Rakamin Academy

kimia farma

Project Based Internship

Exploratory Data Analysis

Multivariate Analysis

Daftar Isi

A. *Multivariate Analysis*

Multivariate analysis digunakan untuk menganalisa lebih dari 2 variabel di waktu yang sama, *trends* yang dihasilkan bisa menjadi multidimensi secara alami, dengan analisis ini akan membantu kita memahami manakah data yang memiliki tren yang kompleks pada kombinasi atribut.

Ketika data melibatkan tiga variabel atau lebih, maka data tersebut dikategorikan sebagai *multivariat*. Contoh dari jenis data ini adalah misalkan seorang pengiklan ingin membandingkan popularitas empat iklan di sebuah situs web, maka rasio klik mereka dapat diukur untuk pria dan wanita dan hubungan antar variabel kemudian dapat diperiksa. Hal ini mirip dengan *bivariat* tetapi mengandung lebih dari satu *variabel dependen*. Cara untuk melakukan analisis pada data ini tergantung pada tujuan yang ingin dicapai. Beberapa tekniknya adalah analisis regresi, analisis jalur, analisis faktor, dan analisis varians multivariat (MANOVA).

Dalam *multivariate analysis*, dapat dilakukan Analisa dengan 2 teknik:

1. *Dependency Techniques*

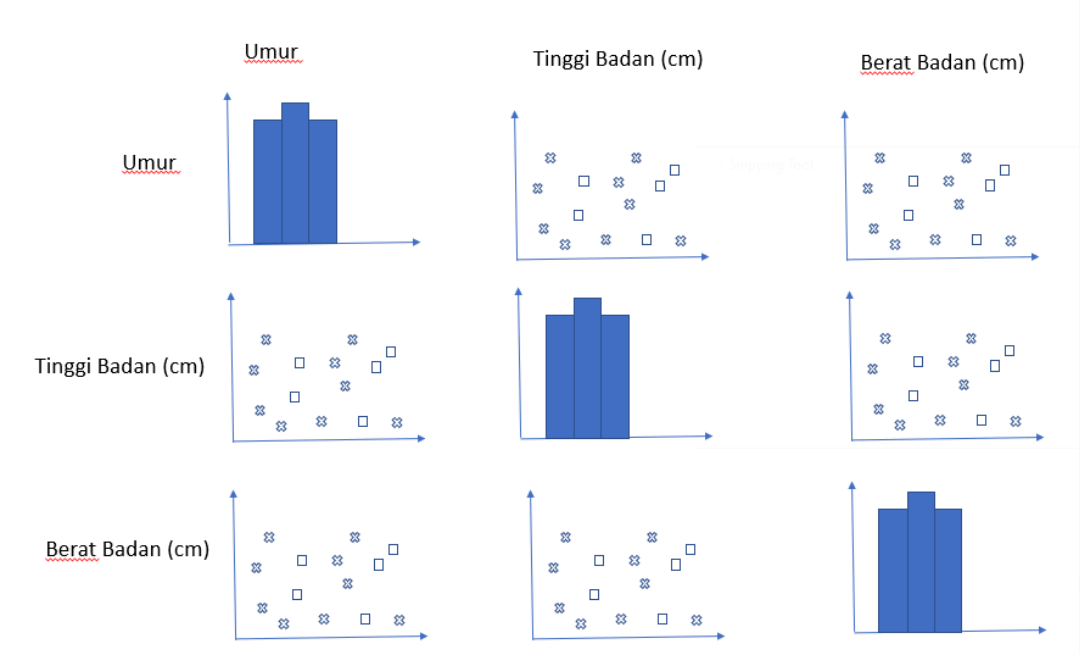
Dependency techniques merupakan teknik analisa ketika satu atau lebih variabel diidentifikasi sebagai variabel dependen dan variabel sisanya diidentifikasi sebagai independen

Analisa bisa menggunakan *multiple regression* dan *multivariate analysis of variance* (MANOVA)

2. *Interdependency Techniques*

Teknik analisis Ketika variabel tidak dapat diklasifikasikan sebagai *independent* atau *dependent*. Analisa pada Teknik ini dapat menggunakan *cluster analysis*.

namun jika data yang kita miliki dalam ukuran kecil, maka analisis *multivariate* bisa menggunakan pairplot seperti berikut:



Pairplot Analysis

Grafik menunjukkan adanya korelasi di tiap *feature*. misalkan hubungan antara umur dan Tinggi Badan (cm). Jika bertambah umur bertambah pula tinggi badan artinya korelasi kedua variabel tersebut adalah positif, jika sebaliknya maka korelasi antara kedua variabel tersebut negatif. Dari *multivariate analisis* ini kita bisa tentukan manakah algoritma yang tepat yang bisa digunakan untuk prediktif analisis.

