Data Science Process and Feature Engineering

Mustakim, S.T., M.Kom.

Tujuan Feature Engineering

1 **Meningkatkan Performa Model**

Dengan menghapus fitur yang tidak relevan dan menambahkan fitur baru yang lebih informatif, kita dapat meningkatkan akurasi model prediksi.

2 **Mengurangi Overfitting**

Dengan melakukan teknik regresi dan regularisasi fitur, kita dapat mengurangi potensi overfitting pada model kita.

3 **Meningkatkan Interpretabilitas Model**

Dengan membuat fitur yang lebih terkait dengan masalah bisnis, kita dapat lebih memahami dan menjelaskan hasil prediksi model.

Proses Feature Engineering

1

Pemahaman Masalah

Menganalisis masalah bisnis dan dataset untuk mengidentifikasi fitur yang penting.

2

Pembersihan Data

Mengatasi missing values, outliers, dan data yang tidak konsisten untuk memastikan kualitas data yang baik.

3

Pemilihan Fitur

Menggunakan metode statistik dan algoritma machine learning untuk memilih fitur yang memiliki hubungan kuat dengan target.

Konsep Feature Engineering

Ekstraksi Fitur (Feature Extraction)

Ekstraksi fitur adalah proses menghasilkan fitur-fitur baru dari data yang sudah ada. Contohnya, jika Anda memiliki data tanggal, Anda dapat mengekstraksi fitur-fitur tambahan seperti hari dalam seminggu, bulan, atau tahun. Teknik seperti Principal Component Analysis (PCA) juga digunakan untuk mengurangi dimensi data.

Pemilihan Fitur (Feature Selection)

Salah satu langkah awal dalam feature engineering adalah memilih subset fitur yang paling relevan dan penting untuk tujuan analisis. Ini membantu mengurangi dimensi data dan menghindari overfitting. Metode seperti analisis korelasi, analisis informasi, dan pemodelan seleksi fitur.

Konsep Feature Engineering (1)

Penanganan Data yang Hilang (Handling Missing Data)

Data yang hilang atau missing data adalah masalah umum dalam sains data. Strategi untuk mengatasi data yang hilang termasuk penghapusan baris atau kolom yang memiliki banyak missing data, mengisi data yang hilang dengan nilai-nilai tertentu (misalnya, rata-rata atau median), atau menggunakan model untuk memprediksi nilai yang hilang.

Scaling dan Normalisasi

Beberapa algoritma pembelajaran mesin, seperti Support Vector Machines (SVM) dan K-Means, sensitif terhadap perbedaan skala antar fitur. Oleh karena itu, scaling (skalasi) atau normalisasi fitur-fitur numerik sering diperlukan untuk menjaga keseimbangan.

Konsep Feature Engineering (2)

Konversi Teks ke Fitur Numerik

Ketika bekerja dengan data teks, seperti teks ulasan produk, Anda perlu mengonversi teks tersebut menjadi fitur numerik yang dapat digunakan oleh model. Metode seperti TF-IDF (Term Frequency-Inverse Document Frequency) dan Word Embeddings seperti Word2Vec dan GloVe digunakan.

Pengurangan Fitur (Feature Reduction)

ika data memiliki banyak fitur dan Anda ingin mengurangi kompleksitas model, teknik pengurangan fitur seperti L1 Regularization atau Recursive Feature Elimination (RFE) dapat digunakan.

Metode Seleksi Fitur

Filter

Menggunakan metrik statistik seperti korelasi dan chi-square untuk memilih fitur yang paling berpengaruh.

Wrapper

Menggunakan algoritma pembungkus seperti Recursive Feature Elimination untuk memilih subset fitur terbaik.

Embedded

Menggunakan algoritma machine learning seperti LASSO dan Random Forest untuk memilih fitur selama proses pembelajaran.

