelah itu, objek ketiga akan bergabung dengan klaster yang ada atau bersama objek lain dan membentuk klaster baru dengan tetap memperhitungkan jarak kedekatan antar objek. Proses akan berlanjut hingga akhirnya terbentuk satu klaster yang terdiri atas keseluruhan objek.

Metode *agglomerative* terdiri atas *linkage*, *variance*, dan *centroid*. Metode *linkage* dibagi lagi menjadi tiga macam, yaitu *single linkage*, *complete linkage*, dan *average linkage*. Metode *single linkage* yaitu proses pembentukan klaster yang didasarkan pada jarak terdekat antar objek. Metode ini sangat bagus untuk analisis pada tiap tahap pembentukan klaster. Metode ini juga sangat cocok dipakai pada kasus *shape independent clustering* karena kemampuannya untuk membentuk pola tertentu dari klaster. Metode ini dikenal juga dengan sebutan metode tetangga terdekat (*nearest neighbors*) dalam klaster yang berbeda. Beda dengan *complete linkage*, pada prosedur *single linkage* jelas dilakukan berdasarkan jarak minimum. Jika individu X dan Y mempunyai jarak d_{XY} terdekat, maka harus dicari jarak minimum XZ dan XY sehingga $d_{(XY)Z} = Min(d_{XZ}, d_{YZ})$. Hasil dari *single linkage* ini dapat ditampilkan dalam bentuk dendrogram atau diagram pohon dengan dahan atau cabang dan diagram pohon tersebut merupakan klasternya.

Jenis metode *linkage* berikutnya yaitu *complete linkage*. *Complete linkage* (*farthest neighbors*) ini merupakan proses pembentukan klaster yang didasarkan pada jarak terjauh antar objek atau pada kesamaan minimum. Metode ini baik untuk kasus klasterisasi pada himpunan data yang mengikuti pola distribusi normal. Namun de-

mikian, metode ini tidak cocok untuk data yang mengandung pencilan (*outlier*).

Jenis metode *linkage* terakhir yaitu *average linkage* yang menggunakan jarak rata-rata antar objeknya sebagai dasar dalam pembentukan klaster. Meskipun metode *average linkage* ini dianggap yang paling baik dibanding dua jenis *linkage* lainnya, jenis ini memerlukan waktu komputasi yang paling lama. Konsep dasar pada *average linkage* ini yaitu bahwa jarak rata-rata antara observasi pengelompokan dimulai dari tengah atau pasangan observasi dengan jarak paling mendekati jarak rata-rata. Selanjutnya, jarak antara klaster ditentukan, termasuk rata-rata jarak seluruh objek suatu klaster lainnya. Cara ini bertujuan untuk meminimalkan rataan jarak semua pasangan pengamatan dan membentuk klaster dengan variansi kecil. Pada berbagai keadaan, *average linkage* ini dianggap lebih stabil. Pada metode ini, tahap pertama yang harus dilakukan adalah sama seperti metode-metode sebelumnya, yaitu menemukan jarak terkecil. Jika kelompok X dan Y mempunyai jarak d_{XY} maka harus dicari jarak rata-rata XZ dan XY menggunakan persamaan $d_{(XY)Z} = \frac{n_X}{n_X+n_Y} + d_{XZ} + \frac{n_Y}{n_X+n_Y} + d_{YZ}$ dengan n_X menyatakan jumlah individu pada kelompok X dan n_Y menyatakan jumlah individu pada kelompok Y.

Jenis metode *agglomerative* berikutnya yaitu *variance* yang mana melalui jenis ini, jarak antara dua klaster, dilakukan melalui Ward, yang ditentukan berdasarkan jumlah kuadrat antara dua klaster untuk seluruh variabel. Metode ini cenderung digunakan untuk mengombinasikan klaster dengan jumlah kecil. Untuk membentuk klaster, metode *variance* meminimalisasi sebuah fungsi objektif berupa ukuran '*squared error*' seperti yang digunakan dalam MANOVA. Prosedur pembentukan klaster yang digunakan dalam metode ini, sekali lagi, didasarkan pada minimum variansi suatu klaster. Ukuran jarak berdasarkan metode ini yaitu $d_{(XY)Z} = \frac{(n_X+n_Z)d_{XZ}+(n_Y+n_Z)d_{YZ}-n_Z d_{XY}}{n_X+n_Y+n_Z}$, dengan n_X menyatakan jumlah individu pada kelompok X, n_Y menyatakan jumlah individu pada kelompok Y, dan n_Z menyatakan jumlah individu pada kelompok Z.

Jenis metode *agglomerative* yang terakhir yaitu *centroid.* Ini merujuk pada rata-rata semua objek di dalam klaster. Pada metode ini, jarak antar klaster adalah jarak antar *centroid*. *Centroid* baru dihitung ketika setiap kali objek digabungkan, sehingga setiap kali anggotanya bertambah, *centroid* akan berubah (Johnson & Wichern, 1992).

- Metode *divisive*

Metode ini dimulai dengan satu klaster besar yang mencakup semua objek pengamatan. Selanjutnya, secara bertahap, objek yang mempunyai ketidakmiripan cukup besar akan dipisahkan ke dalam klaster-klaster yang berbeda. Proses dilakukan terus sehingga terbentuk sejumlah klaster yang diinginkan, misal dua klaster, tiga klaster, dan seterusnya.

Metode non-hierarki pada analisis klaster

Metode non-hierarki sering disebut sebagai metode *k-means*. Prosedur pada metode non-hierarki dimulai dengan memilih sejumlah nilai klaster awal sesuai dengan jumlah yang diinginkan, kemudian objek pengamatan digabungkan ke dalam klaster-klaster tersebut. Dua kelemahan dari prosedur non-hierarki ialah bahwa banyaknya klaster telah ditentukan sebelumnya dan pemilihan pusat klaster ditentukan secara sembarang. Hasil klaster mungkin tergantung pada bagaimana pusat dipilih. Banyak program non-hierarki berfokus pada pemilihan objek atau kasus yang pertama tanpa ada nilai yang hilang sebagai pusat klaster awal. Jadi, hasil klaster mungkin tergantung pada urutan observasi dalam data.

Bagaimanapun juga, klaster non-hierarki lebih cepat (prosesnya) daripada metode hierarki dan lebih menguntungkan apabila jumlah objek atau kasus atau observasi besar sekali (ukuran sampel besar). Tantangan utama pada metode non-hierarki yaitu bagaimana memilih bakal klaster karena ini akan berpengaruh pada hasil akhir analisis klaster. Bakal klaster pertama adalah observasi pertama dalam himpunan data tanpa titik data hilang (*missing value*). Bakal kedua yaitu observasi lengkap berikutnya (tanpa *missing value*) yang dipisahkan dari bakal pertama oleh jarak minimum khusus.

Terdapat dua asumsi dalam analisis klaster menggunakan *k-means* menurut Hair et al. (2010), yaitu sampel bersifat representatif

dan tidak terjadi multikolinearitas. Sampel yang bersifat representatif yaitu sampel yang merepresentasikan atau mewakili populasi yang ada. Tidak ada ketentuan untuk ukuran sampel yang representatif, namun tetaplah diperlukan ukuran sampel yang cukup besar agar proses pembentukan klaster dapat dilakukan dengan benar. Uji Kaiser-Meyer-Olkin (KMO) dapat digunakan untuk mengetahui seberapa jauh sampel bersifat representatif atau untuk mengetahui kecukupan ukuran sampel untuk setiap indikator. Ukuran KMO memiliki nilai 0 sampai dengan 1. Jika nilai KMO berkisar 0,5 sampai 1, maka dapat dikatakan bahwa sampel yang ada bersifat mewakili populasi atau bersifat representatif.

Asumsi yang kedua yaitu tidak adanya multikolinearitas. Multikolinearitas merujuk pada adanya hubungan linear yang sempurna atau pasti antara beberapa atau semua variabel (Gujarati dan Porter, 2009). *Variance inflation factor* (VIF) yang didefinisikan dengan persamaan $VIF = \frac{1}{1-R_j^2}$ dengan R_j^2 menyatakan koefisien determinasi dapat digunakan untuk mengetahui ada atau tidaknya multikolinearitas. Multikolinearitas diindikasikan oleh VIF yang bernilai lebih dari 10.

Metode non-hierarki ini dapat dibagi menjadi tiga jenis, yaitu *sequential threshold, parallel threshold,* dan *optimizing partitioning* (Gudono, 2011). Berikut adalah penjelasan dari ketiga jenis metode non-hierarki tersebut.

- Melalui *sequential threshold*, analisis klaster dimulai dengan memilih satu klaster dan menempatkan semua objek yang berada pada jarak terdekat ke dalam klaster tersebut. Apabila semua objek yang berada pada ambang batas tertentu telah dimasukkan, klaster yang kedua dipilih dan menempatkan semua objek yang berada pada jarak terdekat ke dalamnya. Kemudian klaster ketiga dipilih dan proses dilanjutkan seperti yang sebelumnya.
- Melalui *parallel threshold*, analisis klaster dilakukan melalui pemilihan terhadap beberapa objek awal klaster sekaligus dan kemudian melakukan penggabungan objek ke dalam klaster tersebut secara bersamaan. Saat proses berlangsung, jarak terdekat dapat ditentukan untuk memasukkan beberapa objek ke dalam klaster-klaster.

- Melalui *optimizing partitioning*, analisis klaster dilakukan secara hampir mirip dengan yang dilakukan pada *sequential threshold* dan *parallel threshold*. Pembedanya yaitu *optimizing partitioning* ini memungkinkan untuk menempatkan kembali objek-objek ke dalam klaster yang lebih dekat atau dengan melakukan optimasi pada penempatan objek yang ditukar untuk klaster lainnya dengan pertimbangan kriteria optimasi.

Contoh kasus dan analisis klaster menggunakan program R dan RStudio

Pada bagian ini disajikan suatu kasus yaitu mengenai tingkat kesehatan kabupaten/kota di pulau Jawa. Kasus ini berfokus pada penyelidikan terhadap pembentukan klaster berdasarkan tingkat kesehatan melalui analisis klaster. Sejumlah paket diperlukan untuk melakukan analisis klaster ini. Paket-paket tersebut meliputi 'magrittr' yang digunakan untuk mengurangi waktu pengembangan dan meningkatkan keterbacaan, serta memelihara kode; 'knitr' yang digunakan untuk mengintegrasikan komputasi dan pelaporan; 'ggplot2' yang digunakan untuk memvisualisasikan data atau hasil analisis secara lebih menarik, 'factoextra' yang digunakan untuk mengekstraksi dan memvisualisasikan hasil dari analisis multivariat, yaitu memvisualisasikan klaster dan menentukan jumlah klaster optimum; dan paket yang terakhir yaitu 'cluster' yang digunakan oleh semua *notebook* yang sedang berjalan dan untuk melakukan analisis klaster. Berikut perintah yang digunakan pada analisis klaster dan hasilnya.

```
#Pastikan paket-paket yang diperlukan untuk analisis klaster telah
terpasang
install.packages("magrittr")
install.packages("knitr")
install.packages("ggplot2")
install.packages("factoextra")
install.packages("cluster")
```

```
#Library yang dibutuhkan dalam analisis klaster
library(magrittr)
library(knitr)
library(ggplot2)
library(factoextra)
library(cluster)
```

```
#Import data
Data <- read.csv ("KesehatanJawa.csv", sep = ';')
Summary (Data)

*** Deskripsi Data
      AHH         Rasio.Puskesmas Rasio.Rumah.Sakit Persentase.PHBS Persentase.Sanitasi.Layak
 Min.   :-3.01567 Min.   :-0.5905 Min.   :-0.6242   Min.   :-2.3785 Min.   :-3.9233
 1st Qu.:-0.53077 1st Qu.:-0.4948 1st Qu.:-0.5873   1st Qu.:-0.7378 1st Qu.:-0.3256
 Median : 0.07812 Median :-0.4689 Median :-0.5597   Median : 0.1086 Median : 0.1117
 Mean   : 0.00000 Mean   : 0.0000 Mean   : 0.0000   Mean   : 0.0000 Mean   : 0.0000
 3rd Qu.: 0.61021 3rd Qu.:-0.4430 3rd Qu.: 0.3680   3rd Qu.: 0.6796 3rd Qu.: 0.6640
 Max.   : 1.91026 Max.   : 4.0397 Max.   : 4.2686   Max.   : 2.2534 Max.   : 1.1907
 Persentase.BBLR   Persentase.Asi.Eksklusif Angka.Kesakitan.Diare
 Min.   :-2.0969   Min.   :-2.9227          Min.   :-2.14783
 1st Qu.:-0.5970   1st Qu.:-0.5649          1st Qu.:-0.10167
 Median : 0.1500   Median : 0.1708          Median :-0.09205
 Mean   : 0.0000   Mean   : 0.0000          Mean   : 0.00000
 3rd Qu.: 0.6878   3rd Qu.: 0.7273          3rd Qu.: 0.45125
 Max.   : 2.3610   Max.   : 2.0586          Max.   : 7.53376
```

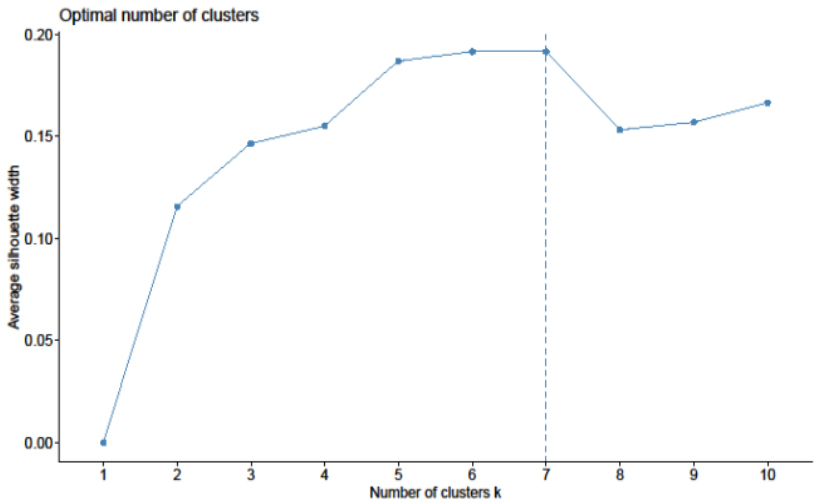
```
#Standardisasi data
Datastand <- scale (Data [2:9])
Datastand %>% head (10) %>% kable (caption = "Hasil Standarisasi
Data Indikator Kesehatan di Pulau jawa")

*** Ringkasan Standarisasi Data
 AHH Rasio.Puskesmas Rasio.Rumah.Sakit Persentase.PHBS
Persentase.Sanitasi.Layak Persentase.BBLR
1 -0.349746950 1.4460149 -0.5820674 -1.06469184 -0.195166781 1.40489149
2 0.001321423 1.0817705 -0.5706045 0.11466181 0.556595341 0.50853687
3 0.323134099 -0.4472397 -0.6129613 -1.46799413 -0.307675398 0.26950898
4 0.462098663 -0.4505211 0.3553041 -1.31522811 0.357148247 -0.32806077
5 0.264622703 -0.4987678 -0.5703171 -0.85081942 0.587279509 0.03048108
6 -0.005992501 -0.4851243 -0.5744254 -0.52084482 -0.261649145 -0.44757472
7 -0.053533010 -0.5254424 -0.5548391 -0.70416404 -0.005947744 -0.62684564
8 -1.011657111 -0.4831618 -0.5792249 -1.91407090 -0.491780407 0.68780780
9 -1.362725484 -0.4998354 -0.5826265 0.28575975 -0.808850145 1.52440544
10 -0.759326718 -0.4646670 -0.5564234 -1.00969608 0.357148247 -0.38781774
```

```
    Persentase.Asi.Eksklusif Angka.Kesakitan.Diare
1                0.503635252            0.45118192
2                1.152894118            0.45124998
3                0.449075683            0.45129225
4                0.279941020            0.45123432
5                1.425691962            0.45129220
6                0.803712879            0.45129122
7                0.438163769            0.45126191
8                1.141982205            0.45124470
9                1.185629860            0.45126558
10               0.923743930            0.45125214
```

```
#Uji representatif data. Uji representatif data menggunakan uji
KMO. Sampel cukup untuk digunakan jika KMO > 0,05. Berdasarkan
hasil uji, KMO = 0,5848527. Ini menunjukkan bahwa data layak untuk
digunakan.

kmo <- function(x)
{
  x <- subset (x, complete.cases(x))
  r <- cor(x)
  r2 <- r^2
  i <- solve(r)
  d <- diag(i)
  p2 <- (-i/sqrt(outer(d, d)))^2
  diag(r2) <- diag(p2) <- 0
  KMO <- sum(r2)/(sum(r2)+sum(p2))
  MSA <- colSums(r2)/(colSums(r2)+colSums(p2))
  return(list(KMO=KMO, MSA=MSA))
}
Kmo (Data[,1:8])

*** Uji KMO
$KMO
[1] 0.5848527

$MSA
AHH                          Rasio.Puskesmas         Rasio.Rumah.Sakit           Persentase.PHBS
              0.6141406                0.4269940                0.4915706                0.6405596
Persentase.Sanitasi.Layak          Persentase.BBLR  Persentase.Asi.Eksklusif    Angka.Kesakitan.Diare
              0.6934455                0.4993522                0.5521318                0.6323414
```

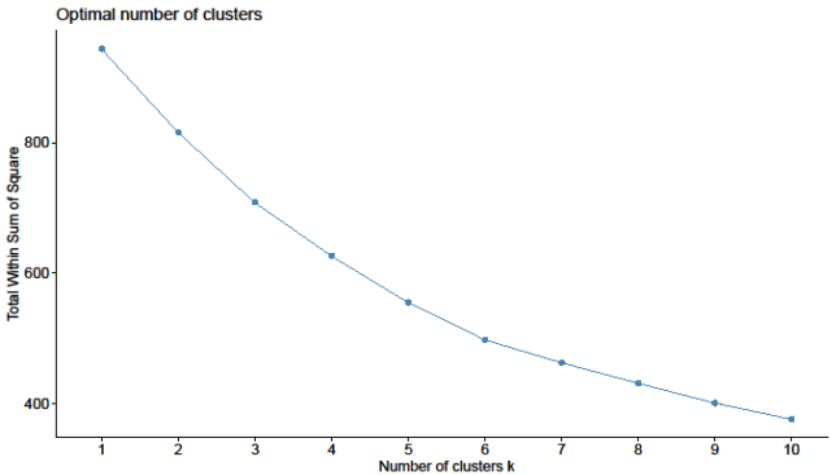
```
#Uji multikolinearitas. Jika korelasi antara variabel lebih dari 0,5, maka terjadi multikolinearitas. Berdasarkan hasil uji
tersebut, didapatkan bahwa tidak terjadi multikolinearitas antar variabel karena koefisien korelasi berada di bawah 0,5.

korelasi = cor (Data, method="pearson")
korelasi

*** Uji korelasi
                              AHH Rasio.Puskesmas Rasio.Rumah.Sakit Persentase.PHBS Persentase.Sanitasi.Layak
AHH                   1.00000000    -0.082462636       0.244143356     0.479176509               0.411990182
Rasio.Puskesmas      -0.08246264     1.000000000       0.271961945    -0.084158634               0.007330331
Rasio.Rumah.Sakit     0.24414336     0.271961945       1.000000000     0.006917752               0.255832965
Persentase.PHBS       0.47917651    -0.084158634       0.006917752     1.000000000               0.248423337
Persentase.Sanitasi.Layak  0.41199018     0.007330331       0.255832965     0.248423337               1.000000000
Persentase.BBLR       0.04458615     0.091049131       0.063197056    -0.028817665               0.173676213
Persentase.Asi.Eksklusif  -0.18227972     0.036627743       0.203892535    -0.356893209              -0.033334935
Angka.Kesakitan.Diare -0.07713513    -0.081691116       0.005372617    -0.092905158              -0.079614593
                      Persentase.BBLR Persentase.Asi.Eksklusif Angka.Kesakitan.Diare
AHH                        0.04458615             -0.18227972          -0.077135131
Rasio.Puskesmas            0.09104913              0.03662774          -0.081691116
Rasio.Rumah.Sakit          0.06319706              0.20389253           0.005372617
Persentase.PHBS           -0.02881767             -0.35689321          -0.092905158
Persentase.Sanitasi.Layak  0.17367621             -0.03333494          -0.079614593
Persentase.BBLR            1.00000000              0.22088176          -0.074378717
Persentase.Asi.Eksklusif   0.22088176              1.00000000           0.082992161
Angka.Kesakitan.Diare     -0.07437872              0.08299216           1.000000000
```

```
#Menentukan jumlah klaster
fviz_nbclust (Datastand, hcut, method = "silhouette") # Menampilkan
grafik jumlah klaster dengan metode “silhouette” sebagaimana yang
tersaji pada Gambar 6.1
```



Gambar 6.1 Grafik banyaknya klaster melalui metode *silhouette*

```
fviz_nbclust (Datastand, hcut, method = "wss") # Menampilkan grafik
jumlah klaster dengan metode “wss” sebagaimana yang tersaji pada
Gambar 6.2. Penentuan klaster menggunakan grafik ‘wss’ dilihat dari
banyaknya klaster yang garisnya berbentuk seperti siku (elbow).
Gambar 6.2 menunjukkan bahwa garis membentuk siku saat berada di
titik yang menunjukkan bahwa banyaknya klaster yaitu 6.
```



Gambar 6.2 Grafik banyaknya klaster melalui metode wss

```
#Klasterisasi untuk kasus bahwa banyaknya klaster yaitu 7. Klaster
dan anggotanya disajikan secara lengkap dalam Tabel 6.1. Hasil
analisis menunjukkan centroid masing-masing variabel pada tiap
klaster.
clus_hier = eclust (Datastand, FUNcluster = "kmeans", k = 7,
hc_method = "complete", graph= F)
clus_hier$cluster
clus_hier$centers

*** Klaster 7
[1] 6 6 6 6 6 6 6 5 5 6 5 5 5 6 4 6 6 5 2 6 7 6 6 6 7 4 6 6 5 1 1 1
6 6 1 1 4 1 3 3 7 3 3 3 6 6 7 7 3 7 3 7 3 3 3
[56] 7 7 7 7 3 3 3 3 3 3 3 6 1 7 7 3 3 3 4 4 2 2 6 2 6 6 4 5 6 3 4
2 2 4 4 4 2 2 6 2 4 4 6 2 2 5 5 5 5 4 4 4 4 1 6
[111] 5 1 1 4 4 4 4 4 2
```

