LAPORAN ANALISIS DATA
BABAK SEMIFINAL STC LOGIKA UI 2023

LOGIKA UI 2023

Nomor Peserta
23-03-006-8

LOMBA DAN KEGIATAN MATEMATIKA UNIVERSITAS INDONESIA
2023

BUSINESS UNDERSTANDING

Yobank adalah perusahaan digital Bank yang ingin memasarkan produk terbarunya, yaitu pinjaman kredit. Langkah yang diambil Yobank sebagai perusahaan digital adalah melakukan uji coba kepada sejumlah *customer* pilihan terhadap produk terbarunya tersebut. tujuan dilakukan uji coba tersebut adalah Yobank berharap bahwa data ini dapat digunakan untuk menganalisis *payment behavior* dan membuat model *credit scoring.*

Pada uji coba tersebut data dikumpulkan dari *customer* pilihan dengan memperhatikan status pembayaran pinjaman *customer* pada bulan tersebut dan bulan-bulan sebelumnya. selanjutnya Yobank memberikan label kepada customer yang terlambat bayar selama 60 hari atau lebih sebagai "*bad" customer* dan *"good" customer* untuk *customer* lainnya.

Tim peserta 23-03-006-8 sebagai *data scientist* ditugaskan Yobank untuk memprediksi *"bad" customer* dan *"good" customer* dengan diberikan tabel informasi *customer.* Data yang diberikan oleh Yobank adalah Tabel informasi *customer* dan tabel status kredit.

Pada analysis ini, tim peserta 23-03-006-8 berencana membuat model *Decision tree* untuk melakukan model yang paling akurat dalam memprediksi *customer behavior*. Alasan dipilihnya model tersebut dijabarkan pada tabel *SWOT Analysis* dibawah ini

STRENGTH	***WEAKNESS***
Model *Decision tree* sudah terbukti keakuratan dan kekuatannya serta memerlukan sedikit usaha saat pre-proses data.	Model *Decision tree* membutuhkan waktu yang lama untuk melatih data dan dapat menjadi kompleks dengan cepat
OPPORTUNITIES	***THREAT***
Model *Decision tree* sangat intuitif dan mudah dijelaskan kepada pihak Yobank	Model *Decision tree* sensitif terhadap data sehingga bisa saja tidak tergeneralisasi terhadap masyarakat umum.

Table 1 SWOT Analysis Decision tree

Kemudian model akan dinilai berdasarkan kriteria parsimoni dan *F1-score* yang dihasilkan dari model tersebut.

DATA UNDERSTANDING

I. *Application Record* File

Application record adalah tabel yang berisi data-data dari *customer* yang terpilih oleh Yobank. kolom-kolom pada *Application record* memiliki 438557 entri kecuali kolom OCCUPATION_TYPE yang hanya memiliki 304354 entri. Penjelasan setiap kolom data *application record* dapat dilihat pada tabel 2.

Nama Variabel	Deskripsi	Tipe Data
ID	ID customer	Numerik
CODE_GENDER	Jenis kelamin customer	Kategorik
FLAG_OWN_CAR	Apakah customer memiliki mobil	Kategorik
FLAG_OWN_REALTY	Apakah customer memiliki properti	Kategorik
CNT_CHILDREN	Jumlah anak	Numerik
AMT_INCOME_TOTAL	Penghasilan per tahun	Numerik
NAME_INCOME_TYPE	Kategori penghasilan	Kategorik
NAME_EDUCATION_TYPE	Tipe Edukasi	Kategorik
NAME_FAMILY_STATUS	Status Pernikahan	Kategorik
NAME_HOUSING_TYPE	Tipe tempat tinggal atau kediaman	Kategorik
DAYS_BIRTH	Tanggal lahir	Numerik
DAYS_EMPLOYED	Tanggal awal bekerja	Numerik
FLAG_MOBIL	Apakah customer memiliki mobile phone	Kategorik
FLAG_WORK_PHONE	Apakah customer memiliki work phone	Kategorik
FLAG_PHONE	Apakah customer memiliki phone	Kategorik
FLAG_EMAIL	Apakah customer memiliki email	Kategorik
OCCUPATION_TYPE	Tipe pekerjaan	Kategorik
CNT_FAM_MEMBERS	Jumlah anggota keluarga	Numerik

Table 2 Penjelasan Application Record File

Pada *file application_record* terdapat beberapa masalah seperti banyak entri kosong pada kolom *"OCCUPATION_TYPE", Customer ID* yang duplikat, dan data-data yang tidak masuk akal nilainya. Oleh karena itu dapat disimpulkan bahwa *file application_record* masih kotor.