Metode Ekstraksi Fitur

Metode	Keterangan
PCA	Menurunkan dimensi fitur dengan mengubahnya ke ruang laten (latent space) dengan varian terbesar.
TF-IDF	Menghitung bobot kata dalam teks berdasarkan frekuensi munculnya dalam dokumen dan koleksi dokumen.
Bag of Words	Mengubah teks menjadi vektor fitur berdasarkan keberadaan kata dalam dokumen.

Teknik Transformasi Fitur

Scaling

Mengubah data ke skala yang serupa, seperti Z-score scaling atau Min-Max scaling.

Polynomial Features

Membuat fitur polynomial dengan memperluas kombinasi fitur asli.

Log Transform

Mengubah data yang tidak terdistribusi normal menjadi distribusi normal dengan logaritma.

One-Hot Encoding

Mengubah data kategorikal menjadi representasi vektor biner.

Evaluasi dan Validasi Fitur

1

Menggunakan Metric Performance

Membandingkan performa model sebelum dan sesudah implementasi fitur engineering dengan metrik seperti akurasi, presisi, dan recall.

2

Validasi Silang

Menerapkan validasi silang untuk memastikan hasil yang stabil dan generalisasi fitur engineering pada data yang tidak terlihat sebelumnya.

3

Analisis Kesalahan

Menganalisis kesalahan model untuk memahami dampak fitur engineering pada prediksi yang salah.

Ekstraksi Fitur dengan PCA

Ekstraksi Fitur dengan PCA

Langkah 1

Menghitung matriks kovarians dan nilai eigen.

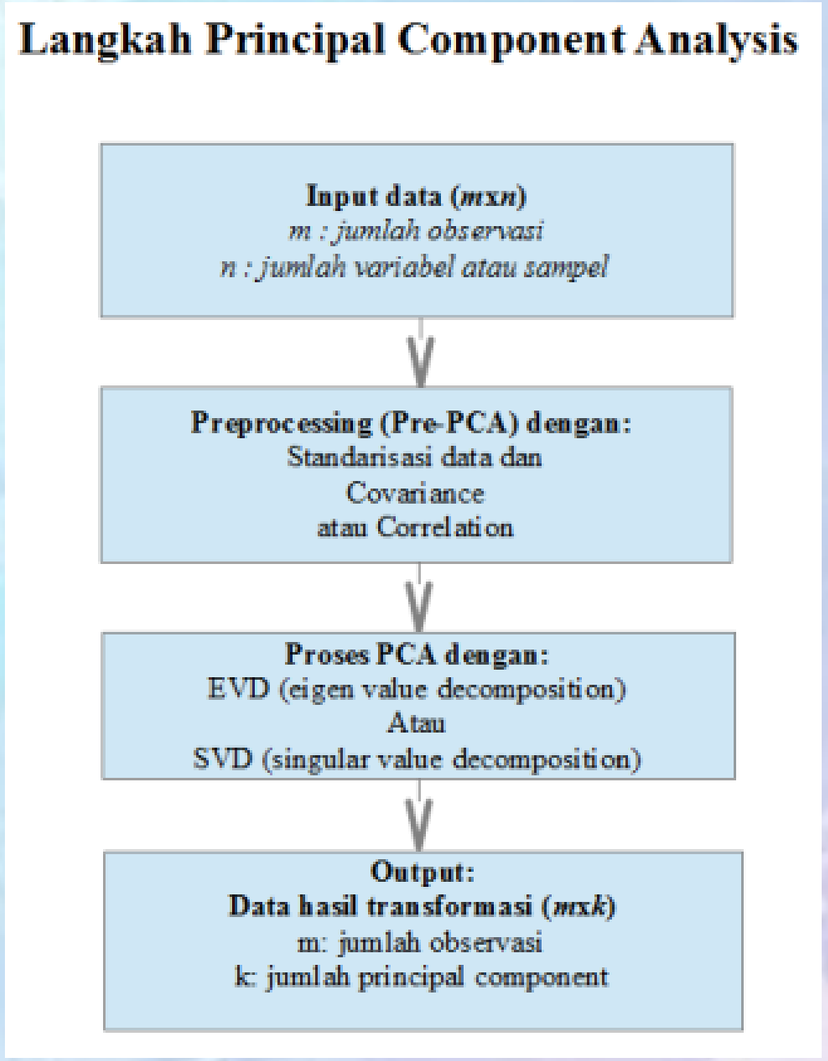
Langkah 2

Mengurutkan nilai eigen dalam urutan menurun dan memilih komponen utama.