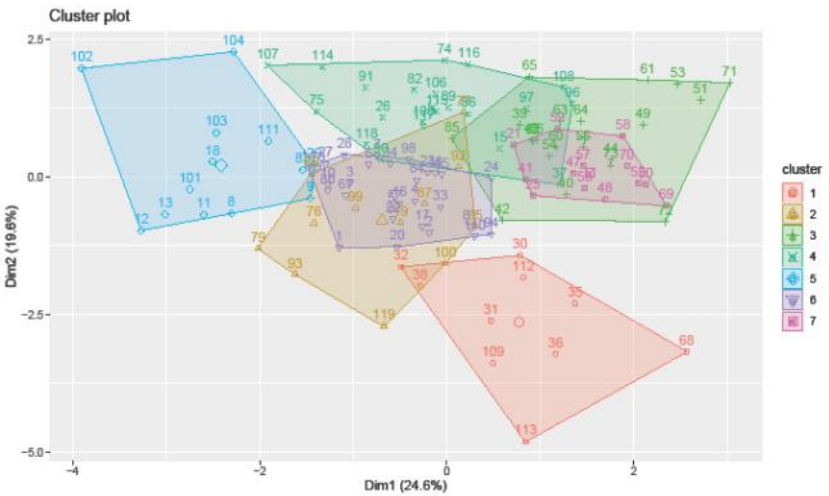
Tabel 6.1 Hasil klaster dan anggotanya yang terbentuk

Kota/Kab	Klaster	Kota/Kab	Klaster
Kab. Pacitan	6	Kab. Gresik	7
Kab. Ponorogo	6	Kab. Bangkalan	4
Kab. Trenggalek	6	Kab. Sampang	6
Kab. Tulungagung	6	Kab. Pamekasan	6
Kab. Blitar	6	Kab. Sumenep	5
Kab. Kediri	6	Kota Kediri	1
Kab. Malang	6	Kota Blitar	1
Kab. Lumajang	5	Kota Malang	1
Kab. Jember	5	Kota Probolinggo	6
Kab. Banyuwangi	6	Kota Pasuruan	6
Kab. Bondowoso	5	Kota Mojokerto	1
Kab. Situbondo	5	Kota Madiun	1
Kab. Probolinggo	5	Kota Surabaya	4
Kab. Pasuruan	6	Kota Batu	1
Kab. Sidoarjo	4	Kab. Cilacap	3
Kab. Mojokerto	6	Kab. Banyumas	3
Kab. Jombang	6	Kab. Purbalingga	7
Kab. Nganjuk	5	Kab. Banjarnegara	3
Kab. Madiun	2	Kab. Kebumen	3
Kab. Magetan	6	Kab. Purworejo	3
Kab. Ngawi	7	Kab. Wonosobo	6
Kab. Bojonegoro	6	Kab. Magelang	6
Kab. Tuban	6	Kab. Boyolali	7

Kota/Kab	Klaster	Kota/Kab	Klaster
Kab. Lamongan	6	Kab. Klaten	7
Kab. Sukoharjo	3	Kab. Pangandaran	4
Kab. Wonogiri	7	Kota Bogor	2
Kab. Karanganyar	3	Kota Sukabumi	2
Kab. Sragen	7	Kota Bandung	6
Kab. Grobogan	3	Kota Cirebon	2
Kab. Blora	3	Kota Bekasi	4
Kab. Rembang	3	Kota Depok	4
Kab. Pati	7	Kota Cimahi	6
Kab. Kudus	7	Kota Tasikmalaya	2
Kab. Jepara	7	Kota Banjar	2
Kab. Demak	7	Kab. Lebak	5
Kab. Semarang	3	Kab. Pandeglang	5
Kab. Temanggung	3	Kab. Serang	5
Kab. Kendal	3	Kab. Tangerang	5
Kab. Batang	3	Kota Tangerang	4
Kab. Pekalongan	3	Kota Cilegon	4
Kab. Pemalang	3	Kota Serang	4
Kab. Tegal	3	Kota Tangerang Selatan	4
Kab. Brebes	6	Kab. Kulon Progo	1
Kota Magelang	1	Kab. Bantul	6
Kota Surakarta	7	Kab. Gunung Kidul	5
Kota Salatiga	7	Kab. Sleman	1
Kota Semarang	3	Kota Yogyakarta	1
Kota Pekalongan	3	Kota Jakarta Pusat	4
Kota Tegal	3	Kota Jakarta Utara	4
Kab. Bogor	4	Kota Jakarta Barat	4
Kab. Sukabumi	4	Kota Jakarta Selatan	4
Kab. Cianjur	2	Kota Jakarta Timur	4
Kab. Bandung	2	Kep. Seribu	2
Kab. Garut	6	Kab. Indramayu	3
Kab. Tasikmalaya	2	Kab. Subang	4
Kab. Ciamis	6	Kab. Purwakarta	2
Kab. Kuningan	6	Kab. Karawang	2
Kab. Cirebon	4	Kab. Bekasi	4
Kab. Majalengka	5	Kab. Bandung Barat	4
Kab. Sumedang	6		

```
Aggregate (Datastand, by =list (cluster=clus_hier$cluster), FUN = mean)
*** Klaster 7 data
  Klaster        AHH Rasio.Puskesmas Rasio.Rumah.Sakit Persentase.PHBS Persentase.Sanitasi.Layak Persentase.BBLR
1       1  0.5579111       0.4469285        2.55861115     -0.57354910                0.8844045       0.6400022
2       2 -0.4969397       2.2987839        0.18426407     -0.39613683               -0.4010064      -0.4331334
3       3  0.7219171      -0.1471994       -0.28316518      0.87884107                0.4723356       0.5341470
4       4 -0.1075167      -0.5169703       -0.09491494      0.09286719               -0.2656917      -1.4992975
5       5 -1.4412095      -0.4875566       -0.59718635     -0.98102307               -1.6837423       0.5958740
6       6 -0.2732029      -0.2470298       -0.37517016     -0.43868831                0.2495773       0.3086601
7       7  1.0877205      -0.3542330       -0.09340313      1.19680927                0.4830320       0.1821719
  Persentase.Asi.Eksklusif Angka.Kesakitan.Diare
1              0.86449224          -0.002428008
2             -0.03150318          -0.198049591
3             -1.30062370          -0.193573135
4             -0.52203268           0.385779119
5              0.57666114           0.076519218
6              0.63070261           0.262724714
7              0.32476380          -0.788401309
```

```
fviz_cluster(clus_hier) #Hasil visualisasi klaster dan anggotanya
tersaji pada Gambar 6.3.
```

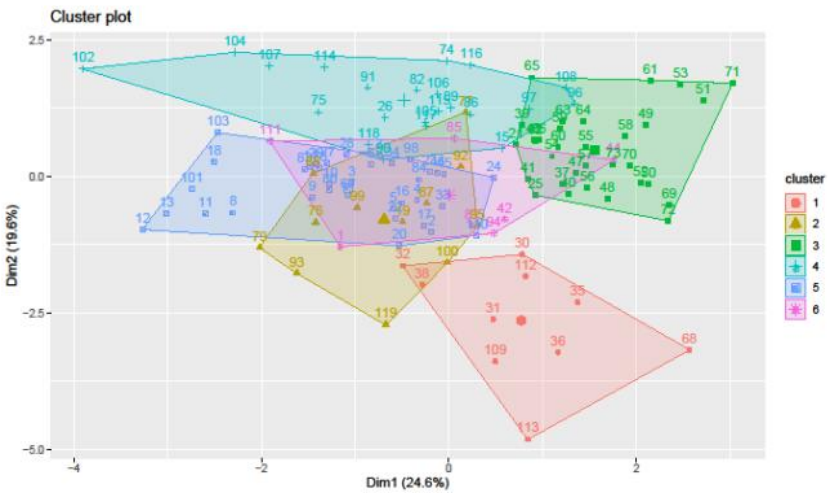


Gambar 6.3 Visualisasi klaster yang terbentuk dan anggotanya dengan banyaknya klaster yaitu tujuh

```
#Klasterisasi untuk kasus bahwa banyaknya klaster yaitu 6.
kmeans <- eclust (Datastand, stand = TRUE, FUNcluster = "kmeans",
k=6, hc_method = "complete", graph = F)
kmeans$cluster
kmeans$centers

*** Klaster 6
[1] 6 5 5 5 5 5 5 5 5 5 5 5 5 5 4 5 5 5 2 5 3 5 5 5 3 4 5 5 5 1 1 1
5 5 1 1 3 1 3 3 3 6 3 6 5 5 3 3 3 3 3 3 3 3 3
[56] 3 3 3 3 3 3 3 3 3 3 3 5 1 3 3 3 3 3 4 4 2 2 5 2 5 6 4 5 5 6 4
2 2 4 4 4 2 2 6 2 4 4 5 2 2 5 4 5 4 4 4 4 4 1 5
[111] 6 1 1 4 4 4 4 4 2
```

```
Aggregate (Datastand, by =list (cluster=kmeans$cluster), FUN = mean)

*** Klaster 6 data
  Klaster        AHH Rasio.Puskesmas Rasio.Rumah.Sakit Persentase.PHBS Persentase.Sanitasi.Layak Persentase.BBLR
1       1  0.5579111       0.4469285         2.5586111     -0.57354910               0.88440454       0.6400022
2       2 -0.4969397       2.2987839         0.1842641     -0.39613683              -0.40100641      -0.4331334
3       3  0.9099623      -0.3640616        -0.1675062      1.04674457               0.53773736       0.2788460
4       4 -0.3126787      -0.5162656        -0.1909334      0.02053017              -0.64092406      -1.4976677
5       5 -0.6570334      -0.3872772        -0.4462902     -0.64120747              -0.09757408       0.3608044
6       6  0.2567864       0.6993004        -0.3100144     -0.05355540              -0.51808112       1.4048915
  Persentase.Asi.Eksklusif Angka.Kesakitan.Diare
1              0.86449224          -0.002428008
2             -0.03150318          -0.198049591
3             -0.56508440          -0.424040627
4             -0.54534685           0.256894830
5              0.69965566           0.289919061
6             -0.48202233          -0.016942461
```

```
fviz_cluster(kmeans) #Hasil visualisasi klaster dan anggotanya
tersaji pada Gambar 6.4.
```

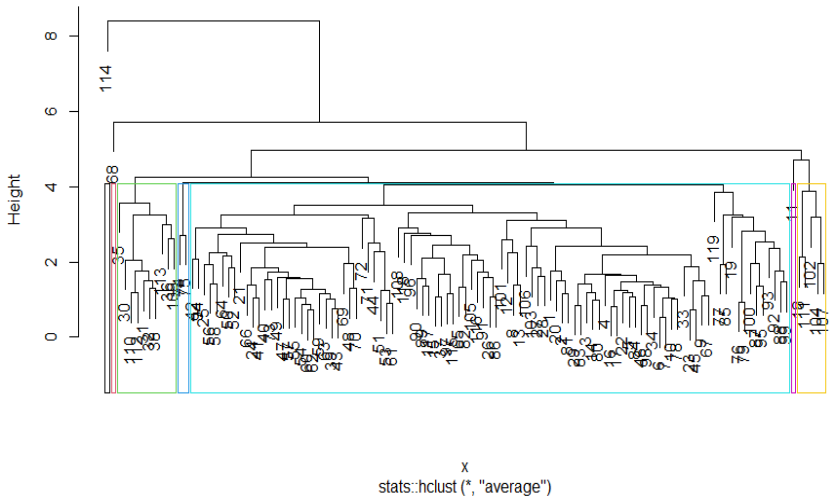


Gambar 6.4 Visualisasi klaster yang terbentuk dan anggotanya dengan banyaknya klaster yaitu enam

```
#Klaster dendrogram. Hasil visualisasi dalam bentuk dendrogram
disajikan pada Gambar 6.5.
clus_hierD <- eclust (Datastand, FUNcluster = "hclust", k = 7,
hc_method = "average", graph = FALSE)
plot(clus_hierD)
rect.hclust (clus_hierD , k = 7, border = 1:8)
```

Gambar 6.5 menunjukkan pembentukan klaster dengan menggunakan metode non-hierarki. Metode ini membebaskan peneliti menentukan jumlah klaster. Pada data kasus yang didemonstrasikan pada analisis klaster ini, klaster yang dibentuk sebanyak tujuh klaster.

Karena data kasus (kabupaten/kota) yang digunakan cukup banyak, klaster tidak terlihat dengan jelas ketika menggunakan representasi berupa dendrogram. Metode non-hierarki melalui penggunaan dendrogram, dengan demikian, tidak disarankan untuk jumlah data atau kasus yang banyak.



Gambar 6.5 Dendrogram yang menyajikan klaster yang terbentuk dan anggotanya untuk kasus banyaknya klaster yaitu tujuh

Halaman ini sengaja dibiarkan kosong

Bab 7
Penskalaan Multidimensi

Banyak masalah dalam kehidupan sehari-hari yang dijadikan penelitian untuk mencari solusinya. Dalam penelitian yang mempunyai banyak unit pengamatan dan variabel sehingga membutuhkan interpretasi hubungan suatu variabel dengan variabel lainya untuk pengambilan keputusan, maka diperlukan teknik analisis yang dapat memudahkan menganalisis variabel-variabel (data) penelitian di antaranya analisis *multidimensional scaling* (MDS).

Multidimensional scaling (MDS) merupakan satu dari banyak teknik analisis multivariat yang dapat membantu kita dalam menginterpretasikan atau menemukan hubungan antara beberapa variabel dengan hanya melihat perkiraan jarak antar variabel tersebut atau dengan melihat peta spasial. MDS dapat pula membantu kita mengidentifikasi atau mengenali dimensi kunci yang mendasari evaluasi objek dari responden tanpa mendeskripsikan sifat atau atribut-atribut terlebih dahulu. MDS menggunakan matriks jarak untuk mereduksi data dan memetakan data dengan melihat kemiripan, baik dari unit pengamatan maupun variabel pengamatan. Hasil *plot* MDS dapat digunakan untuk menentukan banyak pengelompokan objek sebagai salah alternatif dalam melakukan analisis kluster.

Analisis MDS telah banyak digunakan dalam penelitian di berbagai bidang, seperti penelitian yang dilakukan oleh Almeira dan Juanda (2021) yang berfokus pada pengelompokan provinsi berdasarkan tingkat pengangguran. Contoh lainnya yaitu penelitian yang Putri et al. (2018) lakukan, di mana dalam penelitian mereka tersebut, mereka berfokus untuk menyelidiki posisi dari lima merek telepon genggam berdasarkan persepsi yang konsumer miliki pada merek telepon genggam tersebut dan mengeksplorasi keunggulan dari setiap merek telepon genggam tersebut berdasarkan atribut dari produk telepon

genggam pada suatu merek dan persepsi konsumer dengan menggunakan MDS.

Teori dasar pada penskalaan multidimensi

Konsep dasar pada penskalaan multidimensi

Penskalaan multidimensi (*multidimensional scaling*, MDS) adalah teknik analisis untuk mengidentifikasi kesamaan (*similarity*) atau ketidaksamaan (*dissimilarity*) data sejumlah objek (Borg & Groenen, 2005). Penskalaan multidimensi menampilkan kesamaan atau ketidaksamaan ini dalam bentuk titik-titik di ruang geometris. Titik-titik tersebut merupakan data objek-objek yang diteliti, misalnya kesamaan-kesamaan kampus, jenis kejahatan, kota, produk tertentu dan lain sebagainya seperti korelasi antar tes inteligensi. Kedekatan titik-titik itu menunjukkan kesamaan, sedangkan kejauhannya menunjukkan ketidaksamaannya. Kedekatan dan kejauhan dalam ruang geometris lebih mudah dimengerti karena terlihat jelas oleh mata.

Setidaknya terdapat dua pengertian tentang penskalaan multidimensi, yakni pengertian dalam arti luas dan pengertian dalam arti sempit (de Leeuw & Heiser, 1980). Dalam arti luas, penskalaan multidimensi mencakup bermacam-macam bentuk analisis klaster dan analisis multivariat linier. Sementara itu, penskalaan multidimensi dalam arti sempit merepresentasikan kesamaan atau ketidaksamaan data dalam ruang berdimensi rendah atau ruang multidimensi (de Leeuw & Heiser, 1980). Dalam tulisannya itu, de Leeuw dan Hesier kemudian tidak menggunakan dua pengertian tersebut, tetapi mengklasifikasikan penskalaan multidimensi menjadi penskalaan multidimensi metrik dan penskalaan multidimensi non-metrik, penskalaan multidimensi dua arah dan penskalaan multidimensi tiga arah. Ini sejalan dengan pendapat Kruskal (1964).

Penjelasan pada penskalaan multidimensi selanjutnya menyangkut tujuan dari penskalaan multidimensi tersebut. Borg dan Groenen dalam dua bukunya menyebut tujuan-tujuan penskalaan multidimensi secara berbeda. Pada tahun 2005, Borg dan Groenen (2005) mengemukakan bahwa penskalaan multidimensi memiliki empat tujuan. Berikut empat tujuan dari penskalaan multidimensi tersebut.

- *Mengeksplorasi*. Penskalaan multidimensi itu merupakan metode atau teknik yang digunakan untuk merepresentasikan atau menggambarkan kesamaan dan ketidaksamaan data dalam wujud jarak pada ruang dua dimensi sehingga kesamaan dan ketidaksamaan data terlihat secara visual.
- *Menguji hipotesis struktural*. Penskalaan multidimensi merupakan teknik untuk menguji apakah dan bagaimanakah kriteria-kriteria bisa membedakan objek-objek yang berbeda.
- *Mengeksplorasi struktur psikologis*. Penskalaan multidimensi dapat menjadi pendekatan analisis data untuk menemukan dimensi-dimensi yang sebenarnya menentukan kesamaan dan ketidaksamaan data.
- *Model penentu kesamaan*. Penskalaan multidimensi merupakan model psikologis yang menjelaskan penentuan kesamaan dan ketidaksamaan dalam wujud fungsi jarak.

Agak berbeda dengan Borg dan Groenen (2005) dalam hal mengungkap tujuan dari penskalaan multidimensi, Borg et al. (2018) menyebutkan empat tujuan penskalaan multidimensi. *Pertama* yaitu untuk memvisualisasikan data yang berdekatan. Penskalaan multidimensi memvisualisasikan data-data dalam bentuk titik-titik yang berdekatan atau berjauhan di ruang geometris. *Kedua* yaitu menemukan dimensi-dimensi laten (dimensi yang tidak dapat diamati atau diukur secara langsung). Penskalaan multidimensi menemukan dimensi-dimensi laten dengan melihat kedekatan titik-titik datanya. *Ketiga* yaitu membuat model penilaian berdasar jarak. Penskalaan multidimensi dengan jarak antar titik digunakan untuk menilai, misalnya, status keberlanjutan sistem lampu lalu lintas (Mahida, 2020). *Keempat* yaitu untuk menguji hipotesis struktural. Penskalaan multidimensi merupakan teknik untuk menguji apakah dan bagaimanakah kriteria-kriteria bisa membedakan objek-objek yang berbeda. Visualisasi titik-titik data pada ruang geometris dengan penskalaan multidimensi menunjukkan hipotesis struktural terbukti atau tidak.

Tujuan dari penskalaan multidimensi juga diajukan oleh Gudono (2017). Gudono (2017) menyebutkan bahwa penskalaan multidimensi memiliki dua tujuan. Tujuan yang pertama yaitu untuk memperoleh ukuran pembanding antara sejumlah objek ketika dasar pem-

bandingnya belum diketahui atau belum ada. Tujuan yang kedua yaitu untuk mengidentifikasi dimensi yang belum diketahui. Semua tujuan dari penskalaan multidimensi yang telah disebutkan menekankan bahwa penskalaan multidimensi itu memvisualisasikan data dalam bentuk titik-titik di ruang geometris. Kedekatan jarak antar titik memperlihatkan kesamaan titik atau semacamnya.

Teknik-teknik pada penskalaan multidimensi

Teknik-teknik penskalaan multidimensi dapat dibedakan menjadi metrik dan non-metrik. Pertama, penskalaan multidimensi metrik memakai nilai metrik dan bertujuan untuk memperoleh konfigurasi titik-titik data dalam ruang multidimensi, dan kedekatan jarak antar titik itu menunjukkan kemiripan atau kesamaan (Gudono, 2017). Jika kedekatan jarak itu diukur dengan menggunakan *Euclidean distance* (jarak Euclides), teknik penskalaan multidimensi metrik tersebut dinamakan *classical metric multidimensional scaling*. Data penskalaan multidimensi metrik adalah interval atau rasio.

Tabel 7.1 Kriteria nilai stres pada penskalaan multidimensi

Stress	Kriteria
≥ 20%	Kurang (*Poor*)
10% – 20%	Cukup (*Fair*)
5% – 10%	Baik (*Good*)
2,5% – 5%	Sangat Baik (*Excellent*)
< 2,5%	Sempurna (*Perfect*)

Sementara itu, teknik penskalaan multidimensi non-metrik bertujuan untuk menentukan hubungan non-monoton antara kesamaan atau ketidaksamaan data dan jarak di dalam konfigurasi (Kruskal, 1964). Hubungan yang bersifat monoton menjadi tujuan penskalaan multidimensi non-metrik dan sekaligus menjadi hipotesis dalam penskalaan multidimensi non-metrik. Data pada teknik penskalaan multidimensi non-metrik ialah nominal atau ordinal. Dalam penskalaan multidimensi non-metrik, peneliti perlu menghitung nilai stres untuk menentukan kelayakan konfigurasinya. Makin kecil nilai stresnya, makin layak atau bagus konfigurasinya (Kruskal, 1964). Kruskal ke-

mudian membuat tabel nilai stres pada penskalaan multidimensi (lihat Tabel 7.1).

Asumsi-asumsi pada penskalaan multidimensi

Untuk melakukan analisis penskalaan multidimensi beberapa asumsi atau syarat seharusnya terpenuhi. Asumsi-asumsi tersebut antara lain yaitu: (1) semua variabel yang relevan sudah dimasukkan, artinya modelnya sudah dispesifikasi secara tepat; (2) spesifikasi skalanya (ordinal, interval, atau rasio) sesuai dengan teknik penskalaan multidimensi yang digunakan (misalnya, penskalaan multidimensi metrik menggunakan skala rasio atau interval); (3) jumlah objek minimal sebanyak jumlah dimensi; (4) skala yang digunakan setara, di mana jika tidak setara, skalanya distandardisasi (misalnya, sebuah variabel memakai satuan rupiah, maka variabel-variabel lainnya juga memakai satuan rupiah, bukan, misalnya, kilogram, kilometer dan seterusnya); (5) objek-objek yang dibandingkan memiliki kesamaan tertentu sehingga perbandingannya pantas; dan (6) jumlah variabel minimal (Gudono, 2017).

