Rakamin Academy

Homework
Data Pre-Processing

Final Project - Stage 2

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 **laporan homework** (.pdf) 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:

     `Preprocessing - <Nama Kelompok>.zip`

1. Data Cleansing (50 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

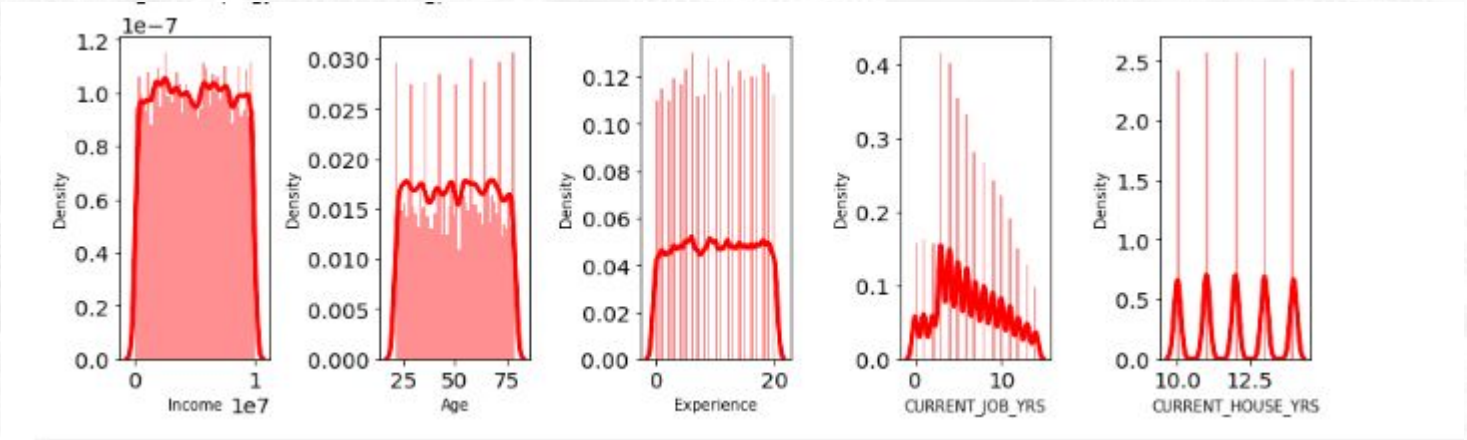
Di laporan homework, tuliskan apa saja yang telah dilakukan dan metode yang digunakan.

* 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”).

1. Data Cleansing

- Handle missing values, Handle duplicated data, Handle outliers : Tidak ada missing value, duplicated data, dan Outlier tidak di-handle karena data sudah bagus.

- Feature transformation: pada stage kali ini kelompok 5 melakukan feature transformation menggunakan pipeline dan juga tanpa pipeline menghasilkan distribusi yang sama. Distribusi datanya uniform dan multimodal, maka data tidak akan ditransformasi, karena tidak akan terjadi perubahan (mendekati normal).



- Untuk memastikan konsistensi data, pipeline harus menyertakan setiap langkah yang diperlukan untuk melatih dan menilai set data pelatihan dan pengujian.

- Setelah *feature engineering pipeline* dibuat untuk data numerik dan kategorikal, kita akan menggunakan kelas ColumnTransformer untuk menggabungkannya menjadi satu preprosesor. Kita akan menggunakan fitur numerik dan kategorik yang kita buat sebelumnya untuk menentukan kolom mana yang harus diteruskan ke pipa mana.

- Feature Encoding : feature encoding yang digunakan adalah binary encoding pada feature state dan profession karena pada feature ini memiliki banyak nilai kategori yang berbeda-beda. Jika menggunakan one-hot encoding untuk kedua feature ini, maka akan menghasilkan banyak fitur yang dapat menyebabkan masalah overfitting. Binary encoding mengonversi setiap nilai kategori menjadi representasi biner (0 dan 1) dan memungkinkan setiap fitur kategori direpresentasikan dalam jumlah kolom yang lebih sedikit. Dengan jumlah kolom yang lebih sedikit, dapat mengurangi dimensi data dan menghindari masalah overfitting

- Selanjutnya dilakukan scalling pada data numerik untuk untuk mengubah rentang nilai dari masing-masing fitur sehingga memiliki skala yang sama. Proses scalling ini dilakukan setelah proses split data train dan data test, karena melakukan scaling sebelum proses split dapat mengakibatkan kebocoran informasi dari data test ke data train, sehingga model yang dihasilkan dapat mengalami overfitting pada data test.

- Handle Class Imbalance : Pada dataset loan prediction, kelas minoritas dapat dianggap sebagai kelas yang mewakili pelanggan yang memiliki risiko kredit yang lebih tinggi (Risk_Flag = 1), sementara kelas mayoritas mewakili pelanggan yang memiliki risiko kredit yang lebih rendah (Risk_Flag = 0). Class imbalance dapat menyebabkan model menjadi bias terhadap kelas mayoritas dan memiliki performa yang buruk dalam mengenali kelas minoritas. Oleh karena itu, strategi yang digunakan untuk menangani masalah class imbalance adalah dengan menggunakan class weighting. Class weighting adalah teknik yang digunakan untuk memberikan bobot yang berbeda pada setiap kelas. Dalam kasus dataset loan prediction, kita dapat memberikan bobot yang lebih tinggi pada kelas minoritas (Risk_Flag = 1) dan bobot yang lebih rendah pada kelas mayoritas (Risk_Flag = 0). Dengan memberikan bobot yang berbeda, model akan lebih sensitif terhadap kelas minoritas dan akan lebih cenderung untuk memprediksi dengan benar kelas yang kurang diwakili dalam dataset. Dalam kasus ini, penggunaan class weighting pada dataset loan prediction dapat membantu meningkatkan performa model dalam mengenali pelanggan yang memiliki risiko kredit yang lebih tinggi, yang merupakan kelas minoritas

2. Feature Engineering (35 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”)

- Feature Selection :
- Feature Extraction : Tidak ada fitur baru yang dibuat, tetapi ada pecahan dari beberapa fitur kategorikal setelah dilakukan encoding
- Feature Tambahan : Fitur yang mungkin perlu ditambahkan : Total amount of loan, Uda berapa kali loan di tempat ini (payment/loan history), Pengeluaran Bulanan, Banyaknya pengajuan pinjaman setiap bulannya., customer creadit score, the lenght of time of the customer has been employed.

3. Git (15 poin)

Upload project teman-teman di sebuah repository git. Berkolaborasilah di Git jika ada perubahan version dari waktu ke waktu.

A. Buat Repository Git
B. Upload file notebook atau file pengerjaan lainnya pada repository tersebut

Untuk file README, dapat merupakan summary dari proses data preproses yang telah dilakukan. Boleh menggunakan repositori yang sama atau membuat baru.

https://github.com/drestantav/raka_project_repo