Praktikum IF3270 Pembelajaran Mesin
Customer Segmentation Classification

Dipersiapkan Oleh Tim Asisten IF3270 2024/2025

Versi: 1.0 10/03/2025

Deadline 1: Kamis, 13 Maret 2025 11.00 WIB
Deadline 2: Kamis, 13 Maret 2025 23:00 WIB

Tujuan

Praktikum pada kuliah IF3270 Pembejalaran Mesin bertujuan untuk memberikan pengalaman langsung kepada peserta kuliah dalam menerapkan rangka kerja CRISP-DM pada permasalahan nyata.

Spesifikasi

Sebuah perusahaan otomotif berencana memasuki pasar baru dengan produk yang sudah ada (P1, P2, P3, P4, dan P5). Setelah melakukan riset pasar yang intensif, mereka menyimpulkan bahwa perilaku pasar baru serupa dengan pasar yang sudah ada.

Di pasar yang sudah ada, tim penjualan telah mengklasifikasikan semua pelanggan ke dalam 4 segmen (A, B, C, D). Kemudian, mereka melakukan pendekatan dan komunikasi yang tersegmentasi untuk setiap segmen pelanggan. Strategi ini terbukti sangat berhasil bagi mereka. Perusahaan berencana menggunakan strategi yang sama di pasar baru dan telah mengidentifikasi 2.627 calon pelanggan baru.

Pada praktikum ini, Anda akan diminta untuk memprediksi **Segmentasi** dari *potential customer*. Berikut deskripsi dari tiap fitur yang ada pada dataset yang akan digunakan:

1. *ID: Unique identifier for each individual.*
2. *Gender: Gender of the individual (Male/Female).*
3. *Ever_Married: Whether the individual has ever been married (Yes/No).*
4. *Age: Age of the individual.*
5. *Graduated: Whether the individual is a graduate (Yes/No).*
6. *Profession: Profession of the individual (e.g., Healthcare, Engineer, Lawyer, etc.).*
7. *Work_Experience: Number of years of work experience.*
8. *Spending_Score: Spending score of the individual (Low/Average/High).*
9. *Family_Size: Size of the individual's family.*
10. *Var_1: Redacted categorical variable with multiple categories (e.g., Cat_1, Cat_2, etc.).*
11. *Segmentation: Target variable indicating the segment to which the individual belongs (A, B, C, D).*

Berikut merupakan beberapa hal yang harus dilakukan oleh praktikan:

EDA (Exploratory Data Analysis)

Exploratory Data Analysis (EDA) adalah langkah krusial dalam proses analisis data yang melibatkan pemeriksaan dan visualisasi dataset untuk mengungkap pola, tren, anomali, dan *insight*. Ini merupakan langkah awal sebelum menerapkan teknik statistik dan machine learning yang lebih lanjut. EDA membantu Anda memahami data secara mendalam, memungkinkan Anda membuat keputusan yang lebih terarah dan dapat merumuskan hipotesis untuk analisis lebih lanjut.

Pada tahap ini, Anda diminta untuk melakukan EDA dengan membuat visualisasi serta penjelasan terkait dataset yang dimiliki. Gunakan visualisasi yang tepat dilengkapi dengan penjelasan untuk setiap pertanyaan atau pernyataan yang ingin disampaikan. Berikan minimal **3 pertanyaan** yang kemudian dijawab dengan analisis dan dibantu dengan visualisasi.

Data Cleaning and Preprocessing

Langkah ini adalah hal pertama yang dilakukan setelah seorang Data Scientist memiliki pemahaman umum tentang data. Data mentah jarang siap untuk training, sehingga perlu dilakukan langkah-langkah untuk membersihkan dan memformat data agar dapat diinterpretasikan oleh model machine learning.