Persamaan penskalaan multidimensi

Untuk membahas persamaan penskalaan multidimensi, kita perlu membicarakan lebih dahulu data multivariat, matriks jarak, dan konfigurasi objek. Data multivariat (Tabel 7.2) diubah jadi matriks jarak (Tabel 7.3) dan matriks jarak direpresentasikan dalam konfigurasi objek di ruang geometris. Konfigurasi objek merupakan hasil analisis penskalaan multidimensi. Penskalaan multidimensi dikategorikan menjadi metrik jika $d_{AB} = 2d_{BC}$ dan dikategorikan menjadi nonmetrik apabila $d_{AB} > 2d_{BC}$.

Tabel 7.2 Contoh struktur data multivariat

	X_1	X_2	...	X_p
A	X_{11}	X_{12}	...	X_{1p}
B	X_{21}	X_{22}	...	X_{2p}
C	X_{31}	X_{32}	...	X_{3p}
D	X_{41}	X_{42}	...	X_{4p}
E	X_{51}	X_{52}	...	X_{5p}

Tabel 7.3 Matriks jarak dari contoh struktur data multivariat

	A	B	C	D	E
A	0	d_{AB}	d_{AC}	d_{AD}	d_{AE}
B	d_{BA}	0	d_{BC}	d_{BD}	d_{BE}
C	d_{CA}	d_{CB}	0	d_{CD}	d_{CE}
D	d_{DA}	d_{DB}	d_{DC}	0	d_{DE}
E	d_{EA}	d_{EB}	d_{EC}	d_{ED}	0

Untuk menghasilkan matriks, kita perlu mengetahui persamaan *Euclidean distance* (jarak Euclides). Kita misalkan terlebih dahulu bahwa objek 1 direpresentasikan oleh $\boldsymbol{x}' = \left(x_1, x_2, \ldots, x_p\right)'$ dan objek 2 direpresentasikan oleh $\boldsymbol{y}' = \left(y_1, y_2, \ldots, y_p\right)'$. Dengan demikian, jarak Euclides antara objek 1 dan objek 2 dinyatakan sebagai berikut.

$$d(x,y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_p - y_p)^2}$$

dengan x_1 merupakan objek 1 pada pengamatan ke-1, y_1 merupakan objek ke-2 pada pengamatan ke-1, dan p merupakan banyaknya pengamatan. Jika persamaan tersebut dibentuk dalam notasi matriks, maka menjadi persamaan berikut.

$$d(x,y) = \sqrt{(\boldsymbol{x} - \boldsymbol{y})'(\boldsymbol{x} - \boldsymbol{y})}$$

Dari persamaan tersebut, kita dapat memperoleh matriks jarak yang menunjukkan jarak antar pasangan objek yang terjadi sebagai berikut.

$$nDn = \begin{bmatrix} d_{11} & d_{12} & \cdots & d_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ d_{n1} & d_{n2} & \cdots & d_{nn} \end{bmatrix}$$

Langkah-langkah penskalaan multidimensi

Langkah-langkah analisis penskalaan multidimensi metrik dan non-metrik sedikit berbeda. Analisis penskalaan multidimensi metrik tidak perlu menghitung nilai stres (S), sedangkan analisis penskalaan multidimensi non-metrik perlu menghitung nilai stres (S). Nilai stres

(S) merupakan ukuran konfigurasi objek. Berikut ini adalah langkah-langkah untuk melakukan penskalaan multidimensi metrik dan non-metrik.

- Langkah analisis penskalaan multidimensi metrik
  1. Menghitung matriks jarak D
  $$d^2{}_{i.} = \frac{1}{n}\sum_j d^2{}_{ij}\,;\; d^2{}_{.j} = \frac{1}{n}\sum_i d^2{}_{ij}\;;\; d^2{}_{..} = \frac{1}{n^2}\sum_{ij} d^2{}_{ij}$$
  2. Menghitung matriks B
  $$b_{ij} = -\frac{1}{2}\left(d^2{}_{ij} - d^2{}_{i.} - d^2{}_{.j} + d^2{}_{..}\right)$$
  3. Menentukan *eigenvalue* dan *eigenvector*
  $$Det(\mathrm{B} - \lambda \mathrm{I}) = 0; Det(\mathrm{B} - \lambda \mathrm{I})\mathrm{E} = 0$$
  4. Menentukan koordinat objek
  $$\mathrm{F} = \tilde{\mathrm{E}}\Lambda^{1/2}; \tilde{\mathrm{e}}_i = \frac{\mathrm{e}_i}{\sqrt{e_i' e_i}}$$
- Langkah analisis penskalaan multidimensi non-metrik

Langkah analisis penskalaan multidimensi non-metrik sama dengan langkah 1 sampai langkah 4 pada analisis penskalaan multidimensi metrik dan ditambah dengan langkah 5. Langkah 5 yang dimaksud yaitu menghitung nilai stres (S) yang diberikan sebagai berikut.

$$S = \frac{\sum_{i=j}^{n}\left(d_{ij} - \hat{d}_{ij}\right)^2}{\sum_{i\neq j}^{n} d_{ij}^2}$$

Contoh kasus dan analisis penskalaan multidimensi menggunakan program R dan RStudio

Terdapat beberapa fungsi atau paket di RStudio di bawah program R yang dapat digunakan untuk melakukan penskalaan multidimensi. Adapun fungsi-fungsi dan paket yang bersesuaian dengan fungsi tersebut yaitu sebagai berikut. Fungsi pertama yang digunakan yaitu `cmdscale()`, di mana fungsi ini tercakup dalam paket "stats". Fungsi ini akan digunakan dalam perhitungan penskalaan multidimensi klasik melalui teknik metrik. Fungsi kedua yang dapat digunakan yaitu `isoMDS()` yang tercakup dalam paket "MASS". Fungsi ini digunakan untuk perhitungan penskalaan multidimensional non-metrik Kruskal. Semua fungsi tersebut mengambil objek jarak sebagai argumen utama dan *k* adalah jumlah dimensi yang diinginkan dalam luaran (*out-*

put) yang diskalakan. Secara pengaturan dasar, mereka mengembalikan solusi dua dimensi. Akan tetapi kita dapat mengubahnya melalui parameter *k* yang secara pengaturan dasar bernilai 2. Selain itu, Fungsi `cmdscale()` yang tercakup dalam paket "stats" dan `isoMDS()` di dalam "MASS" dapat digunakan untuk meminimalkan nilai stres.

Contoh kasus

Pada contoh kasus ini, kami menyajikan dua kasus sesuai dengan teknik-teknik yang ada pada penskalaan multidimensi, yaitu metrik dan non-metrik. *Pertama*, untuk kasus metrik, contoh kasus yang digunakan yaitu berfokus pada analisis kemiripan pada sekolah dasar berstatus negeri (SDN) di suatu kecamatan, misalkan Kecamatan X, berdasarkan capaian nilai rata-rata siswa pada mata pelajaran yang diujikan pada Ujian Sekolah (US). Adapun data terkait capaian tersebut disajikan pada Tabel 7.4.

Tabel 7.4 Data capaian nilai rata-rata siswa pada US

Nama Sekolah	PKn	B.INA	MAT	IPA	IPS	SBK	PJOK
SDN A	81,07	81,11	55,38	75,26	70,15	60,74	64,93
SDN B	75,19	78,06	45,66	68,85	63,70	65,98	65,18
SDN C	79,88	78,95	63,24	73,05	79,74	75,57	69,83
SDN D	79,21	77,50	50,17	66,17	61,90	60,63	68,82
SDN E	81,44	85,25	76,88	68,21	71,27	65,19	71,31
SDN F	79,43	79,31	52,46	66,33	67,23	66,84	67,08
SDN G	78,70	84,10	53,95	68,83	68,99	71,88	65,24
SDN H	85,25	85,50	67,25	72,72	74,30	60,38	66,13
SDN I	80,61	84,33	59,55	70,40	73,17	71,76	67,92
SDN J	80,17	83,75	59,00	70,78	81,34	64,75	66,03
SDN K	81,77	85,11	60,78	68,84	69,11	63,28	65,69
SDN L	81,34	77,89	55,90	65,65	68,98	71,19	66,03
SDN M	74,78	77,44	44,56	66,22	65,56	60,61	65,11
SDN N	64,71	71,50	47,67	55,84	57,40	54,04	52,86
SDN O	80,50	86,50	64,75	64,32	77,68	62,88	67,88
SDN P	75,81	78,17	51,03	61,10	69,70	62,93	65,45
SDN Q	79,61	77,55	52,20	66,66	65,23	62,24	66,23
SDN R	85,46	85,67	59,17	71,57	68,86	61,75	64,63
SDN S	80,13	80,75	61,69	71,03	70,23	65,25	65,25
SDN T	79,00	70,00	63,33	63,49	59,97	71,50	63,58
SDN U	76,63	77,12	57,77	60,72	75,92	59,70	59,90
SDN V	65,35	69,50	44,65	60,53	59,75	54,98	57,55

Kedua, untuk kasus non-metrik, kasus yang digunakan yaitu berfokus pada analisis kemiripan jenis telepon genggam (*handphone*, HP) dari berbagai merek berdasarkan preferensi sejumlah responden terhadap atribut produk HP tersebut. Tabel 7.5 menyajikan data terkait dengan hal itu.

Tabel 7.5 Data preferensi responden terhadap atribut *smartphone*

HP	Merek	Desain	Fitur	Layar	Harga	Kemudahan	Kamera	*Processor*	Memori	Pemakaian
Asus	4	5	4	5	2	4	5	5	5	5
Oppo	2	3	2	2	3	2	1	3	2	2
Samsung	1	1	1	1	5	1	3	1	3	1
Sony	5	4	3	3	4	5	2	2	4	2
Xiaomi	4	4	3	4	1	3	3	4	1	5

Prosedur analisis

Pada bagian ini, kita terlebih dahulu melakukan analisis penskalaan multidimensi pada data metrik sebagai berikut. Ini dilakukan dengan cara menyajikan perintah-perintah yang digunakan dan luaran (*output*) dari perintah-perintah yang bersangkutan. Perintah awal yang jalankan yaitu membaca data yang akan dianalisis dan memeriksa apakah data yang terbaca memang yang diinginkan atau sudah sesuai. Setelah dipastikan bahwa data yang terbaca atau data yang digunakan sudah sesuai dengan yang diinginkan, perintah selanjutnya ditujukan untuk melakukan analisis yang berkaitan dengan penskalaan multidimensi di bawah teknik metrik. Hal yang perlu diperhatikan dalam analisis ini yaitu memastikan bahwa folder direktori untuk kerja (analisis) sudah sesuai dengan yang diinginkan.

```
data <- read.csv('DATA.csv', header = T, sep = ';')
data1<-data[,-1]
rownames(data1) <- c("SDN A","SDN B","SDN C","SDN D","SDN E",
                     "SDN F","SDN G","SDN H","SDN I","SDN J",
                     "SDN K","SDN L","SDN M","SDN N","SDN O",
                     "SDN P","SDN Q","SDN R","SDN S","SDN T",
                     "SDN U","SDN V")
head(data1)
print(head(data1,n=22))
```

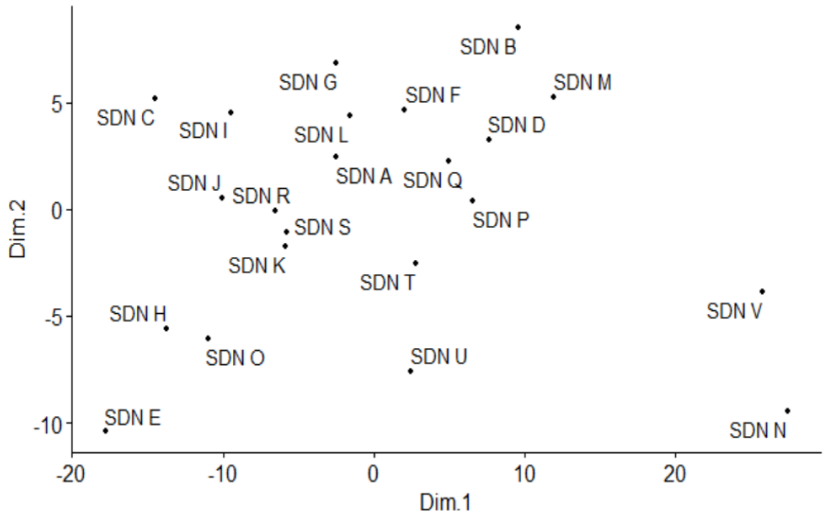
```
#Load required packages
library(magrittr)
library(dplyr)
library(ggpubr)
```

```
#Compute MDS
mds <- data1 %>%
  dist() %>%
  cmdscale() %>%
  as_tibble()
colnames(mds) <- c("Dim.1", "Dim.2")
print(mds,n=22)
#           A           tibble:           22           ×           2
    Dim.1                                                       Dim.2
    <dbl>                                                       <dbl>
 1                        -2.53                                   2.48
 2                              9.55                              8.55
 3           -14.5                                                5.19
 4                              7.62                              3.29
 5                    -17.8                                      -10.4
 6                              2.07                              4.66
 7                        -2.48                                   6.84
 8              -13.7                                            -5.59
 9                        -9.43                                   4.56
10           -10.0                                               0.545
11                            -5.85                              -1.74
12                       -1.55                                    4.37
13                     11.9                                       5.26
14                       27.4                                    -9.47
15              -11.0                                            -6.07
16                            6.57                               0.410
17                            4.96                                2.26
18                           -6.51                             -0.0782
19                            -5.76                              -1.06
20                                  2.78                         -2.53
21                                  2.46                         -7.58
22  25.8   -3.86
```

```
#Plot MDS
ggscatter(mds, x = "Dim.1", y = "Dim.2",
          label = rownames(data1),
          size = 1,
          repel = TRUE)
```

Hasil visualisasi dari penskalaan multidimensi terkait data dan fokus contoh kasus yang pertama disajikan pada Gambar 7.1. Gambar 7.1 menunjukkan posisi dari 22 SDN di Kecamatan X berdasarkan nilai US yang terdiri dari variabel (mata pelajaran yang diujikan pada US) PKn, Bahasa Indonesia (B.INA), Matematika (MAT), Ilmu Pengetahuan Alam (IPA), Ilmu Pengetahuan Sosial (IPS), Seni Budaya dan Keterampilan (SBK), dan Pendidikan Jasmani, Olahraga,

dan Kesehatan (PJOK). Sekolah yang jaraknya berdekatan terindikasi memiliki kemiripan berdasarkan indikator nilai US. Dari hasil pemetaan tersebut, diperoleh hasil bahwa terdapat empat kelompok sekolah yang memiliki kemiripan antar anggotanya, namun berbeda dengan kelompok lainnya. Sebagai contoh, SDN S lebih mirip terhadap SDN K dibandingkan dengan SDN R. Hal ini terlihat jarak dari SDN S ke SDN K lebih dekat dibandingkan jarak dari SDN S ke SDN R.



Gambar 7.1 Peta SDN berdasarkan nilai US

Langkah yang selanjutnya yaitu membuat empat kelompok tersebut dengan menggunakan metode pengelompokan *k-means*. Setelah itu, setiap poin diwarnai sesuai dengan warna kelompoknya.

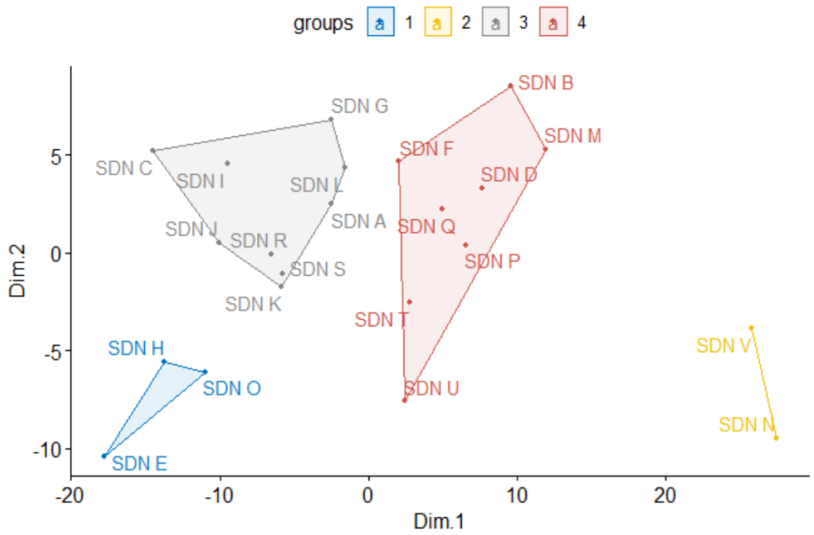
```
#K-means clustering
clust <- kmeans(mds, 4)$cluster %>%
  as.factor()
mds <- mds %>%
  mutate(groups = clust)
print(mds,n=22)
```

```
#           A           tibble:           22           ×           3
    Dim.1                                      Dim.2         groups
    <dbl>                                       <dbl>         <fct>
 1              -2.53                      2.48                    3
 2                   9.55                    8.55                  4
 3       -14.5                             5.19                    3
 4                   7.62                    3.29                  4
 5        -17.8                -10.4                               1
 6                   2.07                    4.66                  4
 7              -2.48                      6.84                    3
 8        -13.7                           -5.59                    1
 9              -9.43                      4.56                    3
10        -10.0                                0.545               3
11                -5.85                  -1.74                     3
12              -1.55                      4.37                    3
13             11.9                          5.26                  4
14              27.4                       -9.47                   2
15        -11.0                           -6.07                    1
16                    6.57                        0.410            4
17                  4.96                      2.26                  4
18                    -6.51                          -0.0782       3
19                -5.76                  -1.06                     3
20                    2.78                -2.53                   4
21                    2.46                -7.58                   4
22  25.8   -3.86   2
```

```
#Plot and color by groups
ggscatter(mds, x = "Dim.1", y = "Dim.2",
          label = rownames(data1),
          color = "groups",
          palette = "jco",
          size = 1,
          ellipse = TRUE,
          ellipse.type = "convex",
          repel = TRUE)
```

Hasil pengelompokan SDN berdasarkan kemiripan nilai US disajikan pada Gambar 7.2. Gambar 7.2 menunjukkan bahwa terdapat empat kelompok yang terbentuk untuk 22 SDN di Kecamatan X. Hasil pengelompokan menunjukkan bahwa terdapat tiga SDN pada kelompok I (yaitu SDN E, SDN H, dan SDN O), dua SDN pada kelompok II (yaitu SDN N dan SDN V), sembilan SDN pada kelompok III (yaitu SDN A, SDN C, SDN G, SDN I, SDN J, SDN K, SDN L, SDN R, dan SDN S), dan delapan SDN pada kelompok IV

(yaitu SDN B, SDN D, SDN F, SDN M, SDN P, SDN Q, SDN T, dan SDN U)



Gambar 7.2 Pengelompokan SDN berdasarkan nilai US dengan menggunakan metode *k-means*

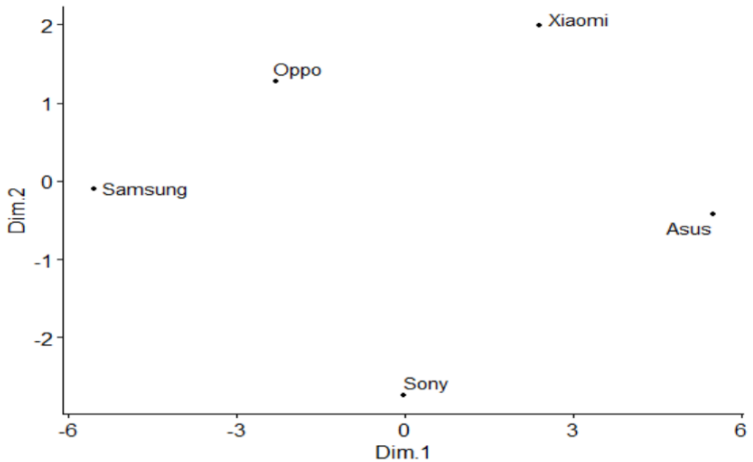
Analisis berikutnya dilakukan pada data yang kedua, yaitu data yang berfokus pada preferensi HP. Analisis pada data terkait dengan hal ini difokuskan pada penskalaan multidimensi dengan teknik non-metrik. Sama halnya dengan langkah-langkah analisis pada penskalaan multidimensi metrik, langkah pertama yang perlu dilakukan pada penskalaan multidimensi non-metrik yaitu memanggil dan menampilkan data yang akan dianalisis, dan melakukan analisis utama.

```
data <- read.csv('Datanonmetrik.csv', header = T, sep = ';')
datas<-data[,-1]
rownames(datas) <- c("Asus","Oppo","Samsung","Sony","Xiaomi")
head(datas)
```

```
#Load general packages:
library(magrittr)
library(dplyr)
library(ggpubr)
```

```
#Kruskal’s non-metric multidimensional scaling
#Compute MDS
library(MASS)
mds <- datas %>%
  dist() %>%
  isoMDS() %>%
  .$points %>%
  as_tibble()
initial                                 value                3.592004
final                                  value                0.000000
converged
colnames(mds) <- c("Dim.1", "Dim.2")
Mds
>                                                                  mds
#           A           tibble:          5          ×            2
    Dim.1                                                       Dim.2
    <dbl>                                                       <dbl>
1                      5.48                                     -0.418
2           -2.30                                                1.27
3             -5.55                                             -0.0977
4                           -0.0189                             -2.74
5  2.39    1.98
```

```
#Plot MDS
ggscatter(mds, x = "Dim.1", y = "Dim.2",
          label = rownames(datas,),
          size = 1,
          repel = TRUE)
```



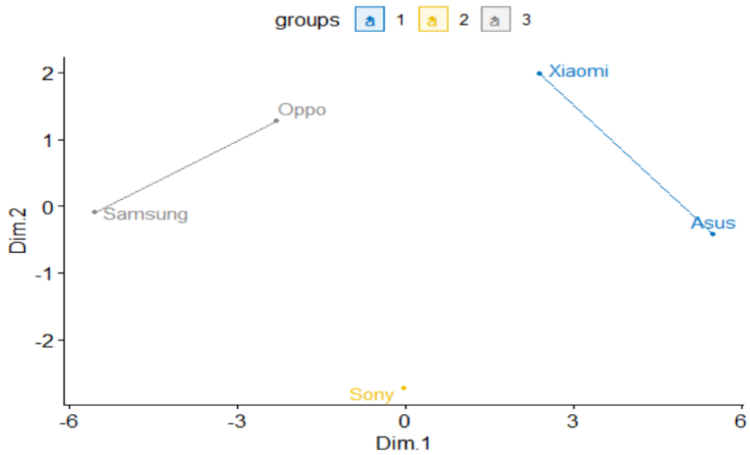
Gambar 7.3 Peta HP berdasarkan preferensinya

Gambar 7.3 menunjukkan posisi dari lima merek HP berdasarkan preferensinya yang terdiri dari variabel merek, desain, fitur, layar, harga, kemudahan, *processor*, memori, dan pemakaian. HP yang jaraknya berdekatan dapat dinyatakan memiliki kemiripan berdasarkan indikator preferensinya. Dari hasil pemetaan tersebut, kita dapat membentuk tiga kelompok HP yang memiliki kemiripan antar anggotanya namun berbeda dengan kelompok lainnya. Sebagai contoh, Oppo lebih mirip dengan Samsung dibandingkan dengan Xiaomi. Hal ini terlihat jarak dari Oppo ke Samsung yang lebih dekat dibandingkan jarak dari Oppo ke Xiaomi.