Langkah 3

Menghitung skor fitur dan memilih fitur-fitur dengan skor tinggi.

Langkah-langkah PCA



Simulasi PCA

Atribut 1	Atribut 2	Atribut 3	Atribut 4	Atribut 5

Component 1	Component 2

Pemahaman

Ekstraksi Fitur (Feature Extraction)

1. Cari dataset di Kaggle, tentukan beberapa atribut awal yang akan anda gunakan (misalnya atribut awal adalah 10)
2. Lakukan proses EF dengan PCA (Silahkan cari lib di Kaggle atau Google secara umum. Gunakan Google Collab
3. Tentukan misalnya dari 10 atribut menjadi 2 atribut atau 3 atribut, 10 Menjadi 2 Komponen dan 10 Menjadi 3 Komponen
4. Output berupa Atribut baru dengan 2 atribut dan 3 atribut disertai dengan plotting visualisasi data

Seleksi Fitur dengan Chi Square

Definisi dari Chi Square

Konsep Dasar

Chi Square adalah uji statistik yang digunakan untuk menguji keberhubungan antara dua variabel kategorikal.

Perhitungan Nilai Chi Square

Perhitungan melibatkan perbandingan antara nilai pengamatan dan nilai yang diharapkan.

Interpretasi Hasil

Nilai Chi Square dapat memberikan informasi tentang tingkat keberhubungan antara dua variabel.

Simulasi Seleksi Fitur

Atribut 1	Atribut 2	Atribut 3	Atribut 4	Atribut 5

Atribut 1	Atribut 4	Atribut 2	Atribut 5	Atribut 3

Penggunaan Chi Square dalam Seleksi Fitur

1

Step 1

Menghitung nilai Chi Square antara setiap fitur dengan variabel target.

2

Step 2

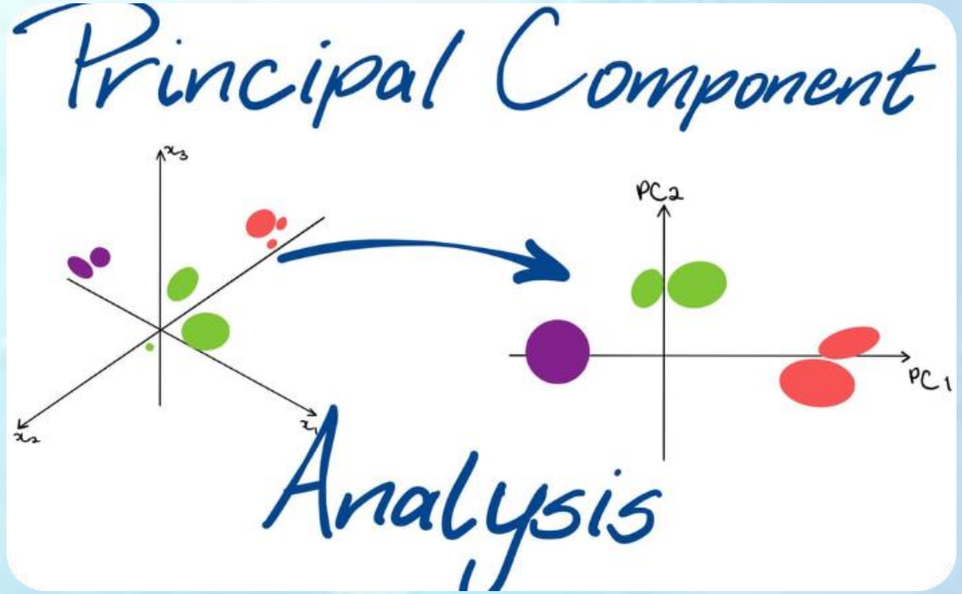
Memilih fitur-fitur dengan nilai Chi Square yang tinggi sebagai fitur yang paling berhubungan dengan variabel target.