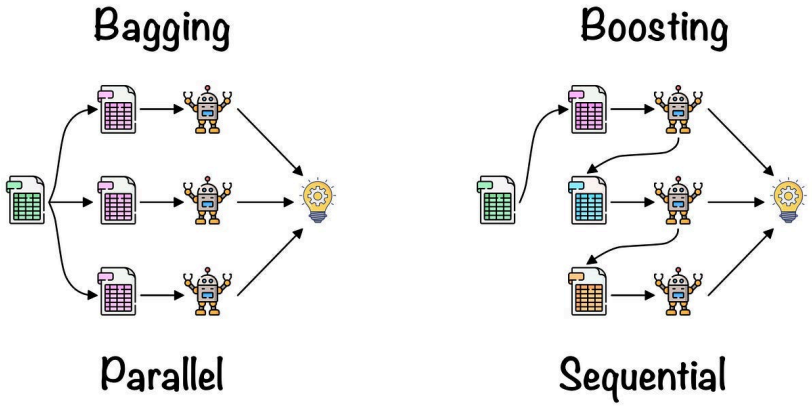
Dengan melakukan data cleaning dan preprocessing, Anda memastikan bahwa dataset siap untuk training model, yang akan menghasilkan hasil machine learning yang lebih akurat dan andal. Langkah-langkah ini sangat penting untuk mentransformasi data mentah menjadi format yang dapat dipelajari secara efektif oleh algoritma machine learning dan digunakan untuk membuat prediksi.

Pada tahap ini, Anda diminta untuk melakukan proses-proses yang sekiranya tepat digunakan untuk dataset yang diberikan. Sertakan penjelasan mengapa Anda memilih untuk melakukan proses tersebut.

Modeling and Validation

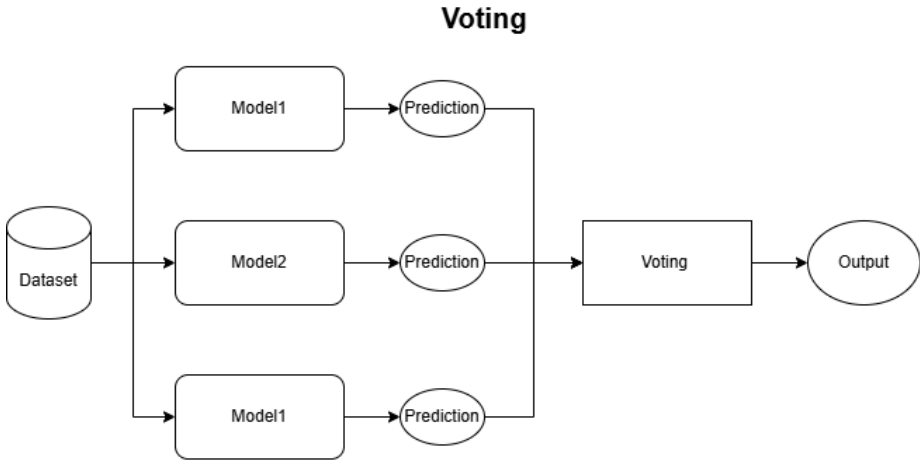
Modeling adalah proses membangun model machine learning untuk menyelesaikan masalah tertentu, atau dalam konteks tugas ini, **memprediksi setiap kelas dalam fitur *Segmentation* menggunakan algoritma Ensemble Learning**.