Langkah selanjutnya yaitu membuat tiga kelompok HP tersebut dengan menggunakan metode pengelompokan *k-means*. Setelah itu, titik yang menunjukkan merek HP diwarnai dan hubungkan dengan garis yang warnanya menunjukkan bahwa merek HP tersebut berada di kelompok yang sama yang mengindikasikan adanya kemiripan yang lebih dibanding dengan merek HP lain.

```
#K-means clustering
clust <- kmeans(mds, 3)$cluster %>%
  as.factor()
mds <- mds %>%
  mutate(groups = clust)
mds
>                                                                    mds
#           A           tibble:           5           ×           3
    Dim.1                                   Dim.2           groups
    <dbl>                                    <dbl>           <fct>
1               5.48                        -0.418               1
2       -2.30                              1.27                  3
3           -5.55                                  -0.0977       3
4           -0.0189           -2.74                              2
5  2.39    1.98   1
```

```
#Plot and color by groups
ggscatter(mds, x = "Dim.1", y = "Dim.2",
          label = rownames(datas),
          color = "groups",
          palette = "jco",
          size = 1,
          ellipse = TRUE,
          ellipse.type = "convex",
          repel = TRUE)
```



Gambar 7.4 Pengelompokan merek HP berdasarkan preferensi dengan menggunakan metode *k-means*

Gambar 7.4 menunjukkan bahwa terdapat tiga kelompok untuk lima merek HP. Hasil pengelompokan menunjukkan bahwa terdapat dua HP pada kelompok I (yaitu Xiaomi dan Asus), satu HP pada kelompok II (yaitu Sony), dan dua HP pada kelompok III (yaitu Samsung dan Oppo).

Bab 8
Analisis Konjoin

Dalam kehidupan sehari-hari, kita sering dihadapkan dengan berbagai persoalan yang menuntut kita untuk memilih salah satu dari berbagai pilihan solusi yang ditawarkan. Dalam memilih solusi yang ditawarkan tentu kita mempertimbangkan berbagai hal secara bersama-sama. Misalnya saja masalah COVID-19 yang menyerang seluruh aspek kehidupan manusia, dua di antaranya adalah aspek kesehatan dan pendidikan. Pemerintah dihadapkan dengan berbagai pilihan jenis vaksin yang dianggap mampu mengatasi masalah pada aspek kesehatan. Pemerintah dalam memutuskan jenis vaksin yang harus dipilih tentu mempertimbangkan preferensi dari masyarakat sebagai pengguna vaksin COVID-19. Begitu juga dalam mengatasi masalah pada aspek pendidikan, pemerintah mempertimbangkan preferensi dari masyarakat seperti pihak sekolah, siswa, dan orang tua dalam membuat kebijakan pembelajaran *online* selama pandemi COVID-19.

Preferensi dari masyarakat tersebut sangat dipengaruhi oleh atribut yang melekat pada produk atau jasa yang ditawarkan. Masyarakat akan mempertimbangkan atribut kehalalan, efek samping, biaya, dan tingkat keberhasilan dari jenis vaksin COVID-19 yang ditawarkan. Begitu juga dengan model pembelajaran online, pihak sekolah, siswa, dan orang tua akan mempertimbangkan atribut metode, aplikasi online, jenis tugas, dan bentuk evaluasi yang digunakan. Setiap atribut yang melekat pada suatu produk atau jasa memiliki level tertentu. Ini berarti, masyarakat dalam mengambil keputusan untuk memilih produk atau jasa tertentu mempertimbangkan atribut beserta level dari produk atau jasa tersebut secara bersama-sama.

Dalam konteks statistik multivariat, kegiatan pengukuran preferensi individu ataupun kelompok dengan mempertimbangkan atribut

dan level secara bersama-sama dikenal dengan istilah analisis konjoin (*conjoint analysis*). Kata '*conjoint*' itu sendiri merupakan akronim dari *considered jointly* (dipertimbangkan bersamaan) (Gudono, 2017b). Orang yang berjasa memperkenalkan pertama kali teknik analisis konjoin salah satunya adalah Paul Green pada tahun 1960-1970, seorang pakar di bidang psikometri. Berbeda dengan alat analisis multivariat lainnya yang biasa mengembangkan sebuah skor dari beberapa individu, analisis konjoin justru mengembangkan sebuah model preferensi untuk setiap individu. Hal ini membuat analisis konjoin banyak diminati oleh para peneliti atau pelaku di bidang pemasaran dan bidang lainnya yang membutuhkan pembobotan beberapa atribut secara bersamaan serta melibatkan pertukaran kepentingan (*trade-off*) antara atribut untuk menilai sesuatu. Beberapa penggunaan analisis konjoin di berbagai bidang di antaranya dalam pemilihan layanan elektronik publik (Pleger et al., 2020), preferensi pada vaksin (Motta, 2021), preferensi pembelajaran mahasiswa keperawatan (Macindo et al., 2019), preferensi instruktur mahasiswa keperawatan klinis (Factor & de Guzman, 2017), dan preferensi pembelajaran *online* untuk siswa SMA selama pandemi COVID-19 (Ong et al., 2022).

Teori dasar pada analisis konjoin

Konsep dasar pada analisis konjoin

Sebagai bagian dari metode *multivariate dependence*, bentuk umum dari analisis konjoin dapat dinyatakan sebagai berikut (Hair et al., 2010):

$$Y_1 = X_1 + X_2 + X_3 + \cdots + X_n$$

dengan Y_1 merupakan variabel dependen yang menggambarkan pendapat atau preferensi keseluruhan (*overal preference*) responden terhadap sekian atribut dan level atribut dari suatu produk atau jasa (datanya berupa metrik atau non-metrik); sementara X merupakan variabel independen atau biasa disebut sebagai faktor, sehingga bentuk $X_1 + X_2 + X_3 + \cdots + X_n$ merupakan atribut level dari suatu produk atau jasa (datanya berupa non-metrik). Dengan penggunaan variabel independen non-metrik, analisis konjoin sangat mirip deng-

an analisis varians (ANOVA) yang memiliki dasar dalam analisis eksperimen. Dengan demikian, analisis konjoin terkait erat dengan eksperimen tradisional.

Dalam melakukan analisis konjoin, seorang peneliti perlu memahami konsep *conjoint measurement* (CM) (Gudono, 2017b). Konsep CM membahas mengenai bagaimanakah preferensi seseorang atas atribut-atribut sebuah benda atau jasa membentuk ukuran *utility* (kebergunaan atau kebermanfaatan) orang tersebut atas benda atau jasa itu. Ini masalah mendasar yang perlu dipahami terlebih dulu karena di dalam analisis konjoin biasanya responden mengevaluasi stimulus yang ditampilkan kepadanya sebagai data nominal, walaupun terkadang ada yang berskala interval, sementara variabel dependen yang diukur dari preferensi responden bersifat ordinal. Hal ini akan memicu munculnya pertanyaan berupa "bagaimanakah efek gabungan (*joint effect*) stimulus tersebut membentuk variabel dependen?". Berkaitan dengan hal ini, maka dalam analisis konjoin umumnya dilakukan transformasi, sebagai berikut.

- $Pref(X) = \sum_{i=1}^{m} \sum_{j=1}^{k} \alpha_{ij} x_{ij} + \varepsilon$, di mana $Pref(X)$ merupakan preferensi sebuah alternatif yang diukur dengan *rank-order* (skala ordinal); m jumlah atribut; k jumlah level atribut; α_{ij} kontribusi sebuah level atribut (*part-worth*) pada tingkat preferensi atau *utility* responden; x_{ij} bernilai 1 jika stimulusnya adalah level j atribut i. Model ini menggunakan pendekatan regresi dengan x_{ij} merupakan variabel *dummy* (buatan). Setelah fungsi regresi ditentukan, maka selanjutnya perlu dilakukan penyesuaian skala (*rescale*) agar level dari masing-masing preferensi berada pada rentang 0 sampai 100 (0 preferensi yang paling tidak disukai dan 100 preferensi yang paling disukai). Dengan demikian, persamaan $Pref(X)$ akan berubah menjadi $U(X) = \sum_{i=1}^{m} \sum_{j=1}^{k} \alpha'_{ij} x_{ij}$, dengan $U(X)$ merupakan *utility* dan α'_{ij} merupakan nilai dari α_{ij} yang telah disesuaikan skalanya (*rescale*).
- $\mu_{ijk} = \mu + \beta_{1i} + \beta_{2j} + \cdots + \beta_{nk} + \varepsilon_{ijk}$, di mana μ_{ijk} merupakan *utility* responden. Model ini merupakan model analisis konjoin metrik dengan n faktor dan setiap faktor memiliki level, misalnya faktor 1 memiliki level i, faktor 2 memiliki level j, dan se-

terusnya; $\sum \beta_{1i} = \sum \beta_{2j} = \cdots = \sum \beta_{nk} = 0$. Apabila diperhatikan secara seksama, maka model ini jelas sekali menggunakan pendekatan ANOVA.

Analisis konjoin memiliki premis dasar bahwa preferensi dibangun atas dasar utilitas (*utility*). Semakin tinggi utilitas, maka semakin tinggi preferensi (Ryan & Deci, 2020). Setiap level atribut dari suatu produk atau jasa itu memiliki nilai utilitas tertentu. Kekuatan preferensi responden atas suatu produk atau jasa disebut juga *worth* atau *utility* (Gudono, 2017b). Hubungan antara masing-masing level sebuah atribut suatu produk atau jasa dengan preferensi responden disebut *part-worth function* (biasa juga disebut *utility function*) (Hair et al., 2010). Dalam analisis konjoin, total *worth* atau *utility* suatu produk atau jasa dianggap sebagai penjumlahan atas semua *part-worth* masing-masing level semua atribut dari sebuah produk atau jasa (Hair et al., 2010).

Menurut Gudono (2017b), seorang peneliti perlu mengetahui tujuan melakukan analisis konjoin karena akan sangat membantu dalam memformulasikan masalah atau faktor-faktor penting bagi responden dalam menilai suatu produk atau jasa. Ada dua tujuan utama kenapa perlu melakukan analisis konjoin, yakni sebagai berikut. Pertama yaitu menentukan kontribusi dari faktor-faktor beserta nilainya yang mempengaruhi profil preferensi responden atas hal (produk atau jasa) tertentu. Tujuan kedua yaitu untuk mengembangkan model *judgment* responden yang berguna untuk menjelaskan keputusan responden atas hal-hal yang dievaluasinya ataupun atas produk atau jasa yang belum ada sekalipun. Secara spesifik, tujuan dari analisis konjoin di antaranya yaitu untuk mengetahui bobot (*weight*) tiap atribut produk atau jasa, nilai (*part-worth*) dari tiap level atribut, meramalkan preferensi konsumen atas beberapa konsep (baru) produk yang telah didefinisikan level atributnya dan meramalkan (simulasi) pangsa pasar dari beberapa konsep (baru) produk atau jasa (Gracía-Pérez & Gil-Lacruz, 2018).

Asumsi-asumsi pada analisis konjoin

Berbeda dengan analisis multivariat lainnya, analisis konjoin tidak memerlukan asumsi normalitas, linearitas, multikolinearitas, heteros-

kedasitas, dan lainnya. Adapun asumsi-asumsi yang perlu diperhatikan dalam analisis konjoin yaitu sebagai berikut (Gudono, 2017b): *Pertama*, konsumen atau subjek (responden) diminta mempertimbangkan atribut-atribut barang dan jasa harus berpikir rasional. *Kedua*, dalam menetapkan preferensinya untuk memilih alternatif (produk atau jasa yang mana), konsumen atau subjek (responden) dipastikan mengevaluasi semua atribut-atribut barang dan jasa serta mampu membuat *trade-off*. *Ketiga*, atribut-atribut dari suatu produk atau jasa dapat diidentifikasi. *Keempat*, sifat preferensi terhadap suatu objek bersifat tambahan (aditif). Ini berarti bahwa preferensi total terhadap sebuah objek adalah penjumlahan preferensi atas semua atribut yang melekat pada objek tersebut. Selain empat asumsi tersebut, oleh karena analisis konjoin melibatkan skenario bagaimana responden merespons atribut-atribut yang diminta oleh peneliti untuk mereka evaluasi, hampir pada setiap analisis konjoin peneliti membuat asumsi-asumsi tambahan mengenai sifat data atau faktor-faktor yang diamatinya serta ketepatan dari model yang digunakannya (Hair et al., 2010).

Langkah-langkah penelitian dengan analisis konjoin

Penelitian dengan analisis konjoin merupakan hal yang unik karena metode pengumpulan data, desain penelitian, dan metode analisis datanya saling terkait sehingga perlu dipertimbangkan secara bersamaan pada tahapan perencanaan penelitian. Sebagaimana diketahui, responden penelitian dalam riset analisis konjoin diskenariokan untuk mengevaluasi sekelompok informasi mengenai atribut (misalnya "bentuk penugasan") berbagai produk (misalnya "pembelajaran *online*"). Masing-masing atribut tersebut tentu memiliki level (misalnya bentuk penugasan "individu" dan "kelompok"). Kemudian responden akan mempertimbangkan berbagai alternatif pilihan (misalnya membeli atau tidak membeli) atau menunjukkan kadar preferensi (kesukaan) mereka atas produk tersebut. Dari penilaian tersebut peneliti kemudian menentukan utilitas (tingkat nilai kesukaan) responden atas masing-masing atribut produk (termasuk misalnya bagaimanakah nilai relatif kesukaan atas "kulit" dibandingkan dengan "imitasi". Setelah itu, analisis atas utilitas responden dimulai

di mana akan digambar dan berbagai simulasi juga bisa dilakukan. Berdasarkan uraian tersebut, langkah-langkah yang dilakukan dalam analisis konjoin yaitu sebagai berikut.

1. *Memformulasi masalah.* Pada tahap ini peneliti perlu mengetahui faktor-faktor apakah yang penting bagi responden dalam menilai suatu produk tahu apa yang akan mereka pertimbangkan saat akan memutuskan membeli atau tidak membeli suatu produk atau menggunakan suatu jasa (Hair et al., 2010). Biasanya dalam menentukan faktor-faktor apa saja penting bagi responden, seorang peneliti perlu melakukan kajian literatur atau juga memanfaatkan informasi dari berbagai macam pengalaman yang pernah dialaminya (Ong et al., 2022).
2. *Merancang stimulus-stimulus*. Stimulus-stimulus adalah faktor yang dipertimbangkan pada saat responden membuat *judgment* mengenai suatu hal (misalnya produk) (Gudono, 2017b). Jelas sekali bahwa konsep stimulus ini dipinjam dari desain penelitian eksperimen. Ini wajar karena proses pemberian stimulus berupa atribut produk dan levelnya seolah-olah dirancang seperti dalam desain eksperimen faktorial penuh (Hair et al., 2010). Penyusunan stimulus dapat diatur dengan tiga macam pendekatan yaitu sebagai berikut.
   - *Pendekatan pasangan (pairwise approach)*. Dalam pendekatan ini, responden menilai dua atribut sekaligus sampai semua kemungkinan pasangan dua atribut telah selesai dievaluasi. Tidak ada keharusan untuk mengevaluasi semua kemungkinan kombinasi karena ini tidak mungkin untuk dilaksanakan. Di dalam pendekatan *pairwise* terdapat kemungkinan untuk mereduksi atau mengurangi jumlah perbandingan pasangan dengan menggunakan *cyclical designs*.
   - *Pendekatan penuh (full-profile approach)*. Dalam pendekatan ini, setiap profil dijelaskan secara terpisah. Ada dua desain yang digunakan dalam pendekatan ini, yakni *full factorial design* dan *fractional factorial design*. Jika menggunakan desain *full fractional*, maka semua kombinasi stimulus-stimulus akan ditampilkan. Menurut Paul Green, jika kombinasi stimulus-stimulus lebih dari 30 macam, maka akan sukat dianalisis oleh

responden (Gudono, 2017b). Olehnya itu, sebisa mungkin ini disederhanakan melalui *fractional factorial design*. Desain *fractional factorial* ini adalah variasi dari desain faktorial dasar yang hanya sebagian (*subset*) dari kombinasi stimulus-stimulus yang dijalankan (Hair et al., 2010). Tentu saja dalam menentukan *subset* kombinasi stimulus-stimulus mana yang akan digunakan peneliti tidak boleh sembarangan. Artinya, ini harus diatur sedemikian rupa sehingga subset yang tidak ditampilkan jangan sampai mengganggu pembandingan antar atribut produk atau jasa. Oleh sebab itu, terdapat saran agar setiap level atribut sebaiknya pernah berpasangan dengan setiap level atribut lainnya.

- *Orthogonal array*. Pendekatan ini merupakan bagian (*subset*) dari *full factorial design* dengan cara memusatkan pada *main effect*. Hal ini karena interaksi atribut dianggap bisa diabaikan (Ong et al., 2022). Dalam *orthogonal array*, setiap level sebuah faktor diatur agar terjadi bersamaan dengan setiap level faktor lainnya dengan frekuensi yang sama atau proporsional dan tetap menjamin independensi *main effect*.

3. *Memutuskan bentuk data masukan (input)*. Umumnya ada dua macam input data yang diharapkan dari responden, yakni data non-metrik dan metrik. Untuk data non-metrik, responden diminta memberi penilaian dalam bentuk data *rank-order* (data dengan skala ordinal) (Hair et al., 2010). Sementara untuk kasus data metrik, responden diminta untuk memberi penilaian dalam bentuk data *rating* (bukan *rank-order*, misalnya data dengan skala rasio atau interval) (Hair et al., 2010).
4. *Memilih desain analisis konjoin*. Desain analisis konjoin itu sendiri pada dasarnya sama dengan rancangan dari desain eksperimen (Gudono, 2017b). Desain analisis konjoin sangat tergantung pada pendekatan yang dipilih dalam merancang stimulus-stimulus dan jenis data yang diolah. Jika kita memilih menggunakan pendekatan *full profile* dalam merancang stimulus-stimulus, maka desain yang memungkinkan untuk dipilih yaitu *full factorial* dan f*ull fractional*.

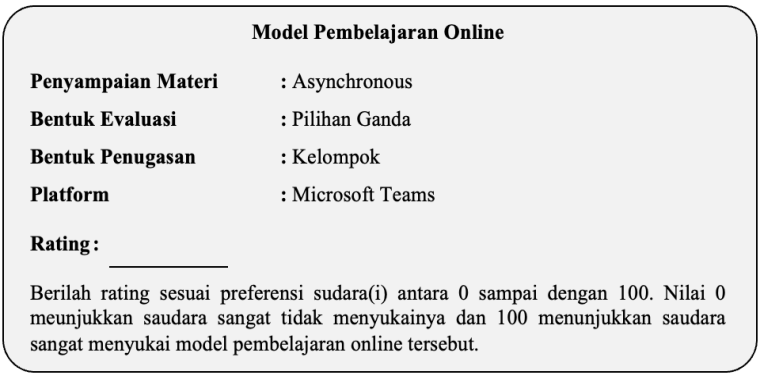
5. *Menyusun instrumen pengumpulan data*. Penyusunan instrumen pengumpulan data menyangkut dua masalah, yaitu bagaimana informasi disajikan pada responden dan bagaimana stimulus-stimulus (gambaran mengenai atribut produk atau jasa) ditampilkan dan diatur. Berikut adalah metode-metode dan contoh penyajian informasi tersebut.
   - *Metode trade-off matrices*. Dalam metode ini, responden hanya mempertimbangkan sepasang atribut dalam setiap waktu. Metode ini sangat sederhana namun kurang realistis karena hanya meminta responden membandingkan sepasang atribut, padahal dalam kenyataannya atribut bisa banyak sekali. Selain masalah tersebut, metode ini juga dapat membuat responden lelah dan bingung karena pada saat bersamaan harus memikirkan banyak kotak penilaian. Gambar 8.1 menunjukkan contoh penggunaan metode *trade-off matrices*.

Jenis Penyampaian Materi

Evaluasi	Asynchronous	Synchronous	Mixed
Pilihan Ganda			
Esai			
Proyek			

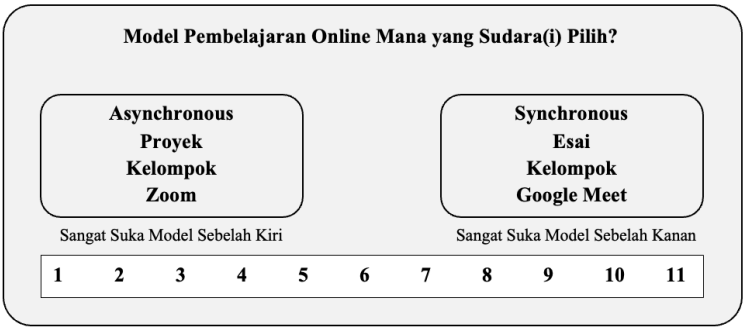
Gambar 8.1 Contoh penggunaan metode *trade-off matrices*

   - *Metode full-profile card sort*. Metode ini memungkinkan responden mengevaluasi banyak atribut secara serentak. Dengan metode ini peneliti harus menyiapkan kartu-kartu (bisa juga berupa lembaran kertas yang dicetak) yang berisi informasi mengenai satu level dari semua atribut produk atau jasa. Setelah itu responden diminta memberikan *rating* pada masing-masing kartu sesuai dengan preferensi mereka atas nilai kombinasi atribut yang ditampilkan. Kombinasi yang ditampilkan dalam kartu mengikuti prinsip-prinsip desain eksperimen. Jika setiap level atribut diberi peluang berpasangan dengan masing-masing level atribut lainnya, maka metode ini disebut *full factorial design*. Gambar 8.2 menunjukkan contoh penggunaan dari metode *full-profile card sort*.