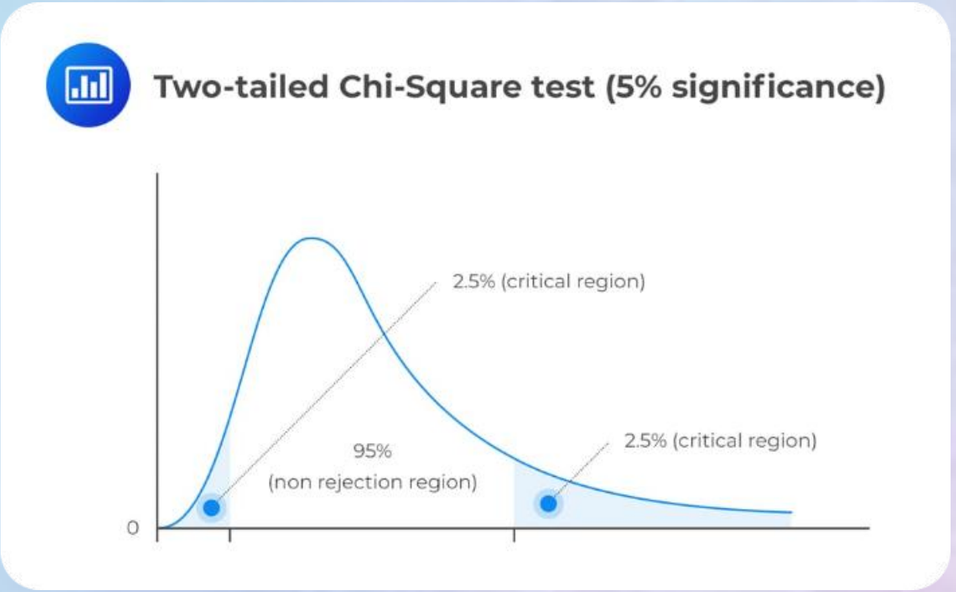
3

Step 3

Mengintegrasikan fitur-fitur terpilih ke dalam model analisis data.

Perbandingan Antara PCA dan Chi Square





PCA

Menggunakan metode statistik untuk mengidentifikasi komponen utama yang menjelaskan variasi dalam data.

Chi Square

Menggunakan uji statistik untuk mengevaluasi keberhubungan antara variabel-variabel kategorikal.

Keuntungan dari Menggunakan Seleksi Fitur dan Ekstraksi Fitur

1

Meningkatkan Performa Model

Dengan menggunakan fitur-fitur yang paling relevan, performa model analisis data dapat meningkat secara signifikan.

2

Mengurangi Biaya

Dengan mengurangi jumlah fitur yang diproses, biaya komputasi dan penggunaan sumber daya dapat berkurang.

