Nama : Satya Athaya Daniswara

NIM : 1103213152

TUGAS PERBAIKAN BAB 4

1. Pentingnya Representasi Data

- Garbage In, Garbage Out: Kualitas model machine learning sangat bergantung pada kualitas data yang digunakan. Jika data tidak diwakili dengan baik, maka performa model akan menurun. Oleh karena itu, data harus diubah menjadi representasi numerik yang dapat diproses oleh algoritma machine learning.

2. Apa Itu Fitur Engineering?

- Feature Engineering: Proses ini melibatkan transformasi data mentah menjadi fitur yang dapat digunakan oleh algoritma. Proses ini mencakup:

  - Seleksi Fitur: Memilih fitur yang relevan untuk model.

  - Transformasi Data: Mengubah atau mengkodekan data menjadi bentuk yang lebih bermakna.

  - Pembuatan Fitur Baru: Menghasilkan fitur baru yang mengandung informasi lebih banyak.

3. Representasi Numerik untuk Data Kategori

- Data Kategori: Data kategori sering muncul dalam dataset, seperti jenis kelamin, kota, atau preferensi. Dua metode umum untuk menangani data kategori adalah:

  - One-Hot Encoding: Mengubah kategori menjadi vektor biner, di mana setiap kategori diwakili oleh satu kolom.

  - Ordinal Encoding: Memberikan nilai numerik pada kategori, yang hanya cocok jika kategori memiliki urutan logis. Misalnya, kategori seperti "Kecil", "Sedang", dan "Besar" dapat direpresentasikan sebagai [1, 2, 3].

4. Binning dan Discretization

- Binning: Proses membagi fitur numerik menjadi beberapa kategori atau interval. Contohnya, usia dapat dibagi menjadi kelompok (0-18, 19-35, 36-50, dst.).

- Discretization: Membantu algoritma dalam menangkap pola non-linear. Model berbasis pohon keputusan tidak memerlukan binning karena dapat menangani fitur dengan skala kontinu.

5. Interaksi dan Polinomial

- Interaksi: Menggabungkan dua atau lebih fitur untuk menciptakan fitur baru. Contohnya, menggabungkan "Lokasi" dan "Waktu" untuk menghasilkan fitur baru seperti "Waktu Lokasi Spesifik".

- Polinomial: Menambahkan pangkat dari fitur, seperti x^2, x^3, dst. Ini memungkinkan model linier sederhana untuk menangkap hubungan non-linear dengan menambahkan fitur polinomial.

6. Transformasi Non-Linear Univariate

- Transformasi Non-Linear: Diterapkan pada fitur individu untuk menangani distribusi yang tidak normal atau pola non-linear. Contoh transformasi meliputi:

  - Log: Untuk data dengan distribusi eksponensial.

  - Akar Kuadrat: Untuk mengurangi efek outlier.

  - Eksponensial: Untuk meningkatkan fitur dengan nilai rendah.

7. Pemilihan Fitur Otomatis

- Proses Pemilihan Fitur: Memilih subset fitur yang paling relevan untuk model secara otomatis. Ini membantu mengurangi kompleksitas model, meningkatkan akurasi, dan mengurangi risiko overfitting. Terdapat tiga metode utama:

  - Statistik Univariate: Memilih fitur berdasarkan hubungan statistik antara fitur dan target.

  - Pemilihan Fitur Berbasis Model: Menggunakan model machine learning untuk menilai pentingnya fitur.

  - Pemilihan Fitur Iteratif: Memilih fitur secara iteratif berdasarkan performa model.

8. Statistik Univariate

- Statistik Univariate: Memilih fitur berdasarkan hubungan statistik antara fitur dan target. Contohnya, menggunakan pengujian statistik seperti χ^2 atau F-test untuk memilih fitur yang paling relevan.

9. Pemilihan Fitur Berbasis Model

- Model-Based Feature Selection: Menggunakan model machine learning untuk menilai pentingnya fitur. Contohnya, pohon keputusan atau model berbasis ensemble seperti Random Forest dapat memberikan nilai penting fitur. Fitur dengan kontribusi kecil dapat dihapus untuk menyederhanakan model.

10. Pemilihan Fitur Iteratif

- Iterative Feature Selection: Memilih fitur secara iteratif berdasarkan performa model. Metode ini mencakup:

  - Forward Selection: Menambahkan fitur satu per satu ke model.

  - Backward Elimination: Menghapus fitur satu per satu dari model.

  - Recursive Feature Elimination (RFE): Menghapus fitur dengan kontribusi terkecil secara iteratif.

11. Representasi untuk Data Teks

- Data Teks: Data teks perlu diubah menjadi vektor numerik untuk digunakan dalam model. Metode yang umum digunakan termasuk:

  - Bag-of-Words: Menghitung frekuensi kata dalam dokumen.

  - TF-IDF (Term Frequency-Inverse Document Frequency): Mengukur relevansi kata berdasarkan frekuensi relatif.

12. Representasi untuk Data Gambar

- Data Gambar: Data gambar sering diubah menjadi array numerik berdasarkan nilai piksel. Teknik preprocessing tambahan, seperti pengubahan ukuran atau normalisasi intensitas piksel, sering digunakan.

13. Pipelines untuk Preprocessing

- Pipelines: Memungkinkan preprocessing dan pelatihan model dilakukan dalam satu langkah. Dengan menggunakan Pipeline dari scikit-learn, Anda dapat memastikan bahwa langkah preprocessing diterapkan secara konsisten.

Kesimpulan

Bab ini menekankan pentingnya representasi data dan rekayasa fitur dalam meningkatkan performa model machine learning. Berbagai teknik, seperti one-hot encoding, interaksi, dan transformasi non-linear, memberikan fleksibilitas dalam menangani berbagai jenis data. Pemilihan fitur, baik secara otomatis maupun manual, memainkan peran penting dalam mengurangi kompleksitas dan meningkatkan efisiensi model.