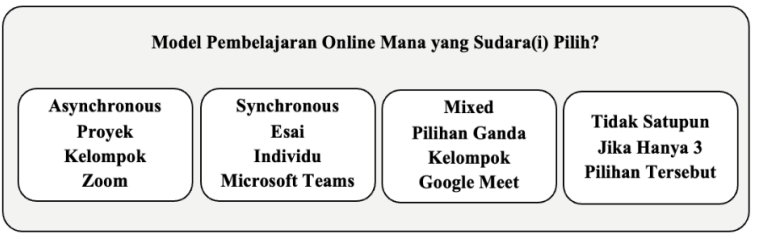
Gambar 8.2 Contoh penggunaan metode *full-profile card sort*

- *Hybrids conjoint dengan metode pairwise comparisons.* Dalam metode ini, responden diminta untuk mengevaluasi dua (sepasang) konsep produk atau jasa dan diminta menunjukkan preferensi mereka dengan menggunakan skala *rating*. Titik tengah skala menunjukkan kedua produk atau jasa yang sama-sama disukai. Gambar 8.3 menunjukkan contoh penggunaan *hybrids conjoint* dengan metode *pairwise comparisons*.



Gambar 8.3 Contoh penggunaan *hybrids conjoint* dengan metode *pairwise comparisons*

- *Metode discrete-choice*. Metode ini disebut juga *discrete-choice modeling*. Metode ini tidak meminta responden menunjukkan preferensinya, melainkan responden diminta memilih satu. Gambar 8.4 menunjukkan contoh penggunaan metode ini.



Gambar 8.4 Contoh penggunaan metode *discrete-choice*

6. *Mengumpulkan dan interpretasi hasil analisis data*. Setelah instrumen pengumpulan data disusun, langkah selanjutnya adalah kegiatan pengumpulan data. Data yang telah terkumpul kemudian dianalisis dengan beberapa macam teknik, misalnya menggunakan ANOVA, *dummy regression*, dan *part-worth function*.
7. *Membuktikan validitas dan estimasi reliabilitas*. Peneliti perlu melakukan evaluasi atas hasil yang diperoleh. Salah satu cara yang dilakukan adalah dengan melakukan *cross-validation* (Ong et al., 2022). Peneliti menerapkan prediksi dari hasil pengolahan data pada sampel *holdout*. Data pada sampel *holdout* adalah data yang sengaja dipisahkan dan tidak dianalisis untuk menghitung parameter statistik atau bagian dari kombinasi faktor (di dalam *factor analysis*) yang dipisahkan dan tidak dianalisis untuk menentukan *part-worth function*. Selain itu, peneliti juga dapat melakukan estimasi reliabilitas dengan cara menghitung korelasi, baik dengan korelasi Spearman atau Kendall's tau, antara data riil dengan data prediksi (Ong et al., 2022).
8. *Melakukan simulasi*. Berdasarkan model dan data yang tergambar dalam *part-worth function*, peneliti dapat mengembangkan simulasi untuk meramalkan preferensi seseorang atas sebuah atau beberapa produk atau jasa yang bersifat baru atau belum ada. Dengan teknik simulasi ini, peneliti dapat memperkirakan atau meramalkan berapa pangsa pasar dari produk atau jasa baru yang akan dikembangkan dengan menganalisis preferensi yang dimiliki oleh seseorang atas produk yang akan dikembangkan tersebut.

Setelah paparan mengenai konsep dasar pada analisis konjoin, asumsi-asumsi pada analisis konjoin, dan langkah-langkah yang diperlukan untuk melakukan analisis konjoin, pada bagian berikutnya

disajikan contoh kasus penggunaan analisis konjoin dengan menggunakan bantuan program R di bawah RStudio. Hal ini dilakukan untuk memberikan pemahaman yang lebih lanjut pada penggunaan atau penerapan dari analisis konjoin.

Contoh kasus dan analisis konjoin menggunakan program R dan RStudio

Contoh kasus

Pandemi COVID-19 telah mengakibatkan pergeseran dari pembelajaran tatap muka ke pembelajaran *online*. Sebuah penelitian dilakukan untuk mengevaluasi preferensi mahasiswa di daerah Provinsi Daerah Istimewa Yogyakarta pada atribut pembelajaran *online* selama pandemi COVID-19 dengan cara memanfaatkan analisis konjoin. Langkah pertama yang dilakukan dalam penelitian ini adalah merumuskan masalah (mengidentifikasi atribut-atribut yang penting dari pembelajaran *online*). Berdasarkan kajian literatur diperoleh informasi bahwa ada empat atribut penting dari pembelajaran *online* yang mempengaruhi preferensi siswa atau mahasiswa. Adapun keempat atribut tersebut beserta levelnya disajikan pada Tabel 8.1. Setelah atribut beserta levelnya ditentukan, langkah yang selanjutnya yaitu merancang stimulus-stimulus.

Tabel 8.1 Atribut pembelajaran *online*

Atribut	Level-atribut
Jenis penyampaian materi	Asynchronous
	Synchronous
	Mixed
Evaluasi	Pilihan ganda
	Esai
	Proyek
Bentuk tugas	Individu
	Kelompok
Platform	Zoom
	Microsoft Teams
	Google Meet

Prosedur analisis

Analisis pada contoh kasus dilakukan sesuai dengan prosedur atau langkah-langkah analisis konjoin yang telah dipaparkan pada bagian sebelumnya. Langkah pertama yaitu merancang stimulus-stimulus. Penelitian ini diilustrasikan menggunakan pendekatan *full-profile* dalam merancang stimulus-stimulus. Penataan stimulus-stimulus akan diuraikan pada bagian desain analisis konjoin. Langkah kedua yaitu menentukan bentuk data masukan (*input*). Variabel dependen dari penelitian ini yakni preferensi mahasiswa terkait atribut dalam pembelajaran *online* selama pandemi COVID-19. Preferensi dari mahasiswa diukur menggunakan *rating scale* sehingga data yang diperoleh berskala interval. Oleh karena itu, data *input* dari penelitian ini bersifat metrik.

Langkah selanjutnya yaitu memilih desain analisis konjoin yang akan digunakan. Contoh kasus yang diberikan menunjukkan bahwa penelitian menggunakan desain *fractional factorial* dengan teknik *orthogonal array* sesuai dengan pendekatan rancangan stimulus-stimulus dan bentuk data *input* yang digunakan. Lebih lanjut, penggunaan desain *fractional factorial* didasarkan pada pertimbangan bahwa kombinasi antar level-atribut yang terbentuk sangat besar jika menggunakan pendekatan *full-profile*, yakni $3 \times 3 \times 2 \times 3 = 54$ kombinasi. Sesuai rekomendasi dari Paul Green, jumlah kombinasi atribut yang lebih dari 30 perlu disederhanakan. Desain *fractional factorial* dengan teknik *orthogonal array* dapat menghasilkan kombinasi atribut yang lebih sederhana dengan hasil yang *robust*. Untuk dapat menata stimulus-stimulus dengan menggunakan teknik *orthogonal array* berbantuan program R dan RStudio, kita perlu memasang dan menjalankan paket analisis konjoin terlebih dahulu sebagai berikut.

```
#Memasang dan menjalankan paket untuk analisis konjoin
> install.packages('conjoint')
> library('conjoint')
```

Setelah berhasil memasang dan menjalankan paket yang diperlukan untuk analisis konjoin, kita selanjutnya perlu mendefinisikan atribut beserta levelnya menggunakan fungsi "`expand.grid`" pada R.

```
#Mendefinisikan atribut
atribut <- expand.grid(
  Peny_Materi = c("Asynchronous", "Synchronous", "Mixed"),
  Evaluasi = c("Pilihan Ganda", "Esai", "Proyek"),
  Tugas = c("Kelompok", "Individu"),
  Platform = c("Zoom", "Microsoft Teams", "Google Meet")
)
```

Variabel atribut yang baru saja didefinisikan akan menghasilkan *dataframe* yang berisikan kombinasi antar level-atribut, yakni sebanyak 54 kombinasi. Untuk memeriksanya, kita dapat menggunakan fungsi "`head`" dan "`tail`" dengan menggunakan perintah berikut sedemikian sehingga diperoleh informasi terkait banyaknya kombinasi antar level-atribut.

```
#Lihat enam data atribut teratas
head(atribut)
> head(atribut)
   Peny_Materi      Evaluasi    Tugas Platform
1 Asynchronous Pilihan Ganda Kelompok     Zoom
2  Synchronous Pilihan Ganda Kelompok     Zoom
3        Mixed Pilihan Ganda Kelompok     Zoom
4 Asynchronous          Esai Kelompok     Zoom
5  Synchronous          Esai Kelompok     Zoom
6        Mixed          Esai Kelompok     Zoom

#Lihat enam data atribut terbawah
tail(atribut)
> tail(atribut)
    Peny_Materi Evaluasi    Tugas    Platform
49 Asynchronous     Esai Individu Google Meet
50  Synchronous     Esai Individu Google Meet
51        Mixed     Esai Individu Google Meet
52 Asynchronous   Proyek Individu Google Meet
53  Synchronous   Proyek Individu Google Meet
54        Mixed   Proyek Individu Google Meet
```

Setelah memastikan variabel atribut, langkah selanjutnya yaitu membuat atau merancang stimulus-stimulus menggunakan desain *fractional factorial* dengan *orthogonal array*. Pada sintaks atau pe-

rintah berikut terdapat fungsi "`seed = 123`" yang bertujuan agar rancangan stimulus-stimulus yang terbentuk tetap sama meskipun diulang berkali-kali. Untuk melihat rancangan stimulus-stimulus yang telah dibuat, kita hanya perlu mengetikkan kembali kata atau perintah "`desain`" kemudian jalankan perintah tersebut.

```
desain <- caFactorialDesign(data = atribut, type = "orthogonal", seed = 123)
desain
> desain
    Peny_Materi      Evaluasi    Tugas        Platform
1  Asynchronous Pilihan Ganda Kelompok            Zoom
8   Synchronous        Proyek Kelompok            Zoom
15        Mixed          Esai Individu            Zoom
21        Mixed Pilihan Ganda Kelompok Microsoft Teams
23  Synchronous          Esai Kelompok Microsoft Teams
34 Asynchronous        Proyek Individu Microsoft Teams
40 Asynchronous          Esai Kelompok     Google Meet
45        Mixed        Proyek Kelompok     Google Meet
47  Synchronous Pilihan Ganda Individu     Google Meet
```

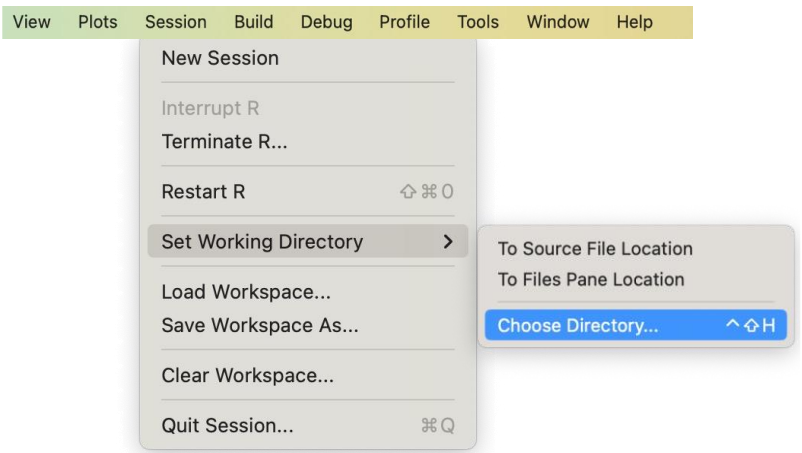
Hasil di atas menunjukkan bahwa 54 kombinasi level-atribut dengan desain *full factorial* berhasil direduksi menjadi 9 kombinasi level-atribut melalui teknik *orthogonal array*. Kita juga dapat memeriksa hubungan antar atribut, namun sebelumnya perlu mengodekan label ke dalam bentuk angka (*integer*) dengan menggunakan fungsi "`caEncodedDesign()`".

```
kode <- caEncodedDesign(desain)
kode
> kode
   Peny_Materi Evaluasi Tugas Platform
1            1        1     1        1
8            2        3     1        1
15           3        2     2        1
21           3        1     1        2
23           2        2     1        2
34           1        3     2        2
40           1        2     1        3
45           3        3     1        3
47           2        1     2        3
```

```
cor(kode)
> cor(kode)
            Peny_Materi Evaluasi Tugas Platform
Peny_Materi           1        0     0        0
Evaluasi              0        1     0        0
Tugas                 0        0     1        0
Platform              0        0     0        1
```

Selanjutnya kita perlu menyimpan desain stimulus-stimulus yang telah dibuat beserta kodenya dalam bentuk file Excel dengan ekstensi `.csv` pada folder tertentu. Desain stimulus-stimulus tersebut digunakan untuk panduan dalam menyusun instrumen pengumpulan data, sedangkan kode dari desain stimulus-stimulus akan digunakan dalam analisis konjoin untuk mengestimasi parameter statistik. Agar kedua data tersimpan pada folder yang sama, maka perlu dilakukan "`Set Working Directory`" pada RStudio seperti yang disajikan pada Gambar 8.5. Setelah menekan "`Choose Directory...`", silakan tentukan di folder mana file-file tersebut akan disimpan. Setelah itu jangan lupa menekan "`Open`". Jika berhasil, maka pada bagian *console* RStudio akan muncul seperti ini, di mana dalam contoh kasus ini, semua file akan disimpan pada folder UAS yang berada di folder KULIAH S3 PEP/SEMESTER-1/STATISTIK MULTIVARIAT).

```
setwd("~/Library/CloudStorage/OneDrive-uny.ac.id/KULIAH S3
PEP/SEMESTER-1/STATISTIK MULTIVARIAT/UAS")
```



Gambar 8.5 Menentukan *working directory*

```
write.csv(desain, 'Desain Stimuli.csv')
write.csv(kode, 'Kode Desain Stimuli.csv')
```

Langkah selanjutnya yaitu menyusun instrumen pengumpulan data. Berdasarkan penataan stimulus-stimulus yang terbentuk, yaitu sembilan kombinasi level-atribut, maka penyajian informasi pada responden dalam penelitian ini diilustrasikan menggunakan metode *full-profile card sort*. Pemilihan metode ini bertujuan agar responden dapat mengevaluasi atribut pembelajaran *online* selama pandemi COVID-19 secara bersamaan. Setiap kombinasi level-atribut disusun dalam satu kartu penilaian. Oleh karena itu, instrumen pengumpulan data dalam penelitian ini berisi sembilan kartu penilaian yang masing-masing akan dievaluasi atau dinilai oleh responden.

Setelah instrumen pengumpulan data berhasil disusun, langkah selanjutnya yaitu mengumpulkan dan interpretasi hasil analisis data. Data preferensi mahasiswa terkait atribut pembelajaran *online* dalam penelitian ini dikumpulkan secara *online* menggunakan instrumen yang telah dibuat sebelumnya. Data yang telah dikumpulkan kemudian disimpan dengan nama "`Data Konjoin`" dalam satu folder bersama dengan file bernama "`Desain Stimuli.csv`" dan "`Kode Desain Stimuli.csv`" dalam bentuk file Excel dengan ekstensi `.csv`. Hal ini bertujuan agar memudahkan dalam pemanggilan dan penyimpanan hasil analisis data.

```
#Memanggil data profil
prof <- read.csv('Kode Desain Stimuli.csv', sep = ";",
header = T)
prof
> prof
   X Peny_Materi Evaluasi Tugas Platform
1  1           1        1     1        1
2  8           2        3     1        1
3 15           3        2     2        1
4 21           3        1     1        2
5 23           2        2     1        2
6 34           1        3     2        2
7 40           1        2     1        3
8 45           3        3     1        3
9 47           2        1     2        3
```

Data yang dianalisis ada dua, yakni data profil yang merupakan kode desain stimulus-stimulus, yaitu `Kode Desain Stimuli.csv`, data preferensi, yaitu `Data Konjoin.csv`, dan data level setiap atribut. Data profil ditampung dalam variabel `prof`, data preferensi ditampung dalam variabel `pref`, dan data level setiap atribut ditampung dalam variabel `level`. Perintah-perintah yang digunakan pada langkah ini bersama dengan hasil-hasil yang diperoleh dari perintah-perintah tersebut disajikan di bagian halaman sebelumnya. Terlihat bahwa dalam data "`prof`" terdapat variabel *X* pada kolom pertama yang merupakan ID atau identitas dari kombinasi level-atribut, sehingga perlu dihapus dengan menggunakan perintah berikut sedemikian sehingga dapat diperoleh hasil sesuai dengan kehendak.

```
#Menghapus kolom pertama (X)
prof <- prof[,-1]
prof
> prof
  Peny_Materi Evaluasi Tugas Platform
1           1        1     1        1
2           2        3     1        1
3           3        2     2        1
4           3        1     1        2
5           2        2     1        2
6           1        3     2        2
7           1        2     1        3
8           3        3     1        3
9           2        1     2        3
```

```
#Memanggil data preferensi
pref <- read.csv('Data Konjoin.csv', sep = ";", header = T)
head(pref) #menampilkan 6 data teratas
tail(pref) #menampilkan 6 data terbawah
> head(pref)
  p1 p2 p3 p4 p5 p6 p7 p8 p9
1  3  2  2  1  4  4  1  5  2
2  4  4  3  3  3  4  4  4  4
3  3  5  5  4  4  4  3  4  3
4  3  4  4  4  4  4  4  4  4
5  5  3  4  5  4  3  4  3  5
6  4  2  3  3  3  3  4  4  4
```

```
> tail(pref)
   p1 p2 p3 p4 p5 p6 p7 p8 p9
25  3  4  2  2  4  2  4  3  2
26  4  4  3  3  3  5  5  4  5
27  3  5  5  1  5  5  5  5  1
28  3  4  1  3  2  4  2  4  4
29  4  4  2  4  4  3  4  4  4
30  3  4  4  3  4  4  4  4  3

#Membuat data level untuk empat atribut
level <- c("Asynchronous", "Synchronous", "Mixed",
         "Pilihan Ganda", "Esai", "Proyek",
         "Kelompok", "Individu",
         "Zoom", "Microsoft Teams", "Google Meet")
level <- data.frame(level)
level
> level
            level
1     Asynchronous
2      Synchronous
3            Mixed
4    Pilihan Ganda
5             Esai
6           Proyek
7         Kelompok
8         Individu
9             Zoom
10 Microsoft Teams
11     Google Meet
```

Setelah ketiga data sudah siap, langkah selanjutnya yaitu memulai analisis konjoin. Untuk memperoleh hasil analisis menggunakan teknik analisis regresi (*dummy variable*), fungsi "`caModel()`" dapat digunakan. Pada bagian berikut dipaparkan secara detail perintah yang digunakan dalam analisis konjoin dan hasil-hasil yang diperoleh dari perintah yang bersesuaian. Perintah dimaksud di sini yaitu perintah yang bertujuan untuk mengestimasi parameter analisis konjoin pada responden pertama.

```
#Estimasi parameter analisis konjoin pada responden pertama
caModel(y=pref[1,], x=prof)
```

```
> caModel(y=pref[1,], x=prof)

Call:
lm(formula = frml)

Residuals:
       1        2        3        4        5        6        7        8        9
1.333e+00 -1.333e+00  4.441e-16 -1.333e+00  1.333e+00  5.551e-16 -1.333e+00  1.333e+00  6.163e-33

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              2.667e+00  1.155e+00   2.309    0.260
factor(x$Peny_Materi)1 -5.861e-17  1.540e+00   0.000    1.000
factor(x$Peny_Materi)2  2.030e-17  1.540e+00   0.000    1.000
factor(x$Evaluasi)1    -6.667e-01  1.540e+00  -0.433    0.740
factor(x$Evaluasi)2    -3.333e-01  1.540e+00  -0.217    0.864
factor(x$Tugas)1        1.128e-16  1.155e+00   0.000    1.000
factor(x$Platform)1    -3.333e-01  1.540e+00  -0.217    0.864
factor(x$Platform)2     3.333e-01  1.540e+00   0.217    0.864

Residual standard error: 3.266 on 1 degrees of freedom
Multiple R-squared:  0.3333,   Adjusted R-squared:  -4.333
F-statistic: 0.07143 on 7 and 1 DF,  p-value: 0.9928
```

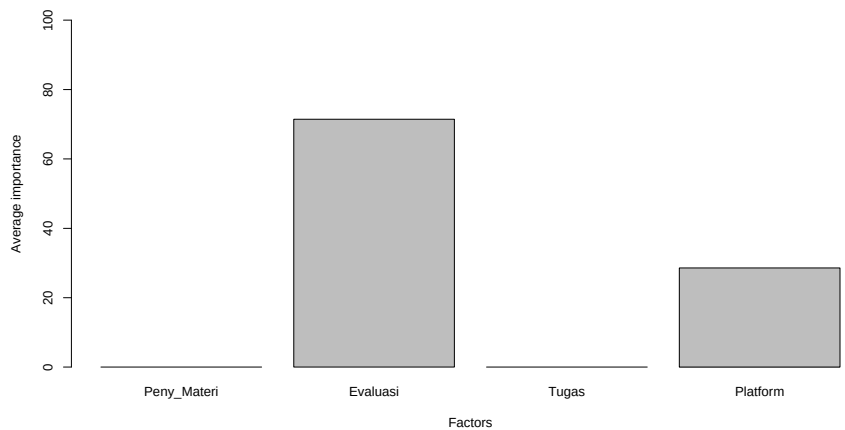
Dengan memperhatikan hasil yang diperoleh atas perintah di atas, diperoleh informasi bahwa setiap atribut ada satu level yang memiliki koefisien sebesar 0 karena dijadikan sebagai model dasar (*based model*). Hal ini karena teknik yang digunakan yaitu teknik analisis regresi dengan variabel *dummy*. Jadi, untuk penyampaian materi, jenis *mixed* merupakan model dasar. Untuk mengetahui *importance value* masing-masing atribut beserta gambar *importance* dan nilai *utility* pada setiap level-atribut beserta gambar *part-worth function*, fungsi "`Conjoint()`" dapat digunakan. Agar luaran *importance value* lebih rapi, kita dapat menggunakan fungsi "`caImportance`", sedangkan untuk menampilkan bentuk visual dari *importance value* setiap atribut, fungsi "`Conjoint()`" dapat digunakan.

```
#Importance responden pertama untuk setiap atribut
nama.atribut <- colnames(atribut)
data.import <- caImportance(y=pref[1,], x=prof)
importance <- data.frame(Atribut = nama.atribut, Importance
= data.import)
importance
```

```
> importance
      Atribut Importance
1 Peny_Materi       0.00
2    Evaluasi      71.43
3       Tugas       0.00
4    Platform      28.57
```