B. Jenis-jenis *multivariate analysis*:

Multivariate analysis merupakan teknik analisis yang digunakan untuk menganalisis hubungan antara dua atau lebih variabel dalam satu dataset. Tujuan dari *multivariate analysis* adalah untuk menemukan pola atau hubungan antar variabel yang tidak dapat ditemukan dengan analisis *univariat* atau *bivariat*.

Jenis-jenis *multivariate analysis*:

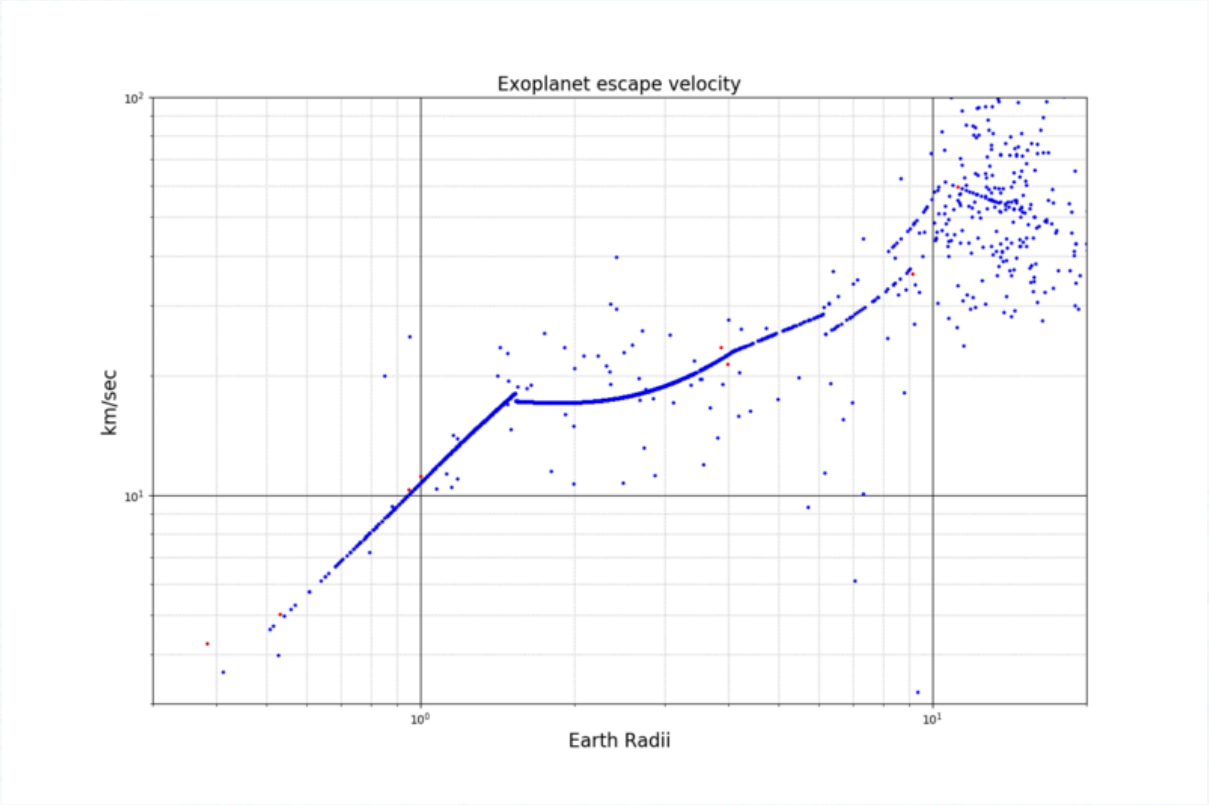
- *Multiple linear regression*
- *Multiple logistic regression*
- *Multivariate analysis of variance* (MANOVA)
- *Principal Component Analysis* (PCA)
- *Factor Analysis* (FA)
- *Cluster Analysis* (CA)
- dan sebagainya