3

Mudah Diterapkan

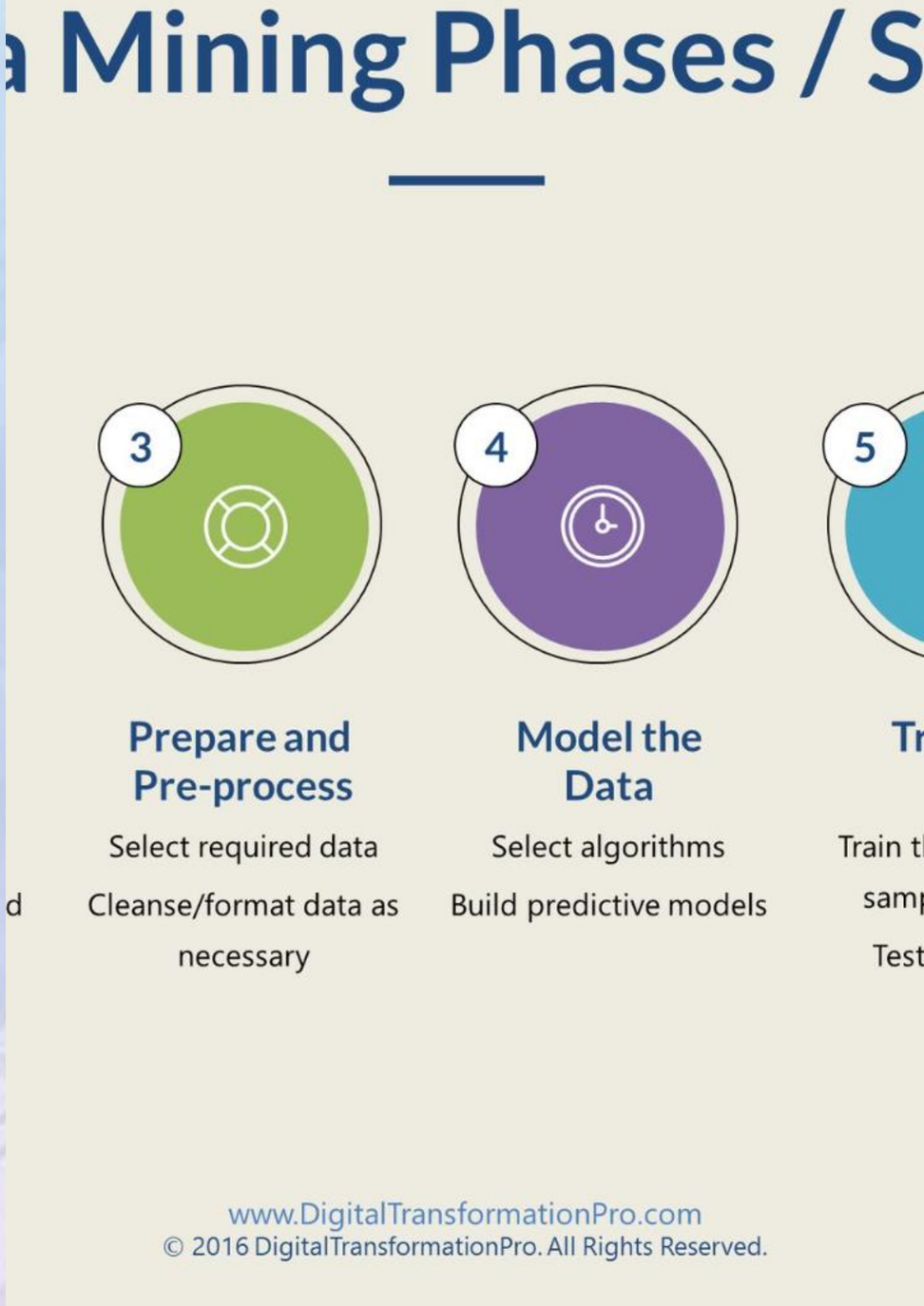
Metode seleksi fitur seperti PCA dan Chi Square dapat dengan mudah diterapkan dalam berbagai kasus analisis data.

Pemahaman

Seleksi Fitur (Feature Selection)

1. Cari dataset di Kaggle, tentukan beberapa atribut awal yang akan anda gunakan (misalnya atribut awal adalah 10)
2. Lakukan proses FS dengan Chi Square (Silahkan cari lib di Kaggle atau Google secara umum. Gunakan Google Collab
3. Tentukan misalnya dari 10 atribut utama menjadi beberapa atribut saja, hasil dari prosess Chi Square
4. Output berupa atribut dengan kelayakan atau yang telah memenuhi standar

Dimensi Reduksi vs Ekstraksi Fitur



Tujuan dari Dimensi Reduksi

Tujuan dari dimensi reduksi adalah untuk menghilangkan atribut yang tidak relevan atau redundan, mempercepat waktu analisis data, dan mempertahankan sebagian besar informasi yang ada.

