

Nama : Naisha Aqshalya Anderindia
NPM : 51423065
Kelas : 2IA11

1

Quis VC3

- 1) Jelaskan apa yang dimaksud
 - Imputasi
 - Handling Outlier
 - Scaling
 - Encoding
- 2) Sebutkan Jenis-jenis Algoritma Regresi !
- 3) Sebutkan kelebihan dan kekurangan
 - Decision Tree
 - Clustering
- 4) Jelaskan mengapa Natural Language Processing penting ?

Jawaban :

- 1) - Imputasi
 - ↳ Imputasi adalah proses menggantikan nilai yang hilang (missing values) dalam dataset dengan nilai tertentu agar analisis data tidak terganggu.
- Handling Outlier
 - ↳ Adalah proses menangani data pencaian (outlier), yaitu data yang nilainya jauh berbeda dari nilai-nilai lainnya dalam dataset.
- Scaling
 - ↳ Proses mengubah skala nilai data sehingga berada dalam rentang tertentu, seperti 0-1 atau -1-1.
- Encoding
 - ↳ Proses mengubah data kategori menjadi representasi numerik agar dapat digunakan oleh algoritma pembelajaran mesin.

- 2)

1. Regresi Linear	6. Regresi Lasso
2. Regresi Polinomial	7. ANN Regression
3. Support Vector Regression	8. k-NN Regression
4. Regresi Decision Tree	
5. Regresi Random forest	

- 3) - Decision Tree
 - Kelebihan :
 - > Dapat menangani variabel bersifat continuous dan categorical
 - > Dapat meniru cara berfikir manusia sehingga membantu dalam pemahaman data dan interpretasi yang baik
 - > Algoritmenya cepat & dapat menangani dataset yang besar dengan akurasi yang cukup dan baik.

- kekurangan :

- > Dapat menjadi tidak stabil rentan mengalami error pada klasifikasi yang memiliki banyak kelas

- Clustering

- kelebihan :

- > tidak membutuhkan label data
- > Mengungkapkan pola dan struktur data
- > Fleksibel untuk berbagai data

- kekurangan :

- > Sensitif terhadap skala data
- > Rentan terhadap outlier
- > Pemilihan jumlah kluster yang sulit

4) Natural Language Processing penting karena mempermudah manusia dalam mengakses dan memanfaatkan teknologi, memaksimalkan pengolahan data teks, serta membuka peluang baru untuk inovasi di berbagai bidang seperti kesehatan, bisnis, pendidikan dan hiburan. Dengan perkembangan teknologi, Natural Language Processing (NLP) semakin relevan dalam kehidupan sehari-hari.

Rangkuman : IMPUTASI - HANDLING OUTLIER - SCALING - ENCODING

1. Imputasi

↳ Proses memperkirakan nilai data yang hilang atau tidak konsisten.

- Teknik Imputasi :
 - > Numerik : Mean, median, nilai arbitrer, nilai ujung (extreme values).
 - > Nominal : kategori yang sering muncul, menambahkan kategori yang hilang
 - > Lainnya : k-NN, regresi deterministik, regresi stokastik.

- Kelebihan / kekurangan :

- > Imputasi sederhana mudah diterapkan tetapi dapat mendistorsi distribusi data jika proporsi data hilang besar (>5%).

2. Handling Outlier

↳ Mengeloa data yang sangat berbeda dari mayoritas (anomali).

- Manfaat :
 - > Deteksi penipuan (misal : kartu kredit).
 - > Analisis medis, segmentasi pelanggan.
- Penyebab : Variasi alami, kesalahan pengukuran, atau data dari kelas berbeda.
- Pendekatan :
 - > Grafis : Representasi visual
 - > Model-Based : pendekatan statistik
 - > Distance-Based : Pendekatan berbasis jarak (clustering).
 - > Deviation-Based : Deteksi deviasi.

3. Scaling

- ↳ Normalisasi data numerik untuk menyetarakan skala antar fitur.
- Metode Umum: Min-max scaling, Z-score standardization.

4. Encoding

- ↳ Perubahan simbol menjadi bentuk lain yang dapat dipahami mesin.
- Sistem pengkodean Umum:
 - > ASCII : kode 7-bit untuk 128 karakter
 - > EBCDIC : kode 8-bit untuk 256 karakter.
 - > Lainnya : Baudot code, Binary Coded Decimal (BCD)

Tahapan Handling Data

1. Identifikasi masalah
2. Pengumpulan Data
3. Pengajian data (tabel, grafik)
4. Analisis Data
5. Kesimpulan

Rangkuman : DASAR REGRESI & PENGENALAN REGRESI LINEAR

1. Dasar-dasar Regresi

- ↳ Regresi adalah metode statistik untuk menganalisis hubungan antara dua atau lebih variabel.
- Variabel Respon (Y) : variabel yang bergantung
- Variabel penduga (X) : variabel independen.
- Tujuan :
 - > Membuat estimasi rata-rata variabel tergantung.
 - > Menguji hipotesis dependensi variabel.
 - > Memamalkan nilai rata-rata berdasarkan variabel bebas.

2. Pengenalan Regresi Linear

- Asumsi dalam Regresi Linear sederhana :
 - > Eksogenitas lemah : Variabel X tetap, Y acak.
 - > Linearitas : Hubungan antara X dan Y bersifat proporsional.
 - > Error Variance : Varians error konstan.
 - > Autokorelasi : Untuk data time-series, harus bebas autokorelasi.
- Asumsi klasik Regresi Linear Berganda :
 - > Data berskala interval/rasio
 - > Tidak ada outlier.
 - > Homoskedastisitas : Varian error konstan
 - > Bebas dari multikolinearitas, residual berdistribusi normal.