Luaran (*output*) di atas menunjukkan bahwa responden pertama lebih mementingkan atribut evaluasi (71,43%) kemudian diikuti tiga atribut lainnya seperti jenis platform (28,57%), penyampaian materi (0,00%), dan bentuk penugasan (0,00%). Selanjutnya, *importance value* dari masing-masing atribut pembelajaran *online* divisualisasikan seperti yang tersaji pada Gambar 8.6 melalui *output* dari fungsi "`Conjoint`" dengan perintah berikut.

```
Conjoint(y=pref[1,], x=prof, z=level)
```

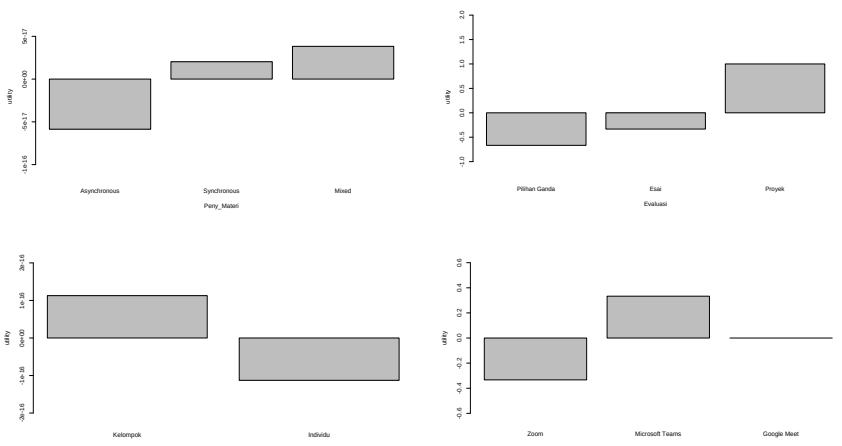


Gambar 8.6 *Importance value* atribut pembelajaran *online* dari responden pertama

Berdasarkan informasi yang diperoleh dari *importance value* setiap atribut, kita dapat mengetahui secara lebih spesifik nilai *utility* dari masing-masing level atribut berdasarkan luaran (*output*) yang dihasilkan menggunakan fungsi "`Conjoint`". Berikut adalah perintah dan luaran yang dihasilkan dari perintah yang bersesuaian tersebut.

```
#Utilitas level-atribut berdasarkan responden pertama
Conjoint(y=pref[1,], x=prof, z=level)
> Conjoint(y=pref[1,], x=prof, z=level)
[1] "Part worths (utilities) of levels (model parameters for whole
sample):"
            levnms    utls
1        intercept  2,6667
2     Asynchronous       0
3      Synchronous       0
4            Mixed       0
5    Pilihan Ganda -0,6667
6             Esai -0,3333
7           Proyek       1
8         Kelompok       0
9         Individu       0
10            Zoom -0,3333
11 Microsoft Teams  0,3333
12     Google Meet       0
```

Agar lebih mudah dalam menginterpretasikan luaran (*output*) di atas, nilai *utility* dari setiap level-atribut dapat disajikan dalam bentuk gambar (lihat Gambar 8.7). Gambar 8.7 menunjukkan bahwa responden pertama lebih mementingkan atribut evaluasi bentuk proyek, menggunakan Microsoft Teams, jenis penyampaian materi secara *Mixed*, dan penugasannya dikerjakan secara berkelompok.



Gambar 8.7 *Part-worth function* atribut pembelajaran *online* dari responden pertama

Kita juga dapat mengetahui nilai utilitas (*utility*) dari sembilan kombinasi level-atribut menggunakan fungsi "`caTotalUtilities`". Dengan menggunakan fungsi tersebut dan menggunakan perintah yang bersesuaian dengan fungsi tersebut dapat diperoleh luaran yang diinginkan ebagai berikut.

```
#Total utilitas responden pertama untuk setiap kombinasi
level-atribut
total.utilitas <- caTotalUtilities(y=pref[1,], x=prof)
colnames(total.utilitas)<- paste('Prof',1:9, sep = '')
total.utilitas
> total.utilitas
     Prof1 Prof2 Prof3 Prof4 Prof5 Prof6 Prof7 Prof8 Prof9
[1,] 1.667 3.333     2 2.333 2.667     4 2.333 3.667     2
```

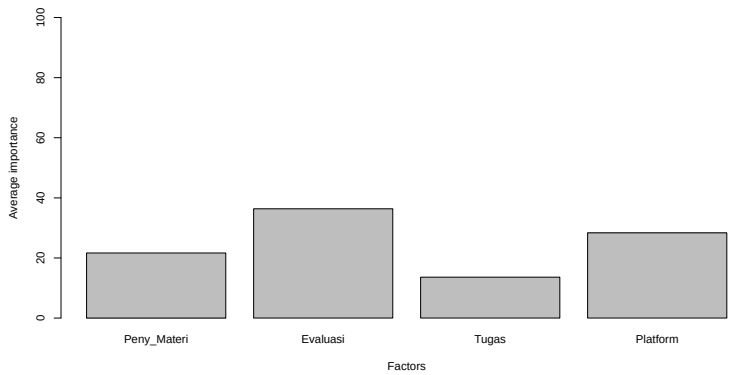
Dari luaran (*output*) di atas, kita dapat mengetahui bahwa responden pertama memiliki nilai utilitas (*utility*) tertinggi pada profil 6 (Prof6), yakni pembelajaran *online* dengan jenis *asynchronous*, evaluasi berbasis proyek, penugasannya secara individu, dan menggunakan Microsoft Teams. Selanjutnya kita dapat melakukan analisis konjoin untuk semua responden menggunakan perintah-perintah sebelumnya dengan sedikit modifikasi sebagai berikut.

```
#Importance untuk setiap atribut
nama.atribut <- colnames(atribut)
data.import <- caImportance(y=pref, x=prof)
importance <- data.frame(Atribut = nama.atribut, Importance
= data.import)
importance
> importance
      Atribut Importance
1 Peny_Materi      21.66
2    Evaluasi      36.36
3       Tugas      13.61
4    Platform      28.37
```

Luaran (*output*) di atas menunjukkan bahwa secara keseluruhan responden lebih mementingkan atribut evaluasi (36,36%) diikuti tiga atribut lainnya yang secara berturut-turut yaitu jenis *platform* yang

digunakan (28,37%), mode penyampaian materi (21,66%), dan bentuk penugasan (13,61%). Selanjutnya, *importance value* dari masing-masing atribut divisualisasikan seperti yang telah disajikan pada Gambar 8.8 yang didapat dengan menggunakan fungsi "`Conjoint`".

```
Conjoint(y=pref, x=prof, z=level)
```

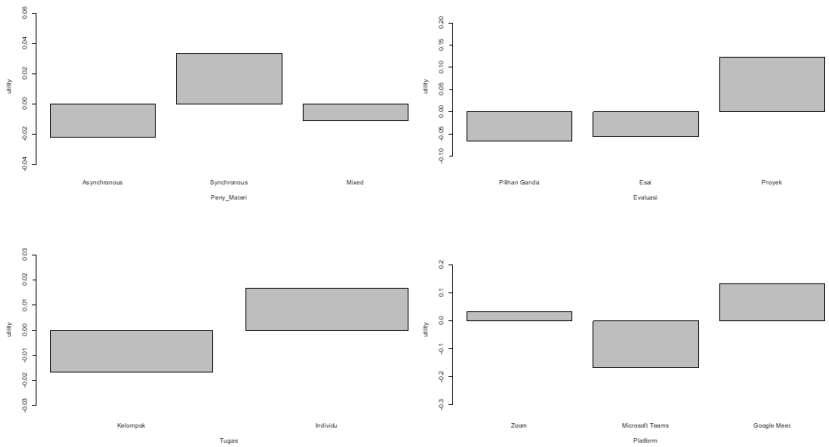


Gambar 8.8 *Importance value* atribut pembelajaran *online* semua responden

Kita juga dapat mengetahui lebih spesifik nilai utilitas (*utility)* dari masing-masing level atribut untuk semua responden berdasarkan luaran (*output*) yang dihasilkan dari fungsi "`Conjoint`". Ini dapat dilakukan dengan menggunakan perintah sebagai berikut.

```
#Utilitas untuk semua level atribut
Conjoint(y=pref, x=prof, z=level)
> Conjoint(y=pref, x=prof, z=level)
[1] "Part worths (utilities) of levels (model parameters for whole sample):"
            levnms    utls
1        intercept    3,45
2     Asynchronous -0,0222
3      Synchronous  0,0333
4            Mixed -0,0111
5    Pilihan Ganda -0,0667
6             Esai -0,0556
7           Proyek  0,1222
8         Kelompok -0,0167
9         Individu  0,0167
10            Zoom  0,0333
11 Microsoft Teams -0,1667
12     Google Meet  0,1333
```

Agar lebih mudah dalam menginterpretasikan luaran (*output*) di atas, nilai utilitas (*utility*) dari setiap level-atribut dapat disajikan dalam bentuk gambar (lihat Gambar 8.9). Gambar 8.9 menunjukkan bahwa, secara keseluruhan, responden sangat mementingkan atribut evaluasi berbasis proyek, *platform* yang digunakan berupa Google Meet, penyampaian materi dengan mode *synchronous*, dan bentuk penugasan yang dikerjakan secara individu.



Gambar 8.9 *Part-worth function* atribut pembelajaran *online* semua responden

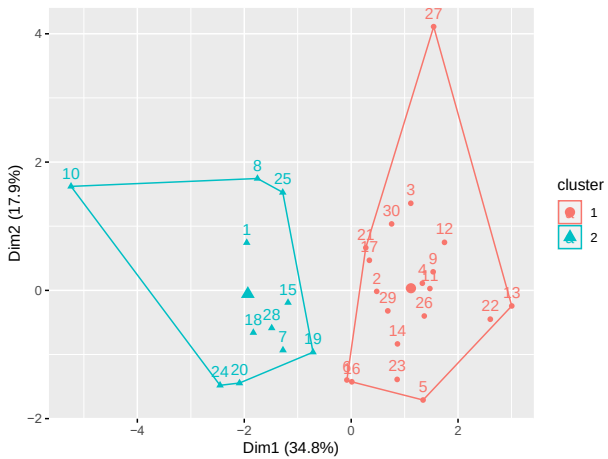
Kita juga dapat mengetahui nilai utilitas (*utility*) dari sembilan kombinasi level-atribut (setiap profil atau butir penilaian) pada setiap responden melalui fungsi "`caTotalUtilities`" sebagai berikut.

```
#Total utilitas untuk kombinasi level-atribut
total.utilitas <- caTotalUtilities(y=pref, x=prof)
colnames(total.utilitas)<- paste('Prof',1:9, sep = '')
head(total.utilitas)
tail(total.utilitas)
> head(total.utilitas)
     Prof1 Prof2 Prof3 Prof4 Prof5 Prof6 Prof7 Prof8 Prof9
[1,] 1.667 3.333     2 2.333 2.667     4 2.333 3.667     2
[2,] 4.000 4.000     3 3.000 3.000     4 4.000 4.000     4
[3,] 3.167 4.833     5 3.833 4.167     4 2.833 4.167     3
[4,] 3.167 3.833     4 3.833 4.167     4 3.833 4.167     4
[5,] 5.000 3.000     4 5.000 4.000     3 4.000 3.000     5
[6,] 3.667 2.333     3 3.333 2.667     3 4.333 3.667     4
```

```
> tail(total.utilitas)
      Prof1 Prof2 Prof3 Prof4 Prof5 Prof6 Prof7 Prof8 Prof9
[25,] 3.000 4.000     2 2.000 4.000     2 4.000 3.000     2
[26,] 4.167 3.833     3 2.833 3.167     5 4.833 4.167     5
[27,] 2.667 5.333     5 1.333 4.667     5 5.333 4.667     1
[28,] 3.000 4.000     1 3.000 2.000     4 2.000 4.000     4
[29,] 4.000 4.000     2 4.000 4.000     3 4.000 4.000     4
[30,] 3.000 4.000     4 3.000 4.000     4 4.000 4.000     3
```

Berdasarkan nilai utilitas (*utility*) level-atribut (profil atribut) pada setiap responden, kita selanjutnya dapat melakukan segmentasi atau pengelompokan mahasiswa menggunakan metode *k-means*. Dalam ilustrasi ini mahasiswa dikelompokkan ke dalam dua klaster (lihat Gambar 8.10).

```
#Segmentasi atau klasterisasi
klaster <- caSegmentation(y=pref, x=prof, c=2)
klaster$segm$centers
> klaster$segm$centers
   [,1]  [,2]  [,3]  [,4]  [,5]  [,6]  [,7] [,8]  [,9]
1 3.702 3.772 4.000 3.667 3.754 3.579 4.088 4.07 3.789
2 2.818 3.364 2.455 2.364 2.364 3.091 2.455 3.00 3.182
Untuk menyajikan hasil pengelompokan dalam bentuk visual, kita dapat
menggunakan fungsi “fviz_cluster”. Namun, untuk bisa menjalakan fungsi
tersebut maka kita perlu menjalankan paket “factoextra”.
#Plot Klasterisasi
library(factoextra)
fviz_cluster(klaster$segm, klaster$util)
```



Gambar 8.10 Segmentasi mahasiswa berdasarkan nilai utilitas (*utility*) profil atribut

Langkah berikutnya yaitu melakukan simulasi. Berdasarkan analisis konjoin yang telah dilakukan, kita dapat membuat simulasi tiga model pembelajaran *online* (lihat Tabel 8.2). Tiga orang responden tambahan diminta untuk memberikan penilaian terhadap kedua model pembelajaran tersebut. Dengan bantuan program R dan RStudio, kita dapat melakukan simulasi tersebut dengan mudah melalui fungsi "`ShowAllSimulations`".

Tabel 8.2 Profil model pembelajaran *online* yang disimulasikan

Pembelajaran *online*	Penyampaian materi	Evaluasi	Tugas	*Platform*
Model A	*Asynchronous* (1)	Pilihan ganda (1)	Kelompok (1)	Microsoft Teams (2)
Model B	*Synchronous* (2)	Proyek (3)	Individu (2)	Google Meet (3)
Model C	*Mixed* (3)	Esai (2)	Individu (2)	Zoom (1)

Sebelum memulai proses simulasi, kita perlu membuat data profil model pembelajaran *online* sesuai dengan petunjuk pada Tabel 8.2. Dengan memperhatikan Tabel 8.2, kita membuat profil model pembelajaran *online* dengan menggunakan perintah berikut sehingga terbentuk profil model pembelajaran *online* sesuai dengan yang dikehendaki.

```
#Membuat profil model pembelajaran online
model.a <- cbind(1,1,1,2)
model.b <- cbind(2,3,2,3)
model.c <- cbind(3,2,2,1)
prof.simul <- data.frame(rbind(model.a,model.b,model.c))
colnames(prof.simul)<-c('Peny_Mat','Evaluasi','Tugas','Platform')
rownames(prof.simul)<-c('Model_A','Model_B','Model_C')
prof.simul
> prof.simul
        Peny_Mat Evaluasi Tugas Platform
Model_A        1        1     1        2
Model_B        2        3     2        3
Model_C        3        2     2        1
```

Kita selanjutnya juga perlu memanggil data preferensi tambahan yang sengaja tidak diikutkan dalam analisis konjoin sebelumnya. Datanya tersebut merupakan "Data Simulasi". Berikut ini merupakan perintah untuk memanggil data preferensi tambahan dan luaran dari data preferensi tambahan tersebut.

```
#Memanggil data prefrensi tambahan
pref.simul <- read.csv('Data Simulasi.csv', sep = ';',
header = T)
pref.simul
> pref.simul
  p1 p2 p3 p4 p5 p6 p7 p8 p9
1  4  4  2  4  4  3  4  4  4
2  3  3  4  3  4  4  4  4  3
3  2  4  3  2  3  4  3  4  4
4  3  4  3  4  4  3  4  4  4
5  3  3  3  4  3  3  3  3  4
```

Setelah data profil simulasi dan preferensi tambahan sudah siap, langkah selanjutnya yaitu melakukan simulasi pangsa pasar (*market share*). Berikut ini merupakan perintah yang digunakan untuk melakukan simulasi dan luaran yang diperoleh dari perintah atau simulasi tersebut.

```
#Melakukan simulasi
simulasi <- ShowAllSimulations(sym =prof.simul,
y=pref.simul, x=prof)
rownames(simulasi)<-c('Model_A','Model_B','Model_C')
simulasi
> simulasi
        TotalUtility MaxUtility BTLmodel LogitModel
Model_A          3.3         40    32.17      31.39
Model_B          4.0         40    38.82      47.73
Model_C          3.0         20    29.01      20.89
```

Berdasarkan luaran (*output*) dari simulasi di atas maka diperoleh informasi bahwa pembelajaran *online* dengan Model B adalah yang paling disukai berdasarkan aturan *share utility*. Hal ini dapat dilihat dari nilai *total utility* Model B yang lebih besar dibanding dua model

lainnya. Jika kita menggunakan aturan *maximum utility*, maka Model A dan Model B memperoleh tingkat kesukaan yang sama. Hal ini dapat dilihat dari nilai *maximum utility* Model A dan B sebesar 40, dan ini lebih besar dari Model C. Selanjutnya, jika kita menggunakan aturan *logit model*, maka Model B merupakan model pembelajaran *online* yang dipilih dibanding dua model lainnya. Dengan demikian, kita dapat menyimpulkan bahwa mahasiswa sangat menyukai model pembelajaran *online* dengan jenis penyampaian materinya dilakukan secara *synchronous*, evaluasinya berbasis proyek, penugasannya dikerjakan secara individu, dan memanfaatkan *platform* untuk pembelajaran *online* berupa Google Meet.

Bab 9
Korelasi Kanonis

Analisis korelasi kanonis dilakukan dengan usaha untuk mengidentifikasi dan menguantifikasi asosiasi atau hubungan antara dua grup variabel. Analisis ini pada dasarnya hampir sama dengan analisis korelasi lainnya yang dilakukan untuk mengetahui hubungan antar variabel. Analisis korelasi kanonis namun demikian lebih berfokus pada dua grup variabel yang masing-masing grupnya terdiri atas beberapa variabel. Ketika ingin meneliti hubungan beberapa variabel dengan beberapa variabel lainnya, analisis korelasi kanonis lebih tepat digunakan dibandingkan analisis korelasi bivariat biasa.

Beberapa contoh penelitian menggunakan analisis korelasi kanonis, di antaranya yaitu penelitian yang dilakukan Srinadi et al. (2014) yang berfokus untuk menyelidiki hubungan antara kepemimpinan atasan dengan motivasi kerja karyawan. Kelompok variabel kepemimpinan atasan terdiri atas sembilan variabel dan kelompok variabel motivasi kerja karyawan terdiri atas enam variabel. Hasil penelitian menunjukkan bahwa faktor perilaku pemimpin yang paling berkorelasi dengan motivasi kerja karyawan yaitu kemampuan mengarahkan dan menghadapi karyawan. Contoh kedua yaitu studi yang dilakukan oleh Walag et al. (2022) yang berfokus pada penyelidikan hubungan literasi sains guru mata pelajaran sains dengan efikasi pengajaran mata pelajaran sains di Filipina. Hasil analisis korelasi kanonis dari enam variabel *scientific literacy* (literasi sains) dan enam variabel untuk *science teaching efficacy* (efikasi pengajaran sains) menunjukkan hubungan positif dan moderat pada variabel literasi sains dan efikasi pengajaran sains. Contoh selanjutnya yaitu studi yang dilakukan oleh Irianingsih et al. (2016) yang ditujukan untuk mengetahui korelasi kanonis perilaku belajar terhadap prestasi belajar siswa SMP dengan studi kasus pada SMPN 1 Sukasari, Pur-

wakarta). Kelompok variabel perilaku belajar terdiri atas variabel intensitas belajar mandiri di luar jam sekolah dan intensitas belajar kelompok. Kelompok variabel prestasi belajar siswa SMP adalah hasil rapor mata pelajaran Matematika, IPA, IPS, Bahasa Indonesia, dan Bahasa Inggris. Hasil dari analisis korelasi kanonis mengindikasikan bahwa perilaku belajar dan prestasi belajar mempunyai keterkaitan yang cukup kuat, namun korelasi terbesar terjadi pada intensitas belajar mandiri di luar jam sekolah pada nilai IPA siswa.

Teori dasar pada analisis korelasi kanonis

Konsep dasar pada analisis korelasi kanonis

Gudono (2017) menjelaskan bahwa analisis korelasi kanonis merupakan studi mengenai hubungan antara sekelompok prediktor (variabel independen) dengan sekelompok variabel dependen atau studi mengenai dua pasang vektor. Secara sederhana, dapat dipahami bahwa analisis korelasi kanonis digunakan untuk menguji korelasi antara lebih dari satu variabel dependen dengan lebih dari satu variabel independen. Hal ini tentu berbeda dengan regresi ganda. Analisis regresi ganda hanya menganalisis satu variabel dependen dengan lebih dari satu variabel independen. Analisis korelasi kanonis bertujuan untuk menentukan sifat hubungan antara dua kelompok variabel, jumlah hubungan yang secara statistik signifikan antara dua kelompok variabel, sejauh mana variansi kelompok variabel dependen bergantung pada variabel independen, serta bobot yang menentukan peran sebuah variabel sehingga korelasi antara dua kelompok variabel memiliki korelasi yang tertinggi.

