Laporan Kerja Individu

Final Project - Data Analytics

Program Zenius Studi Independen Bersertifikat - Angkatan 4

Nama Lengkap: Zulfa Nabilah Nurvitasari
Nomor ID live class: 099
Nomor Kelompok: 30
Mentor: Erwin Fernanda

Deskripsi peran

- Melakukan *cleansing* dan *preparation* pada dataset *credit card balance* dan *previous application*
- Melakukan *cleansing* dan *preparation* pada dataset *application train* (Semua anggota melakukan *cleansing* dan *preparation.* Oleh karena itu, dari semua bagian *cleansing* dan *preparation* data yang telah saya lakukan, yang digunakan ialah bagian *data encoding*.
- Melakukan pemodelan dengan menggunakan *Supervised Learning* menggunakan KNN, *logistic regression,* dan *random forest*
- Membuat dashboard dengan Looker

Lampiran Hasil Kerja

- *Cleansing* dan *Preparation* pada dataset *credit card balance* dan *previous application*: https://colab.research.google.com/drive/1wyx2ypaaOnYMmLx5s3dTkE0JF76isL6y?usp=sharing
- *Cleansing*, *Preparation*, dan *modeling* pada dataset *application train*: https://colab.research.google.com/drive/1ki0chI6qm01JWFk0_sUftZ_e372jSsmc?usp=sharing
- Dashboard: https://lookerstudio.google.com/reporting/b9b058f8-df07-4f8f-ac71-63d985fb2455

Proses Kerja

No	Hari, Tanggal	Pekerjaan	Keterangan
1	Rabu, 31 Mei	*Cleansing* dan *preparation* pada dataset *credit card balance*.	Saya melakukan data understanding seperti melihat jumlah kolom dan baris pada data, melihat kolom pada dataset, melihat tipe data, melihat jumlah missing value di setiap kolom, dan melihat frekuensi setiap *unique value*s pada data.
2	Kamis, 1 Juni	*Cleansing* dan *preparation* pada dataset *credit card balance.*	Saya melakukan *missing value handling* pada data bertipe numerik dengan mengisi *missing value* menggunakan nilai median karena tidak berdistribusi normal. Selain itu, saya juga melakukan *data encoding* dengan menggunakan *frequency encoding.*
3	Jumat, 2 Juni	*Cleansing* dan *preparation* pada dataset *previous application.*	Saya melakukan data understanding seperti melihat jumlah kolom dan baris pada data, melihat kolom pada dataset, melihat tipe data, melihat jumlah missing value di setiap kolom, dan melihat frekuensi setiap *unique value*s pada data.
4	Sabtu, 3 Juni	*Cleansing* dan *preparation* pada dataset *previous application.*	Saya melakukan *missing value handling* pada data bertipe numerik dan kategorik. Saya menghapus baris dengan kolom yang memiliki missing value kurang dari 10 karena jumlah tidak terlalu besar. Selain itu, saya juga menghapus kolom dengan missing value lebih dari 1000000. Untuk data bertipe numerik yang memiliki *missing value* di bawah 1000000 saya isi dengan mean untuk kolom yang memiliki distribusi normal. Sedangkan, *missing value* pada kolom-kolom yang tidak memiliki distribusi normal, saya isi dengan median. Untuk data bertipe

			kategorik, *missing value* saya isi dengan modus.
5	Minggu, 4 Juni	*Cleansing* dan *preparation* pada dataset *previous application.*	Saya melakukan data encoding pada kolom-kolom bertipe data kategori. Untuk kolom-kolom yang memiliki *unique value* kurang dari 5, saya melakukan *label encoding*, sedangkan untuk kolom-kolom yang memiliki *unique value* lebih dari 5 dilakukan *frequency encoding*.
6	Senin, 5 Juni	*Cleansing* dan *preparation* pada dataset *application train*	Saya melakukan data understanding seperti melihat jumlah kolom dan baris pada data, melihat tipe data, melihat statistik deskriptif pada data, melihat jumlah missing value di setiap kolom, melihat jumlah duplikat, melihat frekuensi setiap *unique value*s pada data, dan juga melakukan pengecekan keseimbangan data. Selain itu, saya juga melakukan unusual value handling dengan mengganti salah satu unique value pada kolom bertipe data kategori dengan modus.
7	Selasa, 6 Juni	*Cleansing* dan *preparation* pada dataset *application train*	Saya melakukan *missing value handling* pada data bertipe numerik dan kategorik. Saya menghapus baris dengan kolom yang memiliki missing value kurang dari 1000 karena jumlah tidak terlalu besar. Selain itu, saya juga menghapus kolom dengan missing value lebih dari 70%. Untuk data bertipe numerik yang memiliki *missing value* di bawah 70% saya isi dengan mean untuk kolom-kolom yang memiliki distribusi normal. Sedangkan, *missing value* pada kolom-kolom yang tidak memiliki distribusi normal, saya isi dengan median.

			Untuk data bertipe kategorik yang memiliki *missing value* di atas 50% saya isi dengan 'unk', sedangkan untuk kolom yang memiliki *missing value* di bawah 50% saya isi dengan modus.
8	Rabu, 7 Juni	*Cleansing* dan *preparation* pada dataset *application train*	Saya melakukan data encoding pada kolom-kolom bertipe data kategori. Untuk kolom-kolom yang memiliki *unique value* kurang dari 5, saya melakukan *label encoding*, sedangkan untuk kolom-kolom yang memiliki *unique value* lebih dari 5 dilakukan *frequency encoding*.
9	Jumat, 9 Juni	*Modeling*	Sebelum melakukan modeling, saya mencari nilai korelasi antar semua kolom dengan kolom TARGET terlebih dahulu. Setelah itu, saya mengambil kolom-kolom yang memiliki nilai korelasi dengan TARGET sebesar >= 0.01 dan sebesar <= -0.01. Selanjutnya, saya membagi data menjadi X (Independen) dan Y (Dependen) dan melakukan standarisasi pada data X karena setiap kolomnya memiliki satuan yang berbeda. Setelah itu, saya melakukan *split train test pada data.* Tahap selanjutnya, yaitu melakukan *undersampling* dan *oversampling* pada data X_train dan Y_train karena ada ketidakseimbangan antar data tersebut. Setelah itu, saya melakukan modeling dengan menggunakan *logistic regression*. Setelah melakukan modeling, saya menghitung metric evaluasi untuk mengetahui performa pada model.
10	Sabtu, 10 Juni	*Modeling*	Saya melakukan modeling dengan menggunakan Random Forest dan KNN pada data yang telah di sampling dengan menggunakan *undersampling*. Setelah melakukan modeling, saya menghitung metric

			evaluasi untuk mengetahui performa pada model.
11	Minggu, 11 Juni	*Modeling*	Saya melakukan modeling dengan menggunakan Random Forest dan KNN pada data yang telah di sampling dengan menggunakan *oversampling*. Setelah melakukan modeling, saya menghitung metric evaluasi untuk mengetahui performa pada model.
12	Rabu, 14 Juni	Membuat Dashboard	Membuat pie chart client by target, score card jumlah client dengan target 1 dan 0, membuat dua tabel jumlah client dengan target 1 dan 0 berdasarkan *education type* dan *organization type*, membuat dua barchart jumlah client dengan target 1 dan 0 berdasarkan *contract type* dan gender.
13	Kamis, 15 Juni	Membuat Dashboard	Menampilkan tabel metric evaluation yang terdiri dari precision, recall, f1-score, dan support. Selain itu, saya juga menampilakan kurva ROC dan nilai ROC AUC pada model *logistic regression* tanpa *sampling,* dengan SMOTETomek, dan *over sampling* serta pada model random forest tanpa *sampling,* dengan SMOTETomek, dan *over sampling.*

Persetujuan Anggota

Laporan ini telah disetujui oleh:

- ☑ Christ Jordan Baeha
- ☑ Zulfa Nabilah Nurvitasari
- ☑ Dinny Meilinda Sari
- ☐ Sania Salsabila Agustin
- ☐ Fiqih Imanul Haq