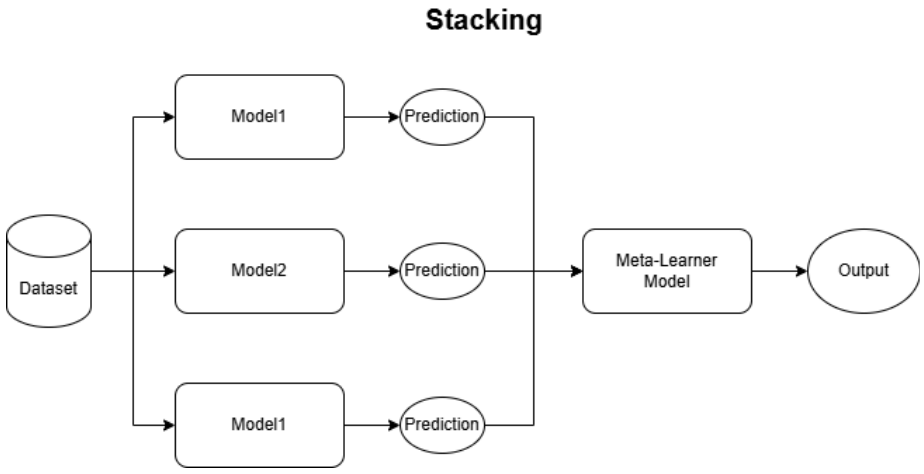
Ensemble Learning terdiri dari 4 teknik berikut:



1. Bagging
   Bagging merupakan teknik yang melatih satu tipe model, dengan setiap model dilatih dengan *subset* data yang berbeda. Dengan prediksi akhir merupakan prediksi paling sering.

2. Boosting
   Boosting merupakan teknik yang melatih model pertama dengan seluruh dataset, lalu dibuat sebuah model kedua yang dibuat untuk menurunkan error model pertama, lalu dibuat model ketiga untuk menurunkan error model kedua, dan seterusnya.





3. Voting
   Voting merupakan teknik yang melatih beberapa model **berbeda**, dengan setiap model dilatih dengan **seluruh** dataset. Prediksi akhir didapatkan melalui **voting.** Voting merupakan suatu heterogeneous ensemble (terdiri dari model yang berbeda-beda) ditambah dengan weighting.

4. Stacking
   Stacking merupakan teknik yang melatih beberapa model **berbeda**, dengan setiap model dilatih dengan **seluruh** dataset. Dengan prediksi akhir didapatkan melalui sebuah **meta-learner model** yang menerima prediksi model-model sebelumnya. Stacking merupakan suatu heterogeneous ensemble ditambah dengan meta learning.

Pada tahap ini Anda diminta untuk melakukan pemodelan **minimal 1 model** untuk setiap algoritma *ensemble learning* berikut:

- Bagging, misalnya Random Forest, dll
- Boosting, misalnya Gradient Boosting, dll
- Stacking
- Voting

Validation adalah proses mengevaluasi model yang telah dilatih menggunakan validation set atau metode cross-validation dan memberikan metrik yang dapat membantu Anda menentukan langkah yang perlu dilakukan pada iterasi pengembangan berikutnya.

Untuk validasi, metrik yang digunakan adalah **macro f1-score**. Hasil validasi yang harus tercantum di notebook adalah **hasil dari pemodelan yang wajib diimplementasikan** dan **hasil dari pemodelan submisi final di kaggle**.

Error Analysis

Error analysis adalah proses menganalisis kesalahan yang dihasilkan oleh model machine learning untuk memahami penyebab utama ketidaktepatan prediksi. Langkah ini bertujuan untuk mengidentifikasi pola dalam kesalahan, seperti kategori data tertentu yang sering salah diklasifikasikan atau fitur yang memberikan kontribusi rendah terhadap akurasi model.

Submisi Kaggle

Pada tugas ini, Anda diminta untuk mengikuti Kaggle Competition berikut dan mengumpulkan hasil prediksi Anda ke kompetisi tersebut. Dataset yang akan digunakan tersedia pada tautan Kaggle Competition yang diberikan. Skor yang Anda raih di kompetisi tersebut akan masuk ke dalam penilaian. Skor yang diambil untuk penilaian adalah skor yang diraih di private leaderboard. Submisi ke Kaggle dibatasi sebanyak **20 kali.** Submisi final yang dikumpulkan ke kaggle harus ***reproducible*** (dapat dihasilkan hasil yang sama dengan notebook yang dikumpulkan). Submisi yang **tidak *reproducible*** akan dikenakan **penalti berupa pengurangan nilai**. Format penamaan kelompok di Kaggle adalah sebagai berikut: **Kelas_NomorKelompok_NamaKelompok** (Contoh: K1_99_gAIB21, nama kelompok opsional)

Asisten telah menyiapkan *notebook template* untuk Anda gunakan dengan tautan sebagai berikut.