1. *Multiple Linear Regression*

Multiple linear regression adalah metode ketergantungan yang melihat hubungan antara satu variabel dependen dengan dua atau lebih variabel independen. Model regresi berganda akan memberitahu kita sejauh mana setiap variabel independen memiliki hubungan linier dengan variabel dependen. Hal ini berguna untuk membantu kita memahami faktor mana yang kemungkinan besar akan mempengaruhi hasil tertentu, sehingga kita dapat memperkirakan hasil di masa mendatang.

Contoh regresi berganda:

Sebagai seorang analis data, kita dapat menggunakan regresi berganda untuk memprediksi pertumbuhan tanaman. Dalam contoh ini, pertumbuhan tanaman adalah variabel dependen kita dan kita ingin melihat bagaimana berbagai faktor mempengaruhinya. Variabel independen kita bisa berupa curah hujan, suhu, jumlah sinar matahari, dan jumlah pupuk yang ditambahkan ke tanah. Model regresi berganda akan menunjukkan kepada kita proporsi varians dalam pertumbuhan tanaman yang dijelaskan oleh masing-masing variabel independen.



Contoh Visualisasi Regresi Berganda

2. *Multiple logistic regression*

Analisis regresi logistik digunakan untuk menghitung (dan memprediksi) probabilitas terjadinya peristiwa biner. Hasil biner adalah hasil yang hanya memiliki dua kemungkinan, yaitu kejadian tersebut terjadi (1) atau tidak terjadi (0). Jadi, berdasarkan satu set variabel independen, regresi logistik dapat memprediksi seberapa besar kemungkinan skenario tertentu akan muncul. Regresi ini juga digunakan untuk klasifikasi. Kita dapat mempelajari perbedaan antara regresi dan klasifikasi di sini.

Contoh regresi logistik:

Bayangkan kita bekerja sebagai analis di sektor asuransi dan perlu memprediksi seberapa besar kemungkinan setiap pelanggan potensial akan mengajukan klaim. Kita dapat memasukkan berbagai variabel independen ke dalam model, seperti usia, apakah mereka memiliki kondisi kesehatan yang serius atau tidak, pekerjaan mereka, dan sebagainya. Dengan menggunakan variabel-variabel ini, analisis regresi logistik akan menghitung probabilitas kejadian (mengajukan klaim) yang terjadi. Contoh lain yang sering dikutip adalah filter yang digunakan untuk mengklasifikasikan email sebagai "spam" atau "bukan spam".

3. *Multivariate analysis of variance (MANOVA)*

Analisis varians multivariat (MANOVA) digunakan untuk mengukur pengaruh beberapa variabel independen terhadap dua atau lebih variabel dependen. Dengan MANOVA, penting untuk diperhatikan bahwa variabel independen bersifat kategorikal, sedangkan variabel dependen bersifat metrik. Variabel kategorikal adalah variabel yang termasuk dalam kategori yang berbeda - misalnya, variabel "status pekerjaan" dapat dikategorikan ke dalam unit-unit

tertentu, seperti "bekerja penuh waktu", "bekerja paruh waktu", "menganggur", dan seterusnya. Variabel metrik diukur secara kuantitatif dan memiliki nilai numerik.

Dalam analisis MANOVA, kita melihat berbagai kombinasi variabel independen untuk membandingkan perbedaan pengaruhnya terhadap variabel dependen.

Contoh MANOVA:

Bayangkan kita bekerja di sebuah perusahaan teknik yang memiliki misi untuk membuat roket super cepat dan ramah lingkungan. Kita dapat menggunakan MANOVA untuk mengukur efek dari berbagai kombinasi desain terhadap kecepatan roket dan jumlah karbon dioksida yang dikeluarkannya. Dalam skenario ini, variabel independen kategorikal kita dapat berupa:

- Jenis mesin, dikategorikan sebagai E1, E2, atau E3
- Bahan yang digunakan untuk eksterior roket, dikategorikan sebagai M1, M2, atau M3
- Jenis bahan bakar yang digunakan untuk menggerakkan roket, dikategorikan sebagai F1, F2, atau F3

Variabel dependen metrik kita adalah kecepatan dalam kilometer per jam, dan karbon dioksida yang diukur dalam bagian per juta. Dengan menggunakan MANOVA, kita akan menguji kombinasi yang berbeda (misalnya E1, M1, dan F1 vs E1, M2, dan F1, vs E1, M3, dan F1, dan seterusnya) untuk menghitung efek dari semua variabel independen. Hal ini akan membantu kita menemukan solusi desain yang optimal untuk roket kita.

