Rakamin Academy

Homework

Machine Learning Preparation

Estimasi Waktu Pengerjaan

 3 - 5 jam

Jumlah Soal

 2 Soal

Total Point

 100 poin

Teknis Pengerjaan

1. Pekerjaan dilakukan secara **berkelompok, sesuai kelompok Final Project**

2. Masing-masing anggota kelompok tetap perlu submit ke LMS (jadi bukan perwakilan)

3. File yang perlu dikumpulkan:
   - File **jupyter notebook** (.ipynb) yang berisi source code.
   - File **slides** (.pdf) simple slides presentasi yang berisi rangkuman dari apa saja yang telah dilakukan.

4. Upload hasil pengerjaanmu melalui LMS.
   - Masukkan semua file ke dalam **1 file** dengan format **ZIP**.
   - Nama File:

     `ML Preparation - <Nama Kelompok>.zip`

Product Classification

- **Deskripsi**:

  Memprediksi apakah suatu produk eksklusif atau tidak berdasarkan fitur yang tersedia

- **Link** : Dataset, template python notebook (optional)

1. Descriptive Statistics (5 poin)

Gunakan function **`info`** dan **`describe`** pada dataset final project kalian. Tuliskan hasil observasinya, seperti:

A. Apakah ada kolom dengan tipe data kurang sesuai, atau nama kolom dan isinya kurang sesuai?
B. Apakah ada kolom yang memiliki nilai kosong? Jika ada, apa saja?
C. Apakah ada kolom yang memiliki nilai summary agak aneh? (min/mean/median/max/unique/top/freq)

* Untuk masing-masing jenis observasi, tuliskan juga jika tidak ada masalah, misal untuk A: “Semua tipe data sudah sesuai”

2. Univariate Analysis (10 poin)

Gunakan visualisasi untuk melihat distribusi masing-masing kolom (feature maupun target). Tuliskan hasil observasinya, misalnya jika ada suatu kolom yang distribusinya menarik (misal skewed, bimodal, ada outlier, ada nilai yang mendominasi, kategorinya terlalu banyak, dsb). Jelaskan juga apa yang harus di-follow up saat data pre-processing.

3. Multivariate Analysis (15 poin)

Lakukan multivariate analysis (seperti correlation heatmap dan category plots, sesuai yang diajarkan di kelas). Tuliskan hasil observasinya, seperti:

A. Bagaimana korelasi antara masing-masing feature dan label. Kira-kira feature mana saja yang paling relevan dan harus dipertahankan?
B. Bagaimana korelasi antar-feature, apakah ada pola yang menarik? Apa yang perlu dilakukan terhadap feature itu?

* Tuliskan juga jika memang tidak ada feature yang saling berkorelasi

4. Data Cleansing (40 poin)

Lakukan pembersihan data, sesuai yang diajarkan di kelas, seperti:

A. Handle missing values
B. Handle duplicated data
C. Handle outliers
D. Feature transformation
E. Feature encoding
F. Handle class imbalance

Di laporan homework, tuliskan apa saja yang telah dilakukan dan metode yang digunakan.

* Tetap tuliskan jika memang ada tidak yang perlu di-handle (contoh: “Tidak perlu feature encoding karena semua feature sudah numerical” atau “Outlier tidak di-handle karena akan fokus menggunakan model yang robust terhadap outlier”).

5. Feature Engineering (30 poin)

Cek feature yang ada sekarang, lalu lakukan:

A. Feature selection (membuang feature yang kurang relevan atau redundan)

B. Feature extraction (membuat feature baru dari feature yang sudah ada)

C. Tuliskan minimal 4 feature tambahan (selain yang sudah tersedia di dataset) yang mungkin akan sangat membantu membuat performansi model semakin bagus (ini hanya ide saja, untuk menguji kreativitas teman-teman, tidak perlu benar-benar dicari datanya dan tidak perlu diimplementasikan)

* Untuk 2A & 2B, tetap tuliskan jika memang tidak bisa dilakukan (contoh: “Semua feature digunakan untuk modelling (tidak ada yang dihapus), karena semua feature relevan”)

Rakamin Academy

Selamat Mengerjakan!