Teknik-teknik Umum Dimensi Reduksi

1 Analisis Komponen Utama (PCA)

Analisis PCA mengidentifikasi kombinasi linier atribut untuk menghasilkan beberapa fitur yang menjelaskan variasi terbesar dalam data.

2 Linier Diskriminant Analisis (LDA)

Analisis LDA mempelajari pemisah yang optimal antara kelas data yang berbeda dengan tujuan menghasilkan fitur-fitur yang memaksimalkan diskriminasi.

Latent Dirichlet Allocation (LDA) pada Natural Language Processing (NLP)

LDA adalah metode yang digunakan untuk memodelkan topik dalam corpus teks. Ini membantu mengurangi dimensi dalam NLP dan menggambarkan topik yang ada dalam dokumen.

Contoh Kasus Penggunaan Teknik Dimensi Reduksi pada Dunia Nyata

1

Penerapan di Bidang Bioinformatika

Contohnya adalah penggunaan PCA untuk memproses data genomik dan mengidentifikasi fitur-fitur kunci yang terkait dengan penyakit.

2

Pengenalan Wajah

Analisis PCA digunakan dalam pengenalan wajah untuk mengurangi dimensi fitur wajah dan meningkatkan akurasi dalam pencocokan.

3

Analisis Sentimen Media Sosial

LDA digunakan untuk mengidentifikasi topik-topik yang dibicarakan secara umum dalam suatu kontroversi atau tren di media sosial.

Perbedaan Reduksi Dimensi dengan Ekstraksi Fitur: Tujuan Utama

Dimensi Reduksi:
Tujuan utama dari dimensi reduksi adalah mengurangi jumlah fitur (dimensi) dalam data tanpa mengorbankan informasi yang signifikan. Hal ini dilakukan dengan menghilangkan beberapa fitur atau merangkumnya ke dalam sejumlah fitur yang lebih kecil.

Fitur Ekstraksi:
Tujuan utama dari fitur ekstraksi adalah menghasilkan fitur-fitur baru dari data asli. Fitur-fitur baru ini dapat berupa kombinasi atau representasi yang lebih informatif dari fitur-fitur awal. Fitur ekstraksi tidak selalu mengurangi dimensi, tetapi lebih fokus pada pembuatan fitur-fitur yang lebih bermakna

Perbedaan Reduksi Dimensi dengan Ekstraksi Fitur: Proses

Dimensi Reduksi:
Dimensi reduksi mencoba untuk menghilangkan fitur-fitur yang kurang penting atau redundan dari data asli. Ini dapat dilakukan dengan teknik seperti Principal Component Analysis (PCA) atau Feature Selection.

Fitur Ekstraksi:
Fitur ekstraksi menciptakan fitur-fitur baru yang tidak ada dalam data asli. Ini sering melibatkan teknik seperti analisis faktor, analisis diskriminan linier, atau transformasi seperti Word2Vec dalam pemrosesan bahasa alami

Perbedaan Reduksi Dimensi dengan Ekstraksi Fitur: Jumlah Fitur Hasil

Dimensi Reduksi:
Dimensi reduksi menghasilkan data dengan jumlah fitur yang lebih sedikit daripada data asli. Biasanya, ini mengurangi dimensi data.

Fitur Ekstraksi:
Fitur ekstraksi menghasilkan fitur-fitur baru yang dapat memiliki jumlah yang sama, lebih banyak, atau lebih sedikit daripada fitur asli, tergantung pada teknik yang digunakan

Perbedaan Reduksi Dimensi dengan Ekstraksi Fitur: Contoh Aplikasi/ Algoritma

Dimensi Reduksi:
PCA adalah contoh dimensi reduksi yang umum digunakan. Ini mengurangi dimensi data dengan mengidentifikasi arah-arah (komponen-komponen) utama variasi dan memproyeksikan data ke ruang tersebut.

Fitur Ekstraksi:
Misalnya, dalam pemrosesan bahasa alami, fitur ekstraksi dapat menggunakan Word2Vec untuk menghasilkan vektor representasi kata yang lebih informatif daripada representasi one-hot encoding

Terima Kasih