4. *Principal Component Analysis (PCA)*

PCA adalah metode untuk mereduksi dimensi dari data dengan memproyeksikan data ke dalam ruang yang lebih rendah. PCA mengidentifikasi variabel yang memiliki hubungan kuat dan menjadikannya sebagai variabel baru yang disebut dengan komponen utama. Komponen utama ini dapat menjelaskan variabilitas dari dataset yang asli.

5. *Factor Analysis (FA)*

FA adalah metode yang digunakan untuk mengidentifikasi variabel laten atau tidak terlihat yang berkontribusi pada pola dalam data. Variabel laten ini diidentifikasi berdasarkan korelasi antar variabel dalam dataset. FA dapat digunakan untuk mengurangi dimensi dari dataset dan mengidentifikasi faktor-faktor yang mempengaruhi hubungan antar variabel.

Overfitting adalah kesalahan pemodelan yang terjadi ketika sebuah model terlalu dekat dan spesifik dengan set data tertentu, sehingga kurang dapat digeneralisasi ke set data di masa depan, dan dengan demikian berpotensi kurang akurat dalam prediksi yang dibuatnya.

Analisis faktor bekerja dengan mendeteksi kumpulan variabel yang berkorelasi tinggi satu sama lain. Variabel-variabel ini kemudian dapat diringkas menjadi satu variabel. Analis data akan sering melakukan analisis faktor untuk mempersiapkan data untuk analisis selanjutnya.

Contoh analisis faktor:

Bayangkan kita memiliki kumpulan data yang berisi data yang berkaitan dengan pendapatan, tingkat pendidikan, dan pekerjaan seseorang. kita mungkin menemukan tingkat korelasi yang tinggi di antara masing-masing variabel ini, dan dengan demikian menguranginya menjadi faktor tunggal "status sosial ekonomi." Kita mungkin juga memiliki data tentang seberapa senang mereka dengan layanan pelanggan, seberapa besar mereka menyukai produk tertentu, dan seberapa besar kemungkinan mereka merekomendasikan produk tersebut kepada teman.

Masing-masing variabel ini dapat dikelompokkan ke dalam faktor tunggal "kepuasan pelanggan" (selama mereka ditemukan berkorelasi kuat satu sama lain). Meskipun kita telah mengurangi beberapa titik data menjadi hanya satu faktor, kita tidak benar-benar kehilangan informasi apa pun-faktor ini cukup menangkap dan mewakili variabel individu yang bersangkutan. Dengan kumpulan data yang telah "disederhanakan", kita sekarang siap untuk melakukan analisis lebih lanjut.

6. *Cluster Analysis (CA)*

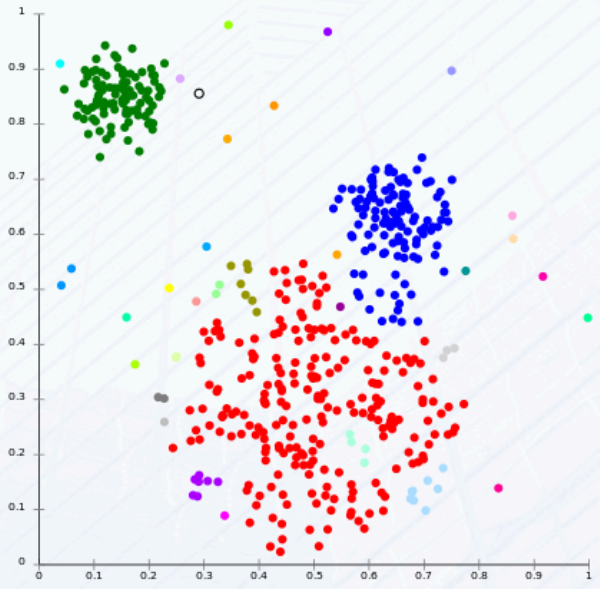
CA adalah metode yang digunakan untuk mengelompokkan objek atau variabel berdasarkan kemiripannya. CA dapat membantu mengidentifikasi pola atau kelompok dalam data yang serupa dalam karakteristik. Dalam CA, objek atau variabel diklasifikasikan ke dalam kelompok atau cluster berdasarkan nilai variabel-variabel yang saling terkait.