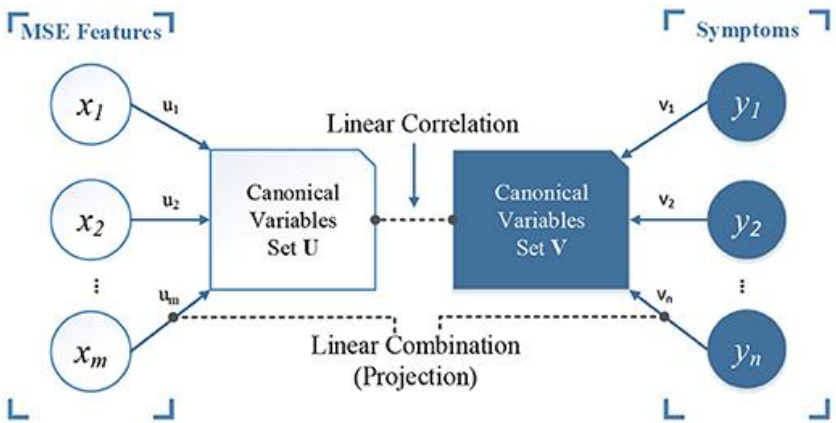
Model persamaan pada analisis korelasi kanonis

Analisis korelasi kanonis menurut Johnson dan Wichern (2007) berfokus pada korelasi linear antara kombinasi variabel dalam satu grup atau himpunan, serta kombinasi linear dari variabel pada grup variabel yang lain. Analisis korelasi kanonis diawali dengan menentukan pasangan kombinasi linear yang memiliki korelasi terbesar, kemudian menentukan pasangan kombinasi yang memiliki kombinasi korelasi linear yang terbesar di antara semua pasangan yang mungkin

ada antar dua grup variabel. Pasangan yang memiliki kombinasi linear ini disebut variabel kanonis, dan korelasinya disebut sebagai korelasi kanonis. Aspek maksimalisasi dari analisis yang dilakukan merupakan upaya untuk menunjukkan pemusatan hubungan antara dua grup variabel yang berdimensi tinggi (*high-dimensional*) menjadi beberapa pasang variabel kanonis. Untuk menggambarkan model bivariat biasa (seperti korelasi Pearson) dapat ditulis sebagai $r: X = Y$, sementara untuk model korelasi kanonikas, model persamaannya dapat ditulis sebagai berikut.

$$R_C = X_1 + X_2 + \cdots + X_n = Y_1 + Y_2 + Y_3 + \cdots + Y_q$$

Korelasi kanonis dapat dimodelkan dalam bentuk sebagaimana yang ditunjukkan oleh Gambar 9.1, dimana U (dapat juga dipandang sebagai X) dan V (dapat juga dipandang sebagai Y) merupakan skor komposit yang merupakan hasil pembobotan dari variabel-variabel pembentuknya. Bobot tetap dihitung agar korelasi antara skor X (U) dan Y (V) dapat maksimal.



Gambar 9.1 Model analisis korelasi kanonis
(sumber: Fan et al., 2018, p. 4, https://doi.org/10.3389/fnins.2018.00685)

Asumsi-asumsi pada analisis korelasi kanonis

Kurniawan dan Yuniarto (2016) menyebutkan bahwa analisis korelasi kanonis tepat diterapkan bila beberapa hal berikut terpenuhi.

- Variabel dari masing-masing grup variabel berskala rasio atau interval. Apabila variabel yang akan dihitung korelasinya memiliki

skala nominal, maka skala variabel tersebut perlu diubah terlebih dahulu menjadi skala buatan (*dummy*).

- Adanya hubungan yang bersifat linear antara dua grup variabel.
- Data memenuhi distribusi normal multivariat. Hal ini diperlukan untuk menguji signifikansi setiap fungsi kanonis. Pemeriksaan asumsi multivariat normal dapat dilakukan dengan analisis grafik dan tes statistik dengan nilai *kurtosis* dan *skewness.*
- Tidak ada multikolinearitas antara anggota grup variabel X dan Y.
- Sebaiknya menghindari *outliers*, karena dapat menyebabkan bias dalam penentuan *canonical weight*.
- Ukuran sampel setidaknya sejumlah 20 kali jumlah variabel.

Berdasarkan sejumlah asumsi tersebut, asumsi berupa adanya hubungan linear antara dua grup variabel, data memenuhi distribusi normal multivariat, dan tidak ada multikolinearitas antara anggota grup variabel X dan Y dianggap sebagai asumsi krusial yang harus dipenuhi untuk melakukan analisis korelasi kanonis.

Prosedur analisis korelasi kanonis

Berikut adalah langkah-langkah yang dapat diikuti dalam melakukan analisis korelasi kanonis.

1. Menentukan tujuan dan spesifikasi masing-masing grup variabel. Data yang baik yaitu data dari grup variabel baik metrik maupun non-metrik. Diasumsikan bahwa setiap grup variabel dapat diartikan secara teoritis.
2. Menentukan jumlah observasi pada setiap grup variabel dan total ukuran sampel. Terlalu kecil ukuran sampel tidak akan merepresentasikan variabel dengan baik. Demikian juga ukuran sampel yang besar akan memiliki kecenderungan signifikansi statistik dalam segala hal, namun secara praktik tidak mengindikasikan suatu signifikansi.
3. Melakukan uji asumsi. Uji asumsi yang perlu dilakukan meliputi asumsi linearitas, normalitas multivariat, homoskedastisitas, dan non-multikolinearitas. Pada tahap ini dilakukan penurunan fungsi kanonis, pemilihan fungsi untuk penginterpretasian (signifikansi statistik, besarnya hubungan). Maksimum fungsi kanonis yang

terbentuk adalah nilai minimum dari jumlah variabel dalam setiap grup variabel.

4. Melakukan interpretasi dari fungsi dan variabel kanonis dengan menggunakan pembobotan kanonis (*canonical weight*), muatan kanonis (*canonical loading*), dan muatan silang kanonis (*canonical cross-loading*).
5. Melakukan validasi atas hasil luaran dari analisis korelasi kanonis tersebut. Validasi biasanya dilakukan dengan membagi dua bagian sampel, kemudian membandingkan hasil yang ada. Bila perbedaan hasil kedua sampel tidak besar, maka bisa dikatakan bahwa hasil analisis korelasi kanonis tersebut valid.

Contoh kasus dan analisis korelasi kanonis menggunakan program R dan RStudio

Contoh kasus

Nilai mata kuliah praktikum biasanya terdiri dari nilai *pretest*, unjuk kerja, laporan, dan responsi. Nilai praktikum yang diraih mahasiswa tentu dipengaruhi oleh faktor internal maupun eksternal. Faktor internal antara lain minat, motivasi, dan kesiapan belajar. Tabel 9.1 merangkum grup variabel dan variabel-variabel yang menjadi fokus pada contoh kasus ini. Pada kasus ini, ada tiga jenis variabel independen dan empat jenis variabel dependen. Dari kasus yang diberikan tersebut, analisis korelasi kanonis dilakukan untuk mengungkap hubungan antara kelompok variabel dan variabel-variabel yang ada.

Tabel 9.1 Variabel-variabel yang menjadi fokus dalam contoh kasus

Variabel	Jenis variabel	Keterangan
X1	Variabel independen 1	Minat
X2	Variabel independen 2	Motivasi
X3	Variabel independen 3	Kesiapan belajar
Y1	Variabel dependen 1	Nilai *pretest*
Y2	Variabel dependen 2	Nilai laporan
Y3	Variabel dependen 3	Nilai unjuk kerja
Y4	Variabel dependen 4	Nilai responsi

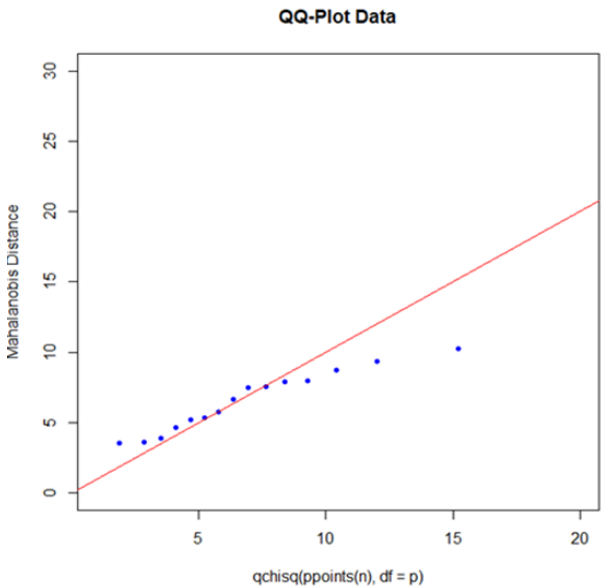
Prosedur analisis

Beberapa paket diperlukan untuk melakukan analisis korelasi kanonis. Paket-paket yang diperlukan tersebut, dan seharusnya dipastikan sudah terpasang, terdiri atas 'openxlsx', 'CCA', 'candisc', 'readr', dan 'car'. Setelah semua paket yang diperlukan tersebut terpasang, semua paket tersebut dapat digunakan atau dijalankan dengan cara menggunakan perintah atau fungsi `library()` sebagai berikut.

```
> library (openxlsx)
> library (CCA)
> library (candisc)
> library (readr)
> library (car)
```

Berikut langkah-langkah untuk melakukan analisis korelasi kanonis yang dimulai dari memanggil data yang dimiliki ke RStudio.

```
#Impor data
data = read.xlsx("C://Users/acer/Downloads//data simulasi kanonik.xlsx")
#nama file dan lokasi file disesuaikan
view(data)
head(data) #untuk menampilkan 6 baris data pertama
   minat motivasi kesiapan.belajar pretest laporan unjuk.kerja responsi
1    36       34               30      75      85          80       80
2    30       30               30      70      85          75       80
3    30       30               30      70      80          75       75
4    32       34               28      60      80          80       70
5    25       28               24      60      75          65       70
6    38       36               30      70      75          80       70
# Mendefinisikan variabel x dan y, kemudian digabungkan
x <- data[,1:3] #disesuaikan dengan jumlah variabel x
y <- data[,4:7] #disesuaikan dengan jumlah variabel y
datajoin <- cbind(y,x)
# Uji asumsi multivariat normal
#membentuk matriks n x p
xx <- as.matrix(datajoin)
#titik pusat
center <- colMeans(xx)
n <- nrow(datajoin)
p <- ncol(datajoin)
cov <- cov(xx)
#jarak mahalanobis
d <- mahalanobis(xx,center,cov)
#membuat Qq-Plot Data
qqplot(qchisq(ppoints(n), df = p), d, xlim = c(1,20), pch = 20,
       col = "blue", ylim = c(0,30), main = "QQ-Plot Data",
       ylab = "Mahalanobis Distance")
abline(a = 0, b = 1, col = "red")
```



Gambar 9.2 Q-Q plot untuk asumsi normal multivariat

Gambar 9.2 menunjukkan Q-Q plot sebaran data yang digunakan dalam studi pada contoh kasus yang diberikan. Gambar 9.2 secara deskriptif mengindikasikan bahwa data menyebar normal ganda karena plot datanya masih berada di sekitar garis lurus. Dengan demikian, dapat dikatakan bahwa asumsi normal multivariat terpenuhi. Selanjutnya yaitu membuktikan apakah asumsi multikolinearitas terpenuhi juga melalui perintah berikut.

```
# Uji multikolinearitas
model1=lm(pretest+laporan+unjuk.kerja+responsi~minat+motivasi+kesia
pan.belajar, data = datajoin)
vif(model1)

   minat         motivasi kesiapan.belajar
4.665669         4.040135         1.572966
```

Nilai VIF yang diperoleh dari uji multikolinearitas di atas untuk variabel-variabel di grup variabel *X* di bawah 10. Hasil ini memiliki arti bahwa tidak terjadi multikolinearitas antar variabel pada grup

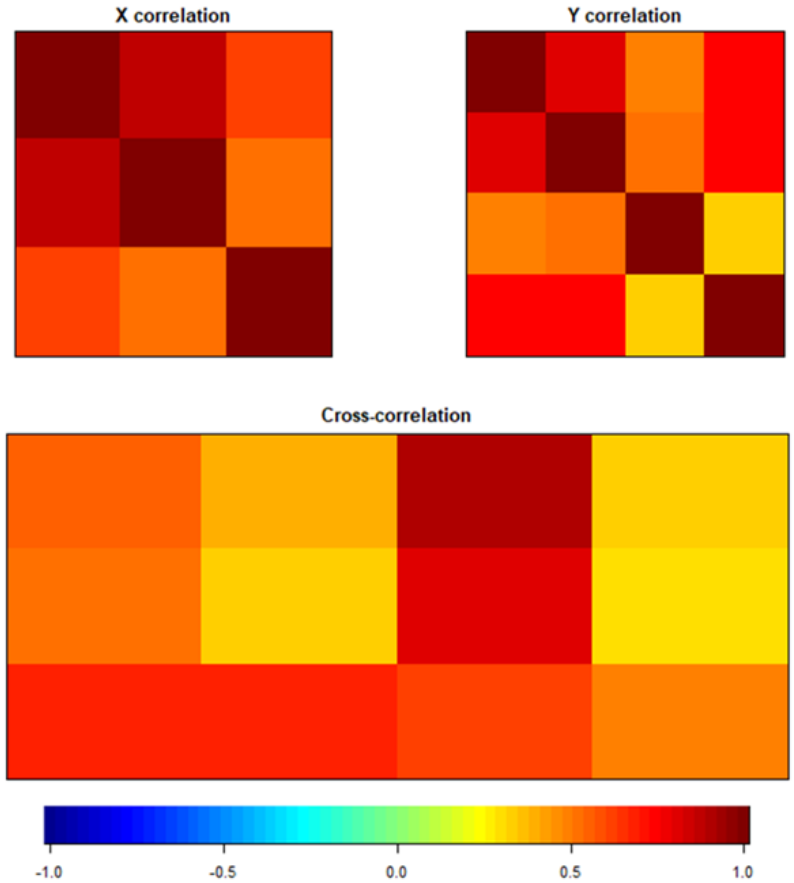
variabel *X* (variabel prediktor). Langkah berikutnya yaitu membentuk matriks korelasi sebagai berikut.

```
# Matrik korelasi
correlation <- matcor(x,y)
Correlation

$Xcor
                     minat   motivasi   kesiapan.belajar
minat            1.0000000  0.8673996   0.6033226
motivasi         0.8673996  1.0000000   0.5152922
kesiapan.belajar 0.6033226  0.5152922   1.0000000

$Ycor
              pretest   laporan  unjuk.kerja   responsi
pretest     1.0000000 0.7941014   0.4906293   0.7234710
laporan     0.7941014 1.0000000   0.5238227   0.7402332
unjuk.kerja 0.4906293 0.5238227   1.0000000   0.3425134
responsi    0.7234710 0.7402332   0.3425134   1.0000000

$XYcor
              minat  motivasi kesiapan.belajar   pretest   laporan
minat      1.0000000 0.8673996        0.6033226 0.5523652 0.3940586
motivasi   0.8673996 1.0000000        0.5152922 0.5268316 0.3317156
Kesiapan   0.6033226 0.5152922        1.0000000 0.6825812 0.6869468
pretest    0.5523652 0.5268316        0.6825812 1.0000000 0.7941014
laporan    0.3940586 0.3317156        0.6869468 0.7941014 1.0000000
unjuk.kerja0.8879546 0.7830791        0.5981313 0.4906293 0.5238227
responsi   0.3239353 0.2848344        0.4712108 0.7234710 0.7402332

                  unjuk.kerja  responsi
minat               0.8879546  0.3239353
motivasi            0.7830791  0.2848344
kesiapan.belajar    0.5981313  0.4712108
pretest             0.4906293  0.7234710
laporan             0.5238227  0.7402332
unjuk.kerja         1.0000000  0.3425134
responsi            0.3425134  1.0000000
```

Setelah membentuk matriks korelasi, baik antara variabel-variabel dalam grup variabel *X*, antara variabel-variabel dalam grup variabel *Y*, dan secara silang antara variabel-variabel dalam grup variabel *X* dengan mereka dalam grup variabel *Y*, langkah yang dapat dilakukan selanjutnya yaitu memvisualisasikan korelasi-korelasi terse-

but. Hal ini dapat dilakukan dengan menggunakan perintah berikut, sehingga dapat diperoleh hasil sebagaimana yang ditunjukkan pada Gambar 9.3. Gambar 9.3 menunjukkan bahwa korelasi antara X dan Y menuju arah yang positif. Hal ini mengindikasikan bahwa semakin tinggi nilai X maka akan semakin tinggi pula nilai Y.

```
#Visualisai korelasi untuk variabel X, Y, dan XY
img.matcor(correlation, type = 2)
```



Gambar 9.3 Visualisasi korelasi

Setelah diperoleh matriks korelasi dan visualisasi dari korelasi tersebut, langkah selanjutnya yaitu melakukan analisis korelasi ka-

nonis. Berikut adalah perintah untuk melakukan analisis korelasi kanonis dan hasil yang diperoleh dari analisis tersebut.

```
#Analisis korelasi kanonis
ccan <- candisc::cancor(x,y)
summary(ccan)

Canonical correlation analysis of:
3   X  variables:  minat, motivasi, kesiapan.belajar
with 4   Y  variables:  pretest, laporan, unjuk.kerja, responsi

    CanR  CanRSQ   Eigen percent    cum                            scree
1 0.9396 0.88287 7.53767 91.4903  91.49 ****************************
2 0.6341 0.40211 0.67255  8.1632  99.65 ***
3 0.1666 0.02775 0.02855  0.3465 100.00

Test of H0: The canonical correlations in the
current row and all that follow are zero

     CanR LR test stat approx F numDF  denDF  Pr(> F)
1 0.93961      0.06809   3.1488    12 21.457 0.009993 **
2 0.63412      0.58130   0.9348     6 18.000 0.494291
3 0.16659      0.97225      NaN     2    NaN      NaN
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Raw canonical coefficients

   X  variables:
                    Xcan1     Xcan2     Xcan3
minat            -0.215701 -0.202148  0.499771
motivasi         -0.077798 -0.022868 -0.860396
kesiapan.belajar -0.017495  0.371485 -0.015857

   Y  variables:
                Ycan1     Ycan2    Ycan3
pretest     -0.0413846  0.011974 -0.13212
laporan      0.1006401  0.282119  0.22103
unjuk.kerja -0.2107418 -0.129885  0.10900
responsi    -0.0021108 -0.085751  0.07851

#Plot korelasi
res.cc <- cc(x,y)
plot(res.cc$cor, type = "b")

# Menampilkan nilai korelasi kanonis dan menampilkan koefisien estimasi
variabel X dan Y
res.cc$cor
res.cc$xcoef
res.cc$ycoef
res.cc$cor
[1] 0.9396127 0.6341215 0.1665920
```

```
# Menampilkan nilai korelasi kanonis dan menampilkan koefisien estimasi
variabel X dan Y
res.cc$cor
res.cc$xcoef
res.cc$ycoef

res.cc$cor
[1] 0.9396127 0.6341215 0.1665920

#Menampilkan penaksir koefisien bagi peubah x dan y
> res.cc$xcoef
                        [,1]       [,2]        [,3]
minat            -0.21570071 -0.2021480  0.49977052
motivasi         -0.07779806 -0.0228685 -0.86039553
kesiapan.belajar -0.01749516  0.3714845 -0.01585678
> res.cc$ycoef
                    [,1]        [,2]        [,3]
pretest     -0.041384571  0.01197425 -0.13211621
laporan      0.100640127  0.28211946  0.22102610
unjuk.kerja -0.210741781 -0.12988482  0.10900460
responsi    -0.002110791 -0.08575117  0.07851022
```

Luaran (*output*) di atas menunjukkan koefisien kanonis variabel X dan Y yang menampilkan tiga variat. Perhatikan variat dengan nilai korelasi kanonis terbesar, yaitu [1] sebesar 0.94. Dapat dilihat bahwa variabel X yang berkontribusi terbesar hingga terkecil adalah minat [$X1$], motivasi [$X2$], dan kesiapan belajar [$X3$]. Sedangkan untuk variabel Y yang berkontribusi terbesar sampai terkecil adalah unjuk kerja [$Y3$], laporan [$Y2$], pretest [$Y1$], dan responsi [$Y4$].

```
#Menampilkan koordinat variat kanonis
res.cc$scores
$xscores
          [,1]       [,2]         [,3]
1  -1.26316067 -0.6950017  0.237748194
2   0.34223580  0.6093604  0.680707185
3   0.34223580  0.6093604  0.680707185
4  -0.36536752 -0.6293786 -1.729620313
5   1.68130643 -0.5630697 -0.002213652
6  -1.85015820 -1.1450347 -0.483501820
7  -0.40035784  0.1135904 -1.761333882
8   0.34223580  0.6093604  0.680707185
9   0.48219710 -2.3625158  0.807561459
10 -1.36813165  1.5339056  0.142607488
11 -0.74614380  1.2409695  0.895571071
12  0.80862754  0.2706874 -0.287120285
13  1.24002895  0.6749835 -1.286661323
14 -0.05417529 -0.5379047  1.711961791
15  0.80862754  0.2706874 -0.287120285
```

```
$yscores
          [,1]        [,2]       [,3]
1  -0.97926374  0.28674204  0.6743745
2   0.28136802  0.87629489  0.7899326
3  -0.21127866 -0.10554653 -0.7077490
4  -0.84058790 -0.44595724  0.7658850
5   1.81733819  0.09171775 -1.9743145
6  -1.75763424 -1.73681204 -1.6604076
7   0.08499913  1.36492199 -0.2631996
8   0.09555308  1.79367784 -0.6557507
9   0.12376608 -1.32687291  0.4368936
10 -1.17563263  0.77536914 -0.3787576
11 -0.96870978  0.71549790  0.2818234
12  0.62696672  0.08372437  1.5420241
13  1.17747499 -0.67744882 -0.1081294
14  0.54816575 -1.01785954  1.3655046
15  1.17747499 -0.67744882 -0.1081294

$corr.X.xscores
                       [,1]       [,2]       [,3]
minat            -0.9949141  -0.04296700   0.09110342
motivasi         -0.9073202  -0.06049023  -0.41606613
kesiapan.belajar -0.6362916   0.77055828   0.03705358

$corr.Y.xscores
                 [,1]       [,2]         [,3]
pretest     -0.5787806 0.41096510 -0.063815877
laporan     -0.4167077 0.54585109  0.032093636
unjuk.kerja -0.8894456 0.03927644  0.051811586
responsi    -0.3392409 0.33104930  0.007067462

$corr.X.yscores
                       [,1]        [,2]        [,3]
minat            -0.9348339 -0.02724630  0.01517710
motivasi         -0.8525295 -0.03835816 -0.06931328
kesiapan.belajar -0.5978676  0.48862761  0.00617283

$corr.Y.yscores
                 [,1]       [,2]        [,3]
pretest     -0.6159779 0.64808570 -0.38306692
laporan     -0.4434888 0.86079884  0.19264814
unjuk.kerja -0.9466088 0.06193835  0.31100888
responsi    -0.3610434 0.52205970  0.04242378
```

Berdasarkan hasil analisis, terlihat bahwa variabel Y yang berhubungan erat dengan fungsi kanonis pertama adalah unjuk kerja $[Y_3]$. Sedangkan variabel X yang berhubungan erat dengan fungsi kanonis pertama adalah minat $[X_1]$ kemudian motivasi $[X_2]$. Untuk korelasi silang antar variabel-variabel respons terhadap fungsi kanonis variabel X yang berhubungan paling erat dengan fungsi kanonik pertama

adalah unjuk kerja [Y_3]. Sedangkan korelasi silang antar variabel X terhadap fungsi kanonis variabel Y yang berhubungan paling erat dengan fungsi kanonis pertama adalah minat [X_1].