3. Karakteristik / Tipe Masalah yang Bisa Ditelesaikan dengan Regresi

- Contoh aplikasi :
 - > Prakiraan penjualan produk
 - > Analisis kepuasan pelanggan
 - > Estimasi harga
 - > Prediksi pendapatan pekerjaan
- Algoritma Regresi
 - > Regresi Linear, Polynomial, Support Vector, Decision Tree, Random forest, LASSO, ANN, k-NN, dll.

4. Kelebihan Regresi Linear

- Ringan dan mudah digunakan
- Tidak memerlukan tuning parameter
- Mudah dipahami dan diinterpretasikan.

Rangkuman : DECISION TREE DAN CLUSTERING

1. Pengenalan

- Decision Tree
 - ↳ Representasi skematik untuk proses pengambilan keputusan melalui diagram bercabang.
 - Komponen Utama:
 - > Titik pilihan : Alternatif solusi yang di pilih
 - > Cabang alternatif : Jalur keputusan yang tersedia
 - > Titik hasil : Nilai akhir yang dihasilkan dari keputusan
 - > Kerusakan : Kombinasi keputusan dan keadaan acak.
 - Tipe Decision Tree
 - > Classification Tree : Hasil kategorikal (contoh : "ya" atau "tidak")
 - > Regression Tree : Hasil kontinu (contoh : angka 1, 2, 3)
 - Kelebihan :
 - > Menangani data kontinu dan kategorikal.
 - > Cepat dan efisien untuk dataset besar.
 - > Mudah dipahami dan diinterpretasikan.
 - Kekurangan :
 - > Rentan terhadap error pada klasifikasi dengan banyak kelas.

• Clustering

↳ Proses mengelompokkan data berdasarkan kemiripan antar objek

□ Tujuan :

- > Menemukan pola tersembunyi dalam data.
- > Membagi data menjadi kelompok-kelompok dengan karakteristik serupa.

□ Kelebihan & kekurangan :

- > Fleksibel menggunakan berbagai algoritma.
- > Bergantung pada parameter seperti nilai k dalam k -Means.

2. Pendekatan dan Aplikasi Clustering

• Pendekatan Clustering :

- > Partisi : Membagi data ke dalam kelompok (contoh: k -Means, k -Medoids).
- > Hirarkis : Membuat struktur hirarkis (Contoh: BIRCH, CAMELEON).
- > Density-Based : Berdasarkan konektivitas dan fungsi densitas (contoh: DBSCAN, OPTICS).
- > Pendekatan lain : Centroid-Based, model-based, constraint-based.

• Aplikasi Clustering :

- > Analisis pola : pengenalan pola dan pemrosesan gambar.
- > Analisis Spasial : Identifikasi cluster geografis.
- > Ekonomi : Penelitian pasar, segmentasi pelanggan.
- > WWW : Pengelompokan hasil pencarian atau data weblog.

3. Karakteristik Masalah yang Bisa Dipecahkan

- Decision Tree : Pengambilan keputusan dengan hasil kategorikal atau kontinu.
- Clustering : Mengelompokkan data untuk menemukan pola atau informasi tersembunyi.

Rangkuman : NATURAL LANGUAGE PROCESSING DAN DEEP LEARNING

1. Natural Language Processing (NLP)

↳ NLP adalah teknologi machine learning yang memungkinkan komputer memahami, menginterpretasikan, dan memproses bahasa manusia.

- Manfaat : Mengelola data teks/ucapan secara otomatis untuk analisis sentimen, chatbot, klarifikasi teks, dll.

- Cara kerja : Kombinasi Linguistik komputasional, machine learning dan deep learning.

Tahapannya meliputi :

- > Preprocessing : Tokenisasi, stemming, lemmatisasi, dan penghapusan kata henti.
- > Training : Melatih model menggunakan data teks.
- > Deployment : Integrasi model ke lingkungan produksi.

- Tugas NLP : Speech recognition, machine translation, sentiment analysis, dan named entity recognition.

2. Deep Learning

- ↳ Subbidang machine learning yang berfokus pada pembelajaran pola kompleks dari data melalui representasi hierarki.
- Kelebihan: Mengotomatiskan ekstraksi fitur dibandingkan metode konvensional
- Arsitektur CNN (Convolutional Neural Network):
 - > Memiliki layer utama: convolutional, pooling, dan fully connected.
 - > Aplikasi: klasifikasi gambar, deteksi objek, segmentasi.
- Arsitektur RNN (Recurrent Neural Network):
 - > Cocok untuk data sequential seperti teks, suara, dan time series.
 - > Varian: LSTM, GRU, Bi-directional RNN, dll.
 - > Aplikasi: klasifikasi sentimen, chatbot, prediksi waktu, dll.

3. Pendekatan Neural Network

- Neural network digunakan untuk menyelesaikan masalah yang kompleks, seperti klasifikasi data, prediksi, dan pengenalan pola.

4. Karakteristik Masalah yang Dapat Diselesaikan

- NLP, Deep Learning, dan Neural Network digunakan untuk menangani data tidak terstruktur seperti teks, ucapan, dan gambar.