Ketika mengelompokkan data ke dalam klaster, tujuannya adalah agar variabel-variabel dalam satu klaster lebih mirip satu sama lain dibandingkan

dengan variabel-variabel dalam klaster lain. Hal ini diukur dalam hal jarak intracluster dan antar klaster. Analisis klaster membantu kita memahami bagaimana data dalam sampel kita didistribusikan, dan menemukan pola.

Contoh analisis cluster:

Contoh utama dari analisis klaster adalah segmentasi audiens. Jika kita bekerja di bidang pemasaran, kita dapat menggunakan analisis klaster untuk menentukan kelompok pelanggan yang berbeda yang dapat memperoleh manfaat dari kampanye yang lebih bertarget. Sebagai analis kesehatan, kita dapat menggunakan analisis klaster untuk mengeksplorasi apakah faktor gaya hidup atau lokasi geografis tertentu terkait dengan kasus penyakit tertentu yang lebih tinggi atau lebih rendah. Karena ini adalah teknik saling ketergantungan, analisis klaster sering kali dilakukan pada tahap awal analisis data.



Contoh Visualisasi Cluster

Glosarium

- ***Multivariate Analysis :*** Pendekatan analisis statistik yang memeriksa hubungan antara dua atau lebih variabel independen terhadap satu atau lebih variabel dependen sekaligus. *Multivariate analysis* digunakan untuk memahami kompleksitas hubungan antara variabel dalam dataset.
- ***Dependency Techniques :*** Teknik analisis yang digunakan untuk menemukan dan mengukur hubungan ketergantungan antara dua atau lebih variabel dalam sebuah dataset. *Dependency techniques* membantu dalam memahami bagaimana perubahan pada satu variabel dapat mempengaruhi variabel lainnya.
- ***Interdependency Techniques :*** Teknik analisis yang digunakan untuk menemukan hubungan timbal balik antara variabel dalam dataset. *Interdependency techniques* mengungkapkan bagaimana variabel-variabel saling mempengaruhi satu sama lain.
- ***Multiple Linear Regression*** **:** Metode statistik yang digunakan untuk mengukur hubungan antara dua atau lebih variabel independen (*predictor*) terhadap satu variabel dependen (*outcome*) dengan model linier.
- ***Multiple Logistic Regression*** **:** Metode statistik yang digunakan untuk menganalisis hubungan antara dua atau lebih variabel independen kategorikal atau kontinu dengan satu variabel dependen biner.
- ***Multivariate Analysis of Variance*** **(MANOVA) :** Teknik analisis statistik yang digunakan untuk menguji perbedaan antara rata-rata dari dua atau lebih variabel dependen dalam satu atau lebih kelompok atau kondisi.

- ***Principal Component Analysis* (PCA) :** Teknik statistik yang digunakan untuk mengurangi dimensi dari dataset yang kompleks dengan mereduksi variabilitas data ke dalam beberapa komponen utama yang paling signifikan.
- ***Factor Analysis* (FA) :** Metode statistik yang digunakan untuk mengidentifikasi pola-pola dalam kumpulan data yang kompleks dengan mengelompokkan variabel-variabel yang berkorelasi tinggi ke dalam faktor-faktor yang lebih sedikit.
- ***Cluster Analysis* (CA) :** Teknik analisis yang digunakan untuk mengelompokkan objek-objek atau individu-individu dalam dataset ke dalam kelompok-kelompok yang serupa berdasarkan kesamaan karakteristik mereka.

References

Multivariate Analysis & Independent Component - Statistics How To

An Introduction to Multivariate Analysis [With Examples]

A Gentle Introduction to Descriptive Analytics | by Angelica Lo Duca