Daftar Pustaka

Anton, H. (1997). *Aljabar linear elementer* (5th ed.). Erlangga.

Agresti, A. (2013). *Categorical data analysis* (3rd ed.). John Willey and Sons.

Agus, W. (2010). *Analisis statistika multivariat terapan* (1st ed.). UPP STIM YKPN.

Akintunde, O & Matthew, S. (2019). Discriminant analysis of psycho-social predictors of mathematics achievement of gifted students in Nigeria. *Journal for the Education of Gifted Young Scientists*, *7*(3), 581–594. https://doi.org/10.17478/jegys.605981

Almeira, D., & Juanda, G. G. (2021). Analisis multidimensional scaling dan k-means clustering untuk pengelompokan provinsi berdasarkan tingkat pengangguran. *Prosiding Seminar Nasional Statistika* (vol. 10, pp. 60–69). Universitas Padjadjaran, Indonesia. https://doi.org/10.1234/pns.v10i.75

Borg, I., & Groenen, P. J. F. (2005). *Modern multidimensional scaling: Theory and applications*. Springer Series in Statistics, Springer. https://doi.org/10.1007/0-387-28981-X

Borg, I., Groenen, P. J. F., & Mair, P. (2018). Applied multidimensional scaling and unfolding (2nd ed.). SpringerBriefs in Statistics, Springer. https://doi.org/10.1007/978-3-319-73471-2

de Leeuw, J., & Heiser, W. (1980). Theory of multidimensional scaling. In P. R. Krishnaiah & L. N. Kanal (Eds.), *Handbook of statistics* (Volume 2). North Holland Publishing Company.

del Prette, Z. A. P., Prette, A. del, de Oliveira, L. A., Gresham, F. M., & Vance, M. J. (2012). Role of social performance in predicting learning problems: Prediction of risk using logistic regression analysis. *School Psychology International*, *33*(6), 615–630. https://doi.org/10.1177/0020715211430373

Erimafa, J. T., Iduseri, A., & Edokpa, I. W. (2009). Application of discriminant analysis to predict the class of degree for graduating students in a university system. *International Journal of Physical Sciences*, *4*(1), 16–21.

Factor, E. M. R., & de Guzman, A. B. (2017). Explicating Filipino student nurses' preferences of clinical instructors'attributes: A conjoint analysis. *Nurse Education Today, 55*, 122–127.

Fuente-Mella, H. D. L., Umaña-Hermosilla, B., Fonseca-Fuentes, M., & Elórtegui-Gómez, C. (2021). Multinomial logistic regression to estimate the financial education and financial knowledge of university students in Chile. *Information*, *12*(9), 379.

Ginanjar, I. (n.d.). *Multidimensional Scaling (MDS)* [PowerPoint slides]. Academia.

Gracía-Pérez, M. L., & Gil-Lacruz, M. (2018). The impact of a continuing training program on the perceived improvement in quality of health care delivered by health care professionals. *Evaluation and program planning*, *66*, 33–38.

Gudono, G. (2011). *Analisis data multivariat* (1st ed.). BPFE.

Gudono, G. (2017). *Analisis data multivariat* (4th ed.). BPFE.

Gujarati, D. N., & Porter, D. C. (2009). *Basic econometric* (5th ed.). McGraw-Hill/Irwin

Hair, J. F., Anderson, R. E., Tatham, R. L., & Black, W. C. (1998). *Multivariate data analysis* (5th ed.). Pearson Education Prentice Hall.

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate data analysis* (7th ed.). Pearson.

Hosmer, David W., & Lemesbow, S. (1980). Goodness of fit tests for the multiple logistic regression model. *Communications in Statistics - Theory and Methods*, *9*(10), 1043–1069.

Huberty, C. J., & Olejnik, S. (2006). *Applied MANOVA and discriminant analysis* (2nd ed.). John Wiley & Sons.

Irianingsih, I., Gusriani, N., Kulsum, S., & Parmikanti, K. (2016). Analisis korelasi kanonik perilaku belajar terhadap prestasi belajar siswa SMP (Studi kasus siswa SMPN I Sukasari Purwakarta).

Prosiding Seminar Nasional Matematika dan Pendidikan Matematika (pp. 693–703). FKIP Universitas Sebelas Maret.

Johnson, R. A., & Wichern, D. W. (2007). *Applied multivariate statistical analysis* (6th ed.). Pearson Education.

Kruskal, J. B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika, 29*(1), 1–27.

Kurniawan, R., & Yuniarto, B. (2016). *Analisis regresi: Dasar dan penerapannnya dengan R*. Kencana

Lemeshow, S., & Hosmer Jr, D. W. (1982). A review of goodness of fit statistics for use in the development of logistic regression models. *American journal of epidemiology, 115*(1), 92–106.

Long, J. S. (1997). *Regression models for categorical and limited dependent variables.* Sage Publications.

Macindo, J. R. B., Danganan, M. P. B., Soriano, C. A. F., Kho, N. S. R., & Bongar, M. V. V. (2019). A conjoint analysis of the acute and critical care experiential learning preferences of Baccalaureate student nurses. *Nurse Education in Practice, 36*, 125–131.

Mahida, M. (2020). Pendekatan multidimensional scaling untuk penilaian status keberlanjutan ATCS kota pintar Semarang. *Warta Penelitian Perhubungan, 32*(2), 103–112.

Matahari, M., Kekenusa, J., & Langi, Y. (2015). Pengelompokan sekolah dasar di Siau berdasarkan indikator mutu sekolah dengan menggunakan analisis gerombol. *D'Cartesian: Jurnal Matematika dan Aplikasi, 4*(2), 188–195.

Motta, M. (2021). Can a COVID-19 vaccine live up to Americans' expectations? A conjoint analysis of how vaccine characteristics influence vaccination intentions. *Social Science & Medicine, 272*, 113642. https://doi.org/10.1016/j.socscimed.2020.113642

Nafkiyah, D., Rifatin, L., Rozikin, M. R., & Pramesti, W. (2022). Analisis cluster dalam pengelompokan kabupaten/kota di Provinsi Jawa Timur berdasarkan indikator pendidikan. *Buana Matematika: Jurnal Ilmiah Matematika dan Pendidikan Matematika, 12*(1), 1–16. https://doi.org/10.36456/buanamatematika.v12i1.6178

Ong, A. K. S., Prasetyo, Y. T., Pinugu, J. N. J., Chuenyindee, T., Chin, J., & Nadlifatin, R. (2022). Determining factors influencing students' future intentions to enroll in chemistry-related courses: integrating self-determination theory and theory of planned behavior. *International Journal of Science Education*, *44*(4), 556–578. https://doi.org/10.1080/09500693.2022.2036857

Ong, A. K. S., Prasetyo, Y. T., Chuenyindee, T., Young, M. N., Doma, B. T., Caballes, D. G., ... & Bautista, C. S. (2022). Preference analysis on the online learning attributes among senior high school students during the COVID-19 pandemic: A conjoint analysis approach. *Evaluation and Program Planning*, *92*, 1–9. https://doi.org/10.1016/j.evalprogplan.2022.102100

Pleger, L. E., Mertes, A., Rey, A., & Brüesch, C. (2020). Allowing users to pick and choose: A conjoint analysis of end-user preferences of public e-services. *Government Information Quarterly*, *37*(4), 1–11. https://doi.org/10.1016/j.giq.2020.101473

Putri, D. S., Wahyuningsih, S., & Goejantoro, R. (2018). Analisis positioning dengan menggunakan multidimensional scaling nonmetrik. *Jurnal Eksponensial*, *9*(1), 85–94.

Rencher, A. C. (1998). *Multivariat statistical inference and application*. John Willey & Sons.

Rencher, A. C. (2002). *Methods of multivariate analysis* (2nd ed.). John Wiley & Sons.

Ryan, R. M., & Deci, E. L. (2020). Intrinsic and extrinsic motivation from a self-determination theory perspective: Definitions, theory, practices, and future directions. *Contemporary Educational Psychology*, *61*, 1–11.

Şemin, F. K. (2020). Examining teachers' disposition towards sustainable education through discriminant analysis. *Research in Pedagogy*, *10*(2), 229–247.

Simamora, B. (2005). *Analisis multivariat pemasaran*. Gramedia Pustaka Utama.

Şirin, Y. E., & Şahin, M. (2020). Investigation of factors affecting the achievement of university students with logistic regression

analysis: School of physical education and sport example. *Sage Open*, *10*(1), 1–9. https://doi.org/10.1177/215824402090208

Srinadi, I. G. A. M., Asih, N. M., & Cahyani, A. D. (2014). Analisis korelasi kanonik hubungan perilaku pemimpin dan motivasi kerja karyawan. *Jurnal Matematika*, *4*(1), 51–62.

Supandi, E. D., Wardati, K., & Kuswidi, I. (2009). Aplikasi multidimensional scalling: studi kasus analisis segmentasi dan peta posisi UIN Sunan Kalijaga terhadap perguruan tinggi di Yogyakarta. *Prosiding Seminar Nasional Matematika dan Pendidikan Matematika* (pp. 599–622). Jurusan Pendidikan Matematika UNY.

Supranto, J. (2004). *Analisis multivariat arti dan interpretasi*. Rineka Cipta.

Wahyuni, T., Agoestanto, A., & Pujiastuti, E. (2018). Analisis regresi logistik terhadap keputusan penerimaan beasiswa PPA di FMIPA UNNES menggunakan software Minitab. *Prisma, Prosiding Seminar Nasional Matematika* (Vol. 1, pp. 755–764).

Walag, A. M. P., Fajardo, M. T. M., Bacarrisas, P. G., & Guimary, F. M. (2022). A canonical correlation analysis of Filipino science teachers' scientific literacy and science teaching efficacy. *International Journal of Instruction*, *15*(3), 249–266.

Wibowo, A. E., & Habanabakize, T. (2022). K-means clustering untuk klasifikasi standar kualifikasi pendidikan dan pengalaman kerja guru SMK di Indonesia. *Jurnal Dinamika Vokasional Teknik Mesin*, *7*(2). 152–163.

Witten, I. H., Frank, E., Hall, M. A. (2011). *Data mining: Pratical machine learning tools and techniques* (3rd ed.) Morgan Kaufmann Publishers.

Biodata Penulis

Heri Retnawati
Penulis merupakan Guru Besar dalam Bidang Ilmu Evaluasi Pendidikan Matematika, Universitas Negeri Yogyakarta (UNY), sejak tahun 2019. Heri menyelesaikan S1 Pendidikan Matematika di IKIP Yogyakarta tahun 1996, S2 Penelitian dan Evaluasi Pendidikan di Universitas Negeri Yogyakarta tahun 2004, dan S3 Penelitian dan Evaluasi Pendidikan di Universitas Negeri Yogyakarta tahun 2008. Sejak tahun 2000 hingga saat ini Heri Retnawati berprofesi sebagai dosen di UNY. Heri telah mempublikasikan beberapa naskah buku seperti Analisis Kuantitatif Instrumen Penelitian, Teori Respons Butir dan Penerapannya, dan Pengantar Analisis Meta. Heri memiliki minat penelitian dalam bidang asesmen dan pendidikan matematika. Berbagai karyanya telah terbit di jurnal internasional bereputasi.

Samsul Hadi
Penulis merupakan Guru Besar dalam bidang Evaluasi Pembelajaran Kejuruan, Universitas Negeri Yogyakarta (UNY). Penulis menyelesaikan studi S1 Pendidikan Teknik Elektro di IKIP Semarang pada tahun 1983, studi S2 Pendidikan Teknologi dan Kejuruan di IKIP Jakarta pada tahun 1991, studi S2 Teknik Elektro di Universitas Gadjah Mada (UGM) pada tahun 1999, dan studi S3 Penelitian dan Evaluasi Pendidikan di Universitas Negeri Yogyakarta pada tahun 2008. Penulis telah memiliki banyak karya ilmiah yang diterbitkan di jurnal internasional bereputasi. Penulis memiliki minat penelitian pada penggunaan teknologi dalam pengukuran, penilaian, dan evaluasi pendidikan.

Kartianom
Penulis merupakan dosen tetap pada program studi Pendidikan Guru Madrasah Ibtidaiyah (PGMI), IAIN Bone Sulawesi Selatan, sejak

tahun 2019. Penulis menempuh pendidikan formal S1 Pendidikan Matematika di Universitas Dayanu Ikhsanuddin Baubau pada tahun 2009-2014. Setahun berselang, penulis melanjutkan studi S2 Penelitian dan Evaluasi Pendidikan di Universitas Negeri Yogyakarta pada tahun 2015-2017.

Krisna Merdekawati
Penulis merupakan dosen tetap di program studi Pendidikan Kimia, Universitas Islam Indonesia. Penulis menempuh studi S1 Pendidikan Kimia di Universitas Sebelas Maret (UNS) pada tahun 2004-2008. Tahun 2010-2011, penulis melanjutkan studi S2 Pendidikan Sains dengan peminatan Pendidikan Kimia di UNS. Penulis melanjutkan studi S3 Penelitian dan Evaluasi Pendidikan di Universitas Negeri Yogyakarta dengan Beasiswa Pendidikan Indonesia (BPI).

Andi Harpeni Dewantara
Penulis merupakan dosen tetap program studi Pendidikan Guru Madrasah Ibtidaiyah (PGMI) IAIN Bone Sulawesi Selatan sejak tahun 2018. Penulis menempuh pendidikan formal S1 Pendidikan Matematika di Universitas Negeri Makassar pada tahun 2008-2012. Penulis melanjutkan studi S2 Pendidikan Matematika di Universitas Sriwijaya Palembang dengan program beasiswa International Master on Mathematics Education (IMPoME) pada tahun 2013-2015. Saat ini penulis sedang menempuh studi S3 Penelitian dan Evaluasi Pendidikan (PEP) Universitas Negeri Yogyakarta (*intake* tahun 2022) dengan beasiswa LPDP Kemenkeu RI.

Yustina Dwisofiani Lawung
Penulis merupakan dosen tetap di Universitas Katolik Widya Mandira, Kupang, Nusa Tenggara Timur. Penulis menempuh studi S1 di Universitas Katolik Widya Mandira pada tahun 2007 sampai 2011. Pada tahun 2013 penulis menyelesaikan studi S2 Pendidikan Sains di Universitas Negeri Surabaya. Pada tahun 2022, penulis melanjutkan studi S3 Penelitian dan Evaluasi Pendidikan di Universitas Negeri Yogyakarta dengan Beasiswa LPDP.

Widayanti
Penulis merupakan dosen tetap di program studi Pendidikan Fisika, Universitas Nurul Huda, sejak tahun 2019. Penulis menempuh studi S1 Pendidikan Fisika di Universitas Islam Negeri Raden Intan Lampung pada tahun 2013-2017. Penulis melanjutkan studi S2 Pendidikan Fisika di Universitas Lampung pada tahun 2017-2019. Saat ini, penulis sedang menempuh studi S3 Penelitian dan Evaluasi Pendidikan di Universitas Negeri Yogyakarta (*intake* tahun 2022) dengan Beasiswa Pendidikan Indonesia.

Artina Diniaty
Penulis merupakan dosen pada program studi Pendidikan Kimia, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Islam Indonesia. Penulis menempuh pendidikan S1 di program studi Pendidikan Kimia di Universitas Negeri Yogyakarta (UNY) dan pendidikan S2 di program studi Pendidikan Sains dengan konsentrasi Pendidikan Kimia di UNY. Saat ini penulis sedang menempuh studi S3 Penelitian dan Evaluasi Pendidikan di UNY.

Yessica Mega Aprita
Penulis merupakan dosen tetap pada program studi Manajemen Universitas Bina Sarana Informatika sejak tahun 2018. Penulis menempuh pendidikan formal S1 Pendidikan Akuntansi kelas Internasional di Universitas Negeri Yogyakarta (UNY) pada tahun 2010-2014. Penulis melanjutkan studi S2 Penelitian dan Evaluasi Pendidikan di UNY pada tahun 2014-2016 dengan beasiswa internal Program Pascasarjana UNY untuk lulusan terbaik dan berprestasi fakultas. Saat ini, penulis sedang menempuh studi S3 Penelitian dan Evaluasi Pendidikan di Universitas Negeri Yogyakarta (*intake* tahun 2022) dengan Beasiswa Pendidikan Indonesia skema Dosen Perguruan Tinggi Akademik.

Widinda Normalia Arlianty
Penulis merupakan dosen pada program studi Pendidikan Kimia, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Islam Indonesia (UII). Penulis menempuh studi S1 Pendidikan Kimia di

Universitas Sebelas Maret (UNS) dan studi S2 Pendidikan Sains dengan konsentrasi pada Pendidikan Kimia di universitas yang sama. Saat ini penulis sedang menempuh studi S3 Penelitian dan Evaluasi Pendidikan di Universitas Negeri Yogyakarta.

Intan Kumala Sari
Penulis merupakan dosen tetap pada program studi Pendidikan Matematika Universitas Bina Bangsa Getsempena di Banda Aceh sejak tahun 2013. Penulis menempuh pendidikan S1 Pendidikan Matematika di Universitas Syiah Kuala Banda Aceh pada tahun 2004-2008. Selanjutnya, Penulis melanjutkan studi S2 Pendidikan Matematika di Universitas Negeri Surabaya melalui Program Beasiswa International Master Program on Mathematics Education (IMPoME) pada tahun 2010-2012.

Okta Alpindo
Penulis merupakan dosen tetap di Universitas Maritim Raja Ali Haji. Penulis menempuh pendidikan S1 Pendidikan Fisika di Universitas Negeri Padang pada tahun 2010 sampai 2014. Pada tahun 2017 penulis menyelesaikan studi S2 Pendidikan Fisika di Universitas Negeri Padang. Penulis melanjutkan studi S3 Penelitian dan Evaluasi Pendidikan di Universitas Negeri Yogyakarta dengan Beasiswa Pendidikan Indonesia tahun 2022.

Mujiyanto
Penulis merupakan dosen tetap di program studi Pendidikan Keagamaan Buddha (STIAB) Smaratungga Boyolali sejak tahun 2021. Penulis menempuh studi S1 (2014-2018) dan S2 (2018-2020) di STIAB Smaratungga Boyolali. Pada tahun 2022, penulis melanjutkan studi S3 Penelitian dan Evaluasi Pendidikan di Universitas Negeri Yogyakarta (UNY) dengan beasiswa dari Dirjen Bimas Buddha Kemenag RI.

Primanisa Inayati Azizah
Penulis merupakan dosen tetap di program studi Pendidikan IPS, Universitas Negeri Yogyakarta (UNY). Penulis menempuh studi S1

Pendidikan IPS di Universitas Negeri Yogyakarta tahun 2008, dan melanjutkan studi S2 Pendidikan IPS di Universitas Negeri Yogyakarta pada tahun 2014. Saat ini, penulis melanjutkan studi S3 Penelitian dan Evaluasi Pendidikan di Universitas Negeri Yogyakarta.

Rizqy Cahyo Utomo
Penulis merupakan dosen tetap pada program studi Psikologi di Universitas Negeri Yogyakarta. Penulis menempuh pendidikan formal S1 di Fakultas Psikologi, Universitas Gadjah Mada (UGM), dilanjutkan dengan pendidikan Magister Profesi Psikologi di bidang pendidikan di UGM. Penulis saat ini menempuh studi S3 Penelitian dan Evaluasi Pendidikan di Universitas Negeri Yogyakarta. Penulis memiliki minat penelitian pada psikologi pendidikan, kepribadian, dan asesmen pendidikan.

Agung Prihantoro
Penulis merupakan dosen tetap di Universitas Cokroaminoto Yogyakarta dan mengajar mata kuliah Metodologi Penelitian, Penelitian Tindakan Kelas, Bahasa Inggris, dan Bahasa Indonesia. Penulis telah menulis buku-buku tentang Bahasa Inggris dan pendidikan, menulis artikel ilmiah di jurnal dan artikel opini di koran, menerjemahkan 43 buku dari bahasa Inggris ke bahasa Indonesia, dan juga menyunting buku. Penulis telah mengelola penerbit buku, Spirit Mondia Corpora. Pendidikan sarjananya diselesaikan pada program studi Pendidikan Bahasa Inggris di Universitas Negeri Yogyakarta dan pendidikan magisternya ditamatkan pada program studi Penelitian dan Evaluasi Pendidikan, Universitas Negeri Jakarta.

Rizki Fitria Setyaningtyas
Penulis saat ini merupakan tutor di Universitas Terbuka (UT) pada program studi Pendidikan Guru Sekolah Dasar (PGSD) sejak tahun 2022. Penulis menempuh studi S1 Pendidikan IPA Universitas Negeri Yogyakarta, lulus pada tahun 2013. Penulis melanjutkan studi S2 di Universitas Sebelas Maret (UNS), lulus tahun 2018. Saat ini penulis sedang menempuh studi S3 Penelitian dan Evaluasi Pendidikan di Universitas Negeri Yogyakarta.

Fitriyani Hali
Penulis merupakan dosen tetap di program studi Pendidikan Matematika, Universitas Sembilanbelas November Kolaka. Penulis menempuh studi S1 Pendidikan Matematika di Universitas Halu Oleo pada tahun 2006-2010. Tahun 2012-2014, penulis melanjutkan studi S2 Pendidikan Matematika di universitas yang sama. Penulis melanjutkan studi S3 Penelitian dan Evaluasi Pendidikan di Universitas Negeri Yogyakarta dengan Beasiswa Pendidikan Indonesia.

Rina Safitri
Penulis saat ini terdaftar sebagai mahasiswa S2 Penelitian dan Evaluasi Pendidikan, Universitas Negeri Yogyakarta dengan fokus studi pada bidang pengukuran pendidikan. Pada tahun 2019, penulis berhasil memperoleh gelar Sarjana Statistika dari program studi Statistika di Universitas Sebelas Maret (UNS) melalui Bidikmisi.