Laporan Kerja Individu

Final Project - Data Analytics

Program Zenius Studi Independen Bersertifikat - Angkatan 4

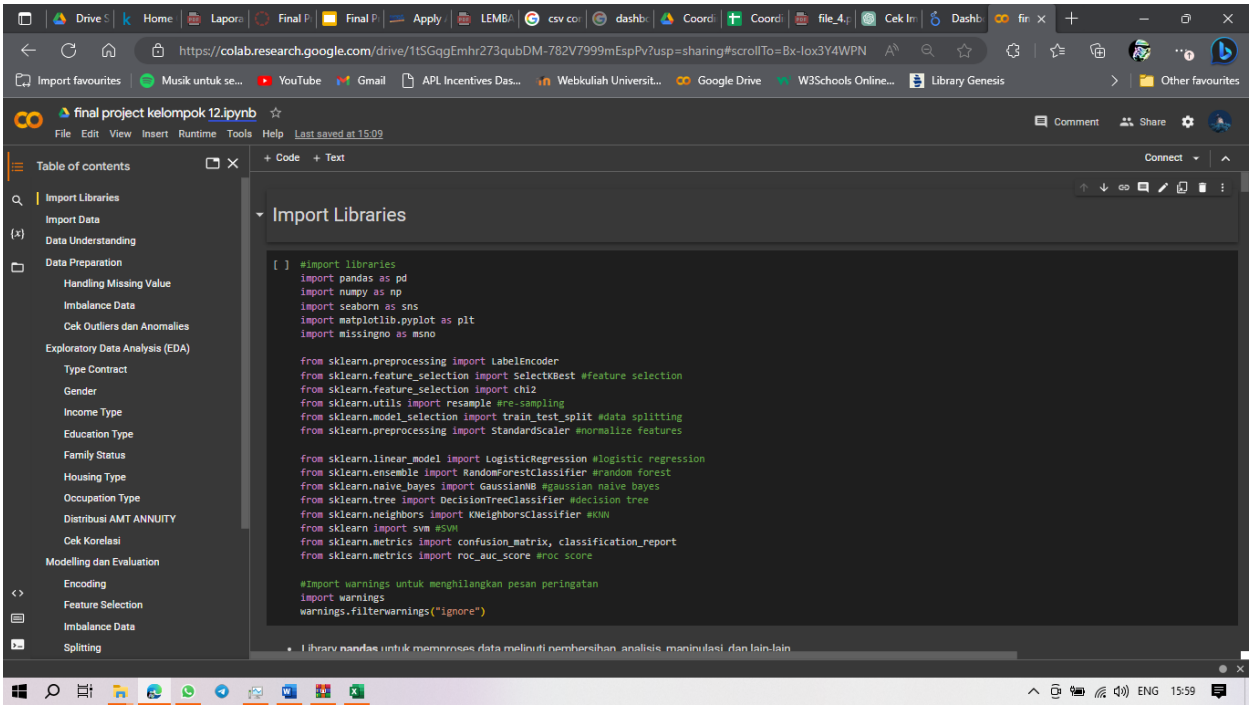
Nama Lengkap: Raihan Tsabita Sabil
Nomor ID live class: 141
Nomor Kelompok: 12
Mentor: Muhammad Verly

Deskripsi peran

Pada bagian saya, saya mengerjakan bagian data understanding dan data preparation. Berikut hal-hal yang saya lakukan:

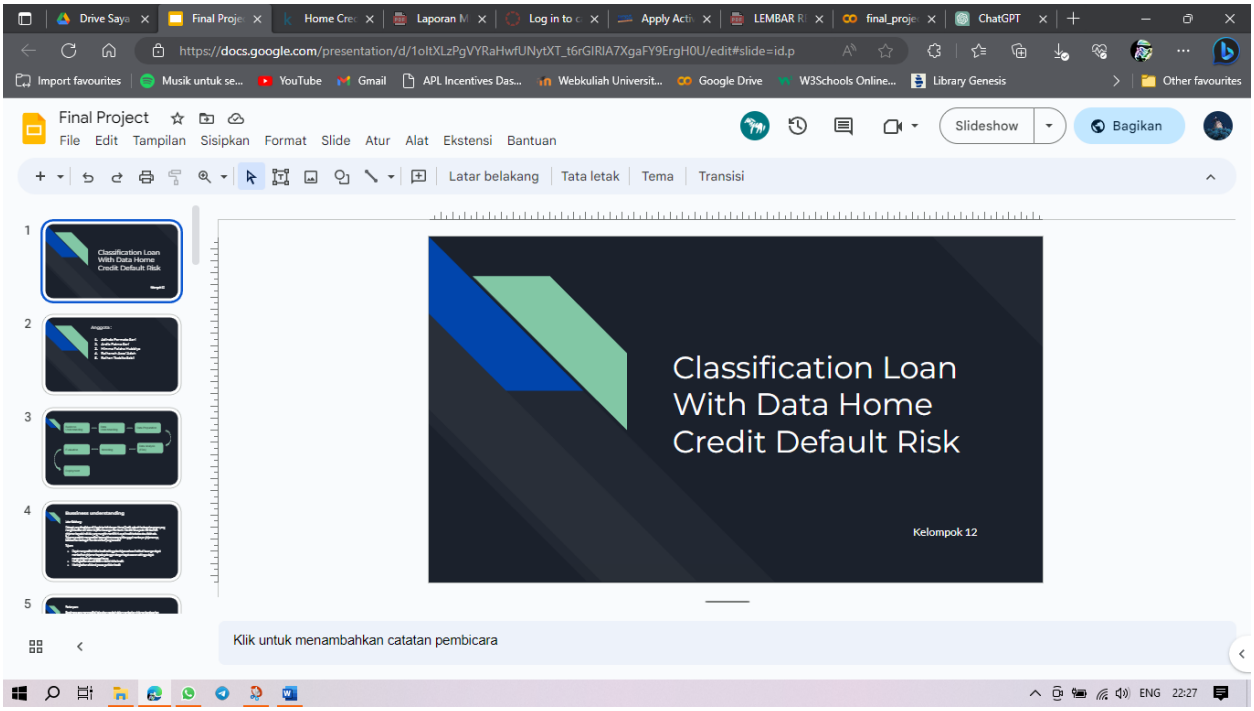
- Mengidentifikasi dan memahami sumber data yang tersedia.
- Menganalisis dan mengevaluasi kualitas data yang ada.
- Menyusun dan membersihkan data agar dapat digunakan dalam analisis.
- Melakukan eksplorasi data untuk mendapatkan wawasan awal dan memahami hubungan antar variabel.
- Menyiapkan data yang relevan dan terstruktur untuk analisis lebih lanjut.
- Mengolah data menjadi format yang sesuai dengan tujuan analisis kelompok.
- Membantu membuat presentasi
- Membantu membuat dashbaord
- Membantu merekam presentasi untuk bukti tugas

Lampiran Hasil Kerja



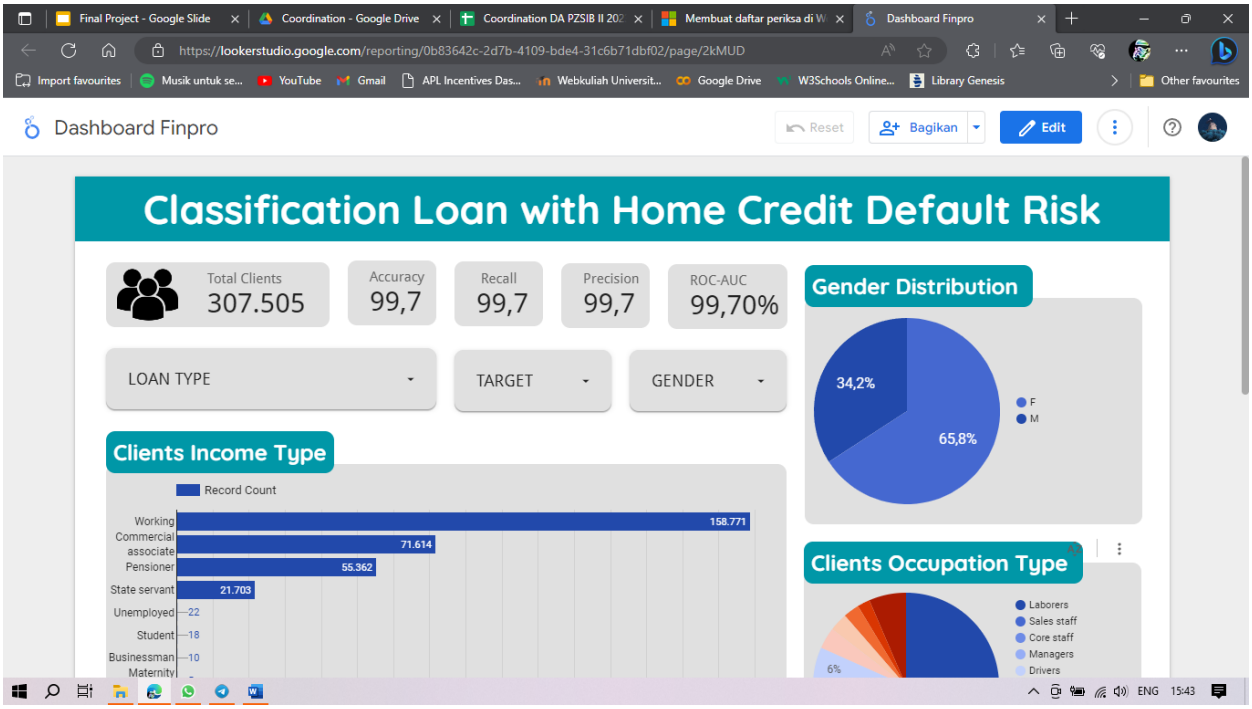
Membantu mengerjakan pada bagian data understanding dan data preparation

Link Gcollab: https://colab.research.google.com/drive/1tSGqgEmhr273qubDM-782V7999mEspPv?usp=sharing



Membantu membuat slide presentasi

Link PPT:
https://docs.google.com/presentation/d/1oItXLzPgVYRaHwfUNytXT_t6rGIRlA7XgaFY9ErgH0U/edit?usp=sharing



Membantu membuat dashboard
Link dashboard: https://lookerstudio.google.com/reporting/0b83642c-2d7b-4109-bde4-31c6b71dbf02

Proses Kerja

Pada proses kerja ini, pertama setelah ditentukannya kelompok, kami berdiskusi terkait project yang akan dilaksanakan. Dimulai dari awal pengerjaan yaitu bussiness understanding. Pada tahap ini kita memahami latar belakang dari dataset perusahaan, tujuan dari menganalisa dataset tersebut, pertanyaan apa saja yang akan ada pada dataset tersebut dan bagi solusi dari pertanyaan tersebut. Dari dataset yang disediakan, memiliki latar belakang sebagai perusahaan yang menyediakan layanan kredit dimana kita diminta untuk memprediksi kemungkinan klien gagal bayar berdasarkan fitur-fitur yang ada. Adapun tujuan lain dari menganalisa data ini adalah meningkatkan akurasi dalam menilai risiko kredit dan meningkatkan efisiensi proses penilaian kredit.

Setelah berdiskusi mengenai bussiness understanding, dilanjut ke tahap data understanding. Dari data understanding ini, kita memahami data mana saja yang akan kita pakai dalam menganalisa nanti. Berdasarkan hasil diskusi, menggunakan dataset utamanya yaitu application train. Sebelum masuk ke pengolahan data, kita perlu mengimpor library apa

saja yang digunakan. Ada beberapa library utama yang kita digunakan untuk mengulah dan memvisualisasikan data seperti numpy, pandas, seaborn dan matplotlib. Serta menggunaka beberapa library lain tambahan untuk tahap modeling dan prediksi nanti. Setelah itu, barulah masuk ke pengolahan data. Kita impor dataset application yang akan digunakan melalui metode import dari google drive, kemudian membuat variabel yang memuat data tersebut dengan nama 'train_data'. Setlah import, dilanjutkan melihat isi data dahulu, adapun perintah yang digunakan seperti:

- *.head()* : melihat 5 baris data dari atas
- *.shape* : melihat ukuran dataset
- *.columns.values* : melihat nama-nama kolom
- *.duplicated()* : mengecek data yang duplikat
- *.info()* : melihat info data serta tipe data

Setelah dilihat isi datanya, kita bagi data tersebut menjadi 2 jenis yaitu data kolom categorical untuk tipe data object dan data kolom numerik untuk data tipe angka. Tidak lupa juga untuk mengecek adakah data yang duplikat menggunakan perintah *train_data[train_data.duplicated()]* .Kemudian dicek kembali menggunakan *train_data[col].value_counts()* untuk menghitung nilai unique values pada variabel kategorik.

Setelah itu masuk kedalam data preparation, disini pertama kita cek terlebih dahulu berapa banyak data kolom yang kosong dalam dataset train lalu urutkan data yang kosong tersebut dari yang terbesar ke terkecil. Kemudian hitung berapa persen data kolomnya yang hilang dan total nya. Setelah dicek berapa banyak data yang kosong, kita handle missing valuenya dengan langkah pertama kita drop terlebih dahulu data kolom yang missingnya lebih dari 50%. Setelah langkah tersebut, dilanjutkan mengisi data kolom kosong yang dibawah 50% dengan ketentuan untuk tipe kolom kategorikal diisi data kosongnya menggunakan modus dan untuk numerik diisi dengan menggunak mean. Kemudian dicek kembali untuk memastikan bahwa data kosong tersebut telah terisi. Langkah selanjutnya kita melihat imbalance data dalam dataset tersebut dengan kolom utamanya yaitu 'TARGET' untuk melihat seberapa jauh ketimpangan dalam data tersebut. Dan langkah selanjutnya adalah cek outlier dan anomalinya dari dataset tersebut.

Pada tahap selanjutnya iyalah Exploratory Data Analysis (EDA) yang dilakukan oleh teman saya. Pertama dilakukan mencari berapa banyak data yang terbagi pada tipe kontrak, gender, tipe income, tipe edukasi, status keluarga, tipe rumah, dan tipe okupasi. Pada 7 data yang diexplore tersebut semua diurutkan berdasarkan target dan yang terbesar, kemudian divisualisasikan menggunakan table, bar plot dan pie chart sesuai tipe datanya. Kemudian selanjutnya distribusi kredit yang dieksplor, dicari rentang rata-rata kreditnya kemudian divisualisasikan dengan plot. Dengan dieksplornya distribusi kredit tersebut, bisa dilihat yang gagal bayar berdasarkan pengelompokan kreditnya. Begitupun sama yang di ekslpor pada data berdasarkan days from birth dan AMT ANNUITY. Dan terakhir pada tahap ini adalah mengecek

korelasi antar fitur dimana batas tinggi korelasi adalah 0.05 dan korelasi rendah adalah -0.2 dan dikelompokan berdasarkan yang tinggi dan yang rendah.

Tahap selanjutnya yang dikerjakan oleh teman saya adalah modelling dan evaluasi. Pertama dilakukan encoding untuk mengubah data yang bertipe object ke numerik, kemudian melakukan feature selection dimana cara ini digunakan untuk mengambil data yang berkorelasi tinggi. Kemudian cara selanjutnya melakukan imbalanced data dengan agar data menjadi seimbang, dan terakhir sebelum melakukan modelling melakukan splitting pada data. Pada tahap selanjutnya adalah melakukan modeliing, disini kita menggunakan 4 modelling yaitu:

- Logistic Regression
- Random Forest Classifier
- Decision Tree
- KNN

Setelah dilakukan 4 modeliing tersebut, langkah selanjutnya adalah mengevaluasi model. Jadi bandingkan model mana yang memiliki tingkat akurasi, presisi, recall dan ROC_AUC yang baik.

Setelah dilakukan pengolahan data, dibuatkan lah presentasi dan dashboard untuk menjabarkan hasil dari pengolahan dan eksplorasi data tersebut.

Persetujuan Anggota

Laporan ini telah disetujui oleh:

- ✓ Adinda Permata Sari
- ✓ Ardia Fatma Sari
- ✓ Himma Faicha Hubbiya
- ✓ Raihanah Assa`Adah