Kelompok

Pembagian kelompok ditentukan sendiri oleh mahasiswa dengan mengisi sheets kelompok berikut ini dengan 1 kelompok terdiri dari dengan **maksimal anggota sebanyak 2 orang**.

QnA

Pertanyaan dapat ditanyakan pada **link QnA** berikut. Pastikan pertanyaan yang ditanyakan tidak berulang.

Aturan

Terdapat beberapa hal yang harus diperhatikan dalam pengerjaan tugas ini, yakni:

1. Jika terdapat hal yang tidak dimengerti, silahkan ajukan pertanyaan kepada asisten melalui **link QnA** yang telah diberikan di atas. Pertanyaan yang diajukan secara

personal ke asisten **tidak akan dijawab** untuk menghindari perbedaan informasi yang didapatkan oleh peserta kuliah.

2. Dilarang melakukan **plagiarisme, menggunakan AI dalam bentuk apapun secara tidak bertanggungjawab, dan melakukan kerjasama antar kelompok**. Pelanggaran pada poin ini akan menyebabkan pemberian **nilai E** pada setiap anggota kelompok.
3. Tidak ada batasan untuk *library* yang boleh digunakan.
4. Dilarang untuk menggunakan **SampleSubmission.csv** sebagai **submisi utama**. Submisi utama harus menggunakan hasil prediksi dengan model sendiri. Penggunaan SampleSubmission.csv sebagai submisi final akan diberikan penalti berupa **pemberian nilai 0 untuk praktikum**.

Deliverables

- **Deadline 1 (Kamis, 13 Maret 2025 11.00 WIB)**
  - Link ke notebook pengerjaan dengan format penamaan file NIM1_NIM2_Kelas_Deadline1.ipynb
    (Contoh: 13521998_13521999_K1_Deadline1.ipynb)
  - Minimal mengumpulkan **1 kali submisi** ke kaggle, dibuktikan dengan *screenshot* leaderboard yang menunjukkan kelompok Anda sudah tercantum pada leaderboard.
- **Deadline 2 (Kamis, 13 Maret 2025 23:00 WIB)**
  - Link ke notebook pengerjaan dengan format penamaan file NIM1_NIM2_Kelas_Deadline2.ipynb
    (Contoh: 13521998_13521999_K1_Deadline2.ipynb). Jika terdapat revisi atau tambahan dari notebook pertama, silakan tambah markdown cell di notebook pengerjaan yang berisi daftar perubahan yang dilakukan.
  - Notebook yang dikumpulkan harus sudah lengkap untuk seluruh bagian beserta dengan analisisnya:
    - EDA
    - Data Preprocessing
    - Modeling and Validation
    - Error Analysis
- Notebook pengerjaan yang dikumpulkan harus memiliki akses **Editor** untuk **Anyone with the link**.
- Pastikan notebook yang dikumpulkan untuk Deadline 1 dan Deadline 2 merupakan **notebook yang berbeda**.
- Pengumpulan dilakukan melalui form dengan tautan sebagai berikut.
- Tugas yang terlambat dikumpulkan tidak akan diterima.
- Pengumpulan dilakukan oleh NIM terkecil.

Referensi

- https://scikit-learn.org/stable/modules/ensemble.html#ensemble
- https://scikit-learn.org/stable/api/sklearn.ensemble.html
- https://scikit-learn.org/stable/modules/ensemble.html#random-forests-and-other-randomized-tree-ensembles
- https://scikit-learn.org/stable/modules/ensemble.html#gradient-boosted-trees
- https://scikit-learn.org/stable/modules/ensemble.html#stacked-generalization
- https://scikit-learn.org/stable/modules/ensemble.html#voting-classifier