II. Credit Record File

Credit Record adalah tabel yang berisi data-data histori pembayaran pinjaman dari *customer* yang terpilih oleh Yobank. Semua kolom pada *Credit record* memiliki 1048575 entri. Penjelasan setiap kolom data *credit record* dapat dilihat pada tabel 3.

Nama Variabel	Deskripsi	Tipe Data
ID	ID customer	Numerik
MONTHS_BALANCE	Nomor bulan ketika data status pembayaran diambil. Dimulai dari 0 ke belakang. 0: bulan ini -1: 1 bulan yang lalu -2: 2 bulan yang lalu	Numerik
STATUS	Status pembayaran pinjaman di bulan tersebut. 0: 1-29 hari terlambat 1: 30-59 hari terlambat 2: 60-89 hari terlambat 3: 90-119 hari terlambat 4: 120-149 hari terlambat 5: lebih dari 150 hari terlambat, write-offs C: sudah dibayar di bulan tersebut X: tidak ada pinjaman di bulan tersebut	Kategorik

Table 3 Penjelasan Credit Record File

File application_record tidak memiliki entri kosong dan tidak terlihat seperti ada nilai yang tidak masuk akal ataupun kesalahan penginputan data. Oleh karena itu, *File application_record* disimpulkan bersih tetapi diperlukan sedikit persiapan untuk mengkategorikan *bad* dan *good customer*.

DATA PREPARATION

I. Membersihkan Application Record File

Sesuai yang sudah dijelaskan pada bagian data understanding, data pada Application Record File mempunyai entri kosong pada kolom "*OCCUPATION_TYPE*". Kami melakukan pengisian entri kosong pada kolom "*OCCUPATION_TYPE*" dengan dua kemungkinan yaitu "*Pensioner*" atau "*Other*". Hal ini dapat dijustifikasi dengan melihat bahwa kebanyakan baris-baris dengan entri kosong pada kolom "*OCCUPATION_TYPE*" memiliki nilai "*Pensioner*" pada kolom "*NAME_INCOME_TYPE*". maka pengisian dilakukan dengan sistem berikut:

- jika nilai "*NAME_INCOME_TYPE*" bernilai "*Pensioner*", isi "*OCUPPATION_TYPE*" dengan "*Pensioner*".
- Isi "*Other*", jika nilai "*NAME_INCOME_TYPE*" bukan bernilai "*Pensioner*".

Selain itu, kolom "*DAYS_EMPLOYED*" dan "*DAYS_BIRTH*" akan diubah satuannya menjadi satuan tahun. Kita beri label baru untuk kedua kolom tersebut: "*AGE_EMPLOYED*" dan "*AGE_BIRTH*" secara berurutan. Lebih lanjut lagi, pada kolom "*AGE_EMPLOYED*" terdapat entri yang nilainya besar, yaitu 1000.66 tahun, dibandingkan entri-entri lainnya. Oleh, karena itu akan dilakukan *encoding* nilai tersebut dengan suatu angka unik (nilainya tidak ada di dalam kolom "AGE_EMPLOYED"), katakanlah 60.

Selanjutnya, akan dilakukan pembersihan baris-baris yang duplikat. Pembersihan duplikat dilihat melalui semua kolom, kecuali kolom "ID". Dua buah baris, dikatakan duplikat, jika setiap entri pada semua kolom (kecuali kolom "ID") bernilai sama. untuk setiap baris yang saling duplikat, pada akhir proses hanya akan terdapat satu baris dengan entri tersebut yang unique.

II. Mempersiapkan Credit Score File

Data pada *Credit Score File* sudah bersih. Untuk setiap "ID"-nya, kita hanya perlu mengklasifikasikan "ID" tersebut ("good" customer atau "bad" customer) berdasarkan time series pada data *Credit Score File*. Pengklasifikasian dilakukan sesuai dengan kriteria yang sudah ditetapkan oleh Yobank, yaitu, Jika terdapat keterlambatan pembayaran lebih dari 60 hari (nilai "2", "3", "4", 5" pada kolom STATUS), Customer dengan ID tersebut adalah "bad" customer.

III. Penggabungan File dan Pembersihan Duplikat.

Pada bagian ini, dilakukan penggabungan Credit Score File dengan Application Record File (selanjutnya, akan disebut *Clean File*). Variabel terikat (CUSTOMER_TYPE) merupakan klasifikasi customer yang dilakukan pada Bagian B. Variabel penjelasnya adalah kolom-kolom pada *Application Record File* (kecuali kolom "ID"). Pengabunggan dilakukan berdasarkan kecocokan entri pada kolom "ID" antara kedua file tersebut. nilai "ID" yang tidak cocok pada kedua file, tidak dapat

dicari nilai variabel terikatnya, sehingga tidak akan dimasukkan pada *Clean File*. Selanjutnya, hilangkan juga kolom “ID”-nya.

Lalu, pada Clean File, akan dilakukan dilakukan penghapusan duplikasi. Dua baris dikatakan duplikat jika kedua baris tersebut memiliki kesamaan nilai pada setiap kolomnya.

Terakhir, kita lakukan penyesuaian tipe data pada setiap kolom.

IV. Feature Engineering and Feature Selection.

Feature Selection untuk data bertipekan numerik dilakukan dengan melihat korelasi pearson antar variabel independen tersebut. pada variabel independen yang saling berkorelasi, akan dipilih salah satu variabelnya saja. hal ini karena variabel dengan korelasi tinggi merupakan variabel yang saling bergantung (linear). menambahkan kedua variabel tersebut pada model hanyalah menambah kompleksitas dari model.

	CNT_CHILDREN	AMT_INCOME_TOTAL	CNT_FAM_MEMBERS	AGE_EMPLOYED	AGE_BIRTH
CNT_CHILDREN	1.000000	0.033290	0.890282	-0.231068	-0.322966
AMT_INCOME_TOTAL	0.033290	1.000000	0.028501	-0.156733	-0.067906
CNT_FAM_MEMBERS	0.890282	0.028501	1.000000	-0.213091	-0.275532
AGE_EMPLOYED	-0.231068	-0.156733	-0.213091	1.000000	0.665230
AGE_BIRTH	-0.322966	-0.067906	-0.275532	0.665230	1.000000

dari tabel diatas, dapat dilihat bahwa “CNT CHILDREN” dan “CNT_FAM_MEMBERS” saling berkorelasi, dan “AGE_EMPLOYED” dan “AGE_BIRTH” saling berkorelasi juga. Oleh karena itu, kita hanya akan memilih fitur “AMT_INCOME_TOTAL”, “AGE_BIRTH”, dan “CNT_FAM_MEMBERS” untuk variabel bertipekan numeriknya sebagai fitur.

Untuk variabel bertipekan kategorikal, akan dipilih melalui mekanisme uji khi-kuadrat yang dilakukan pada bagian Exploratory Data Analysis.

PREDICTION MODEL AND EVALUATION

A. Exploratory Data Analysis

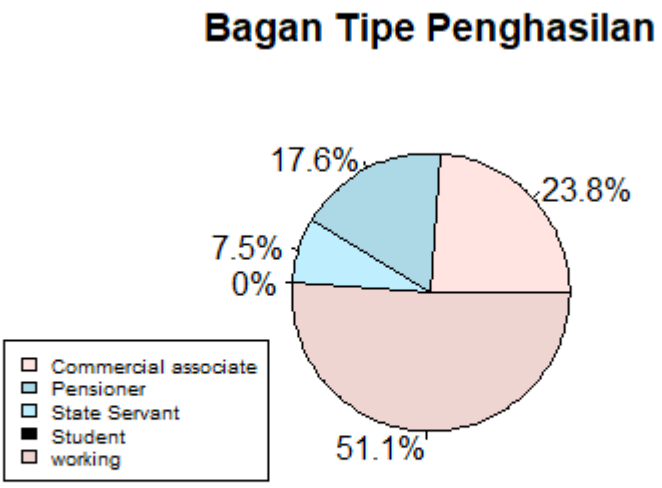
I. *Analisis Data Kategorik*

Dari 18 fitur yang terdapat pada data mayoritas dari data tersebut adalah data kategorik, lebih tepatnya sebanyak 13 fitur merupakan data kategorik. Dari 13 fitur-fitur tersebut dapat dilihat bahwa Rangkuman modus pada data kategorik dapat dilihat pada tabel dibawah ini

Fitur	Mayoritas Data	Kuantitas	Persentase
CODE_GENDER	Wanita	6510	65%
FLAG_OWN_CAR	No	6323	63%
FLAG_OWN_REALTY	Yes	6693	67%
NAME_INCOME_TYPE	Working	5109	51%
NAME_EDUCATION_TYPE	Secondary / secondary special	6949	70%
NAME_FAMILY_STATUS	Married	6728	67%
NAME_HOUSING_TYPE	House / apartment	8941	89%
FLAG_MOBIL	Yes	9997	100%
FLAG_WORK_PHONE	No	7809	78%
FLAG_PHONE	No	7113	71%
FLAG_EMAIL	No	9119	91%
OCCUPATION_TYPE	Laborers	1771	18%
CUSTOMER_TYPE	good	9555	95%

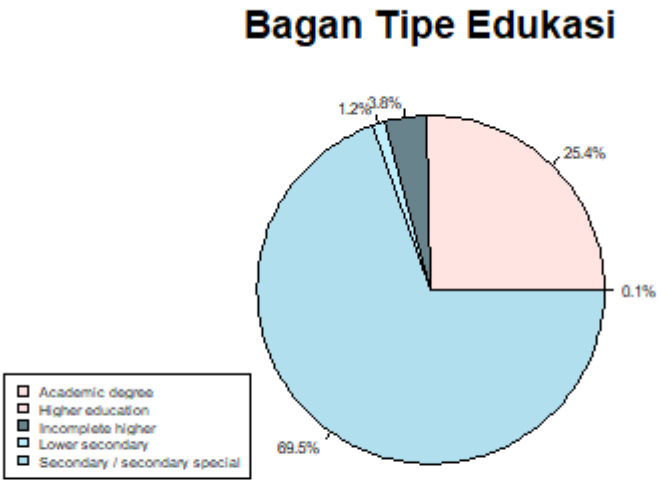
Table 4 Mayoritas Data Kategorik

Kemudian akan diselidiki lebih lanjut data-data yang memiliki variasi kategorik lebih dari 2 dengan menggunakan *pie chart* dan tabel frekuensi untuk data yang memiliki variasi sangat banyak.



Gambar 1 Bagan Tipe Penghasilan

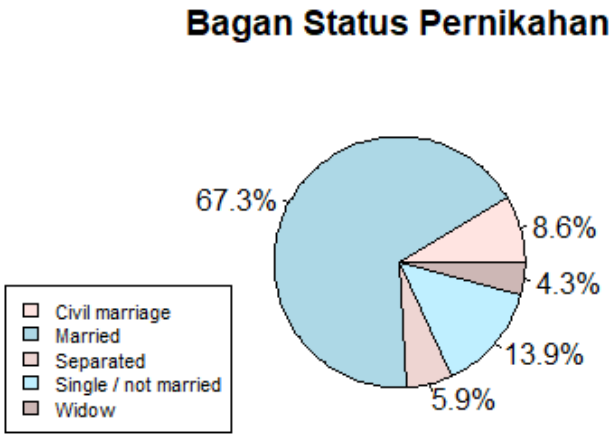
Pada bagan Tipe penghasilan dapat dilihat bahwa jumlah murid yang terpilih sangatlah sedikit dan memiliki proporsi yang cukup sesimbang pada datanya.



Gambar 2 Bagan Tipe Edukasi

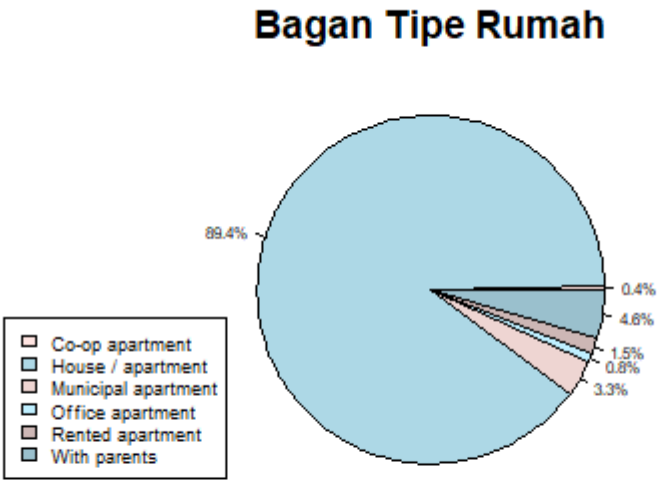
Pada bagan Tipe edukasi dapat dilihat bahwa hanya sekitar 25% orang yang mengambil pinjaman dengan edukasi tinggi atau diatasnya dan 75%

customer yang meminjam memiliki edukasi incomplete higher atau dibawahnya.



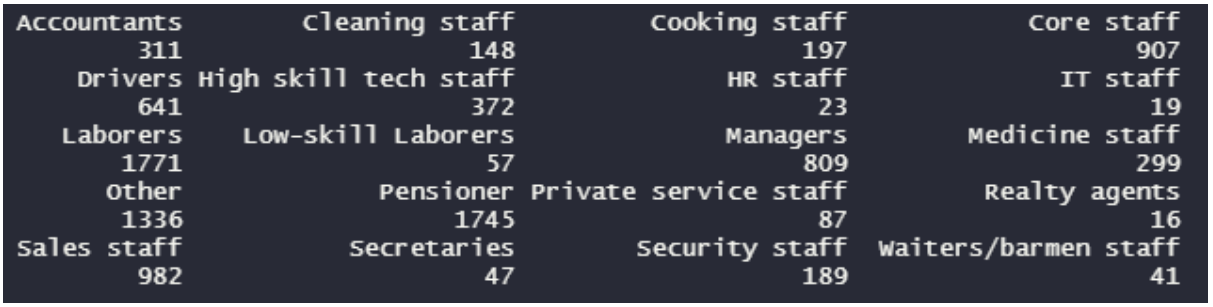
Gambar 3 Bagan Status Pernikahan

Pada bagan status pernikahan dapat dilihat sampel didominasi oleh orang yang memiliki status menikah, data juga dapat dikatakan cukup seimbang.



Gambar 4 Bagan Tipe Rumah

Pada bagan status Tipe rumah sangat didominasi oleh *customer* yang tinggal di rumah / apartment.



Gambar 5 Bagan Tipe Perkerjaan

Pada bagan tipe perkerjaan dapat dilihat bahwa pensioner, laborers dan other relatif lebih tinggi dari yang lain sedangkan perkerjaan-perkerjaan seperti HR staff, IT staff, dan waiters memiliki jumlah yang sangat sedikit.

Lalu akan dilakukan analisis *chi square* untuk melihat apakah terdapat data yang dependen terhadap tipe *customer* atau tidak

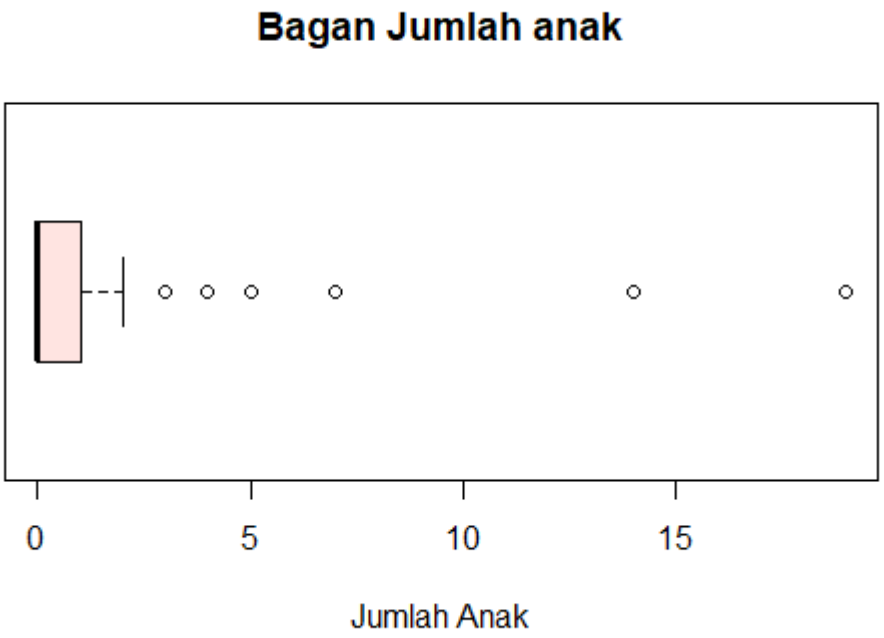
Fitur	p-value chi square	Kesimpulan
CODE_GENDER	0.05459	independent
FLAG_OWN_CAR	0.4528	independent
FLAG_OWN_REALTY	$5.687x10^{-5}$	dependent
NAME_INCOME_TYPE	0.1411	independent
NAME_EDUCATION_TYPE	0.5066	independent
NAME_FAMILY_STATUS	0.01181	dependent
NAME_HOUSING_TYPE	0.256	independent
FLAG_MOBIL	-	tidak bisa disimpulkan
FLAG_WORK_PHONE	0.3929	independent
FLAG_PHONE	0.362	independent

FLAG_EMAIL	0.7545	independent
OCCUPATION_TYPE	0.2946	independent

Table 5 Hasil Uji Chi Square

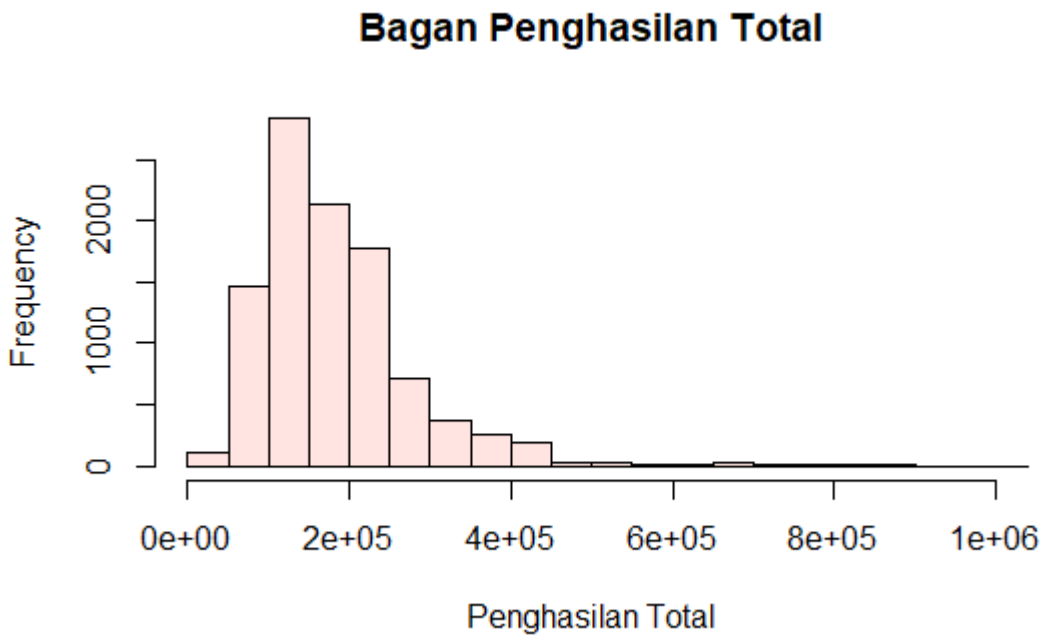
Hanya 2 fitur yang dependent terhadap customer type sedangkan majoritas fitur tidak dependent. untuk fitur FLAG_MOBIL tidak dapat ditentukan karena FLAG_MOBIL tidak memiliki variansi data sama sekali.

II. Analisis Data Numerik



Gambar 6 Bagan Jumlah anak

Pada bagan jumlah anak didapatkan bahwa bagan memiliki skewness kanan sebesar 3.56 bahkan terdapat 5 pencilan atas.



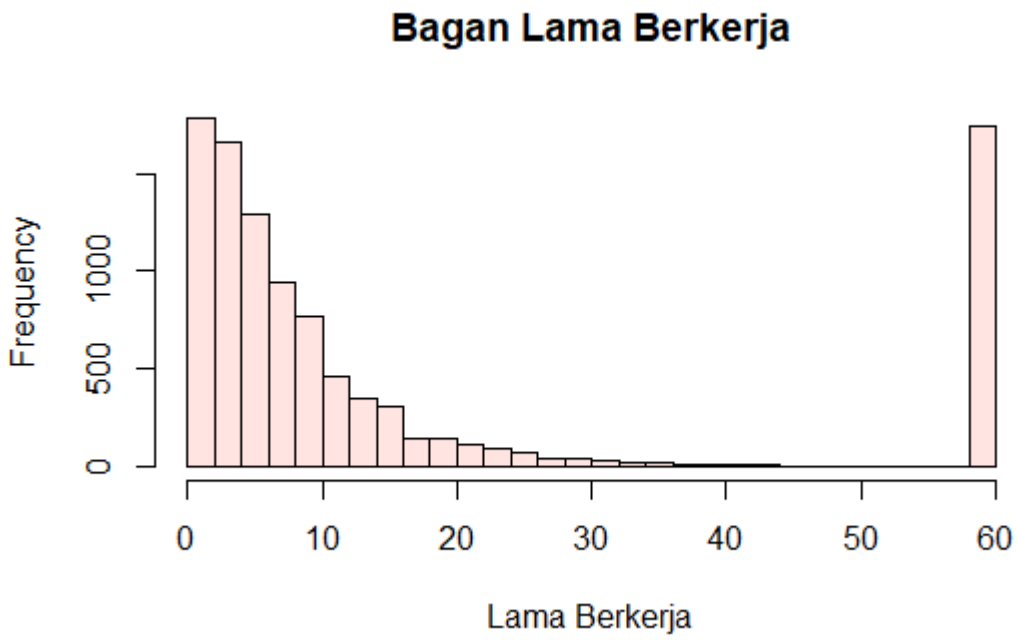
Gambar 7 Bagan Penghasilan Total

Pada bagan penghasilan total didapatkan bahwa bagan memiliki skewness kanan sebesar 2.65.

1	2	3	4	5	6	7	9	15	20
2005	5333	1684	826	123	19	4	1	1	1

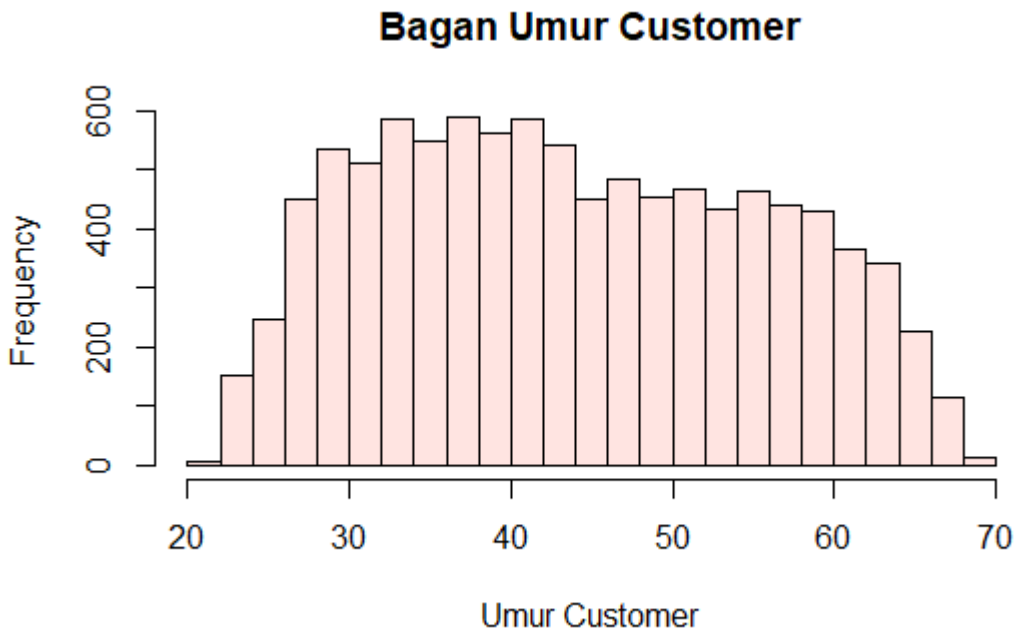
Gambar 8 Bagan Jumlah Anggota Keluarga

Pada bagan penghasilan total didapatkan bahwa bagan memiliki skewness kanan sebesar 1.83.



Gambar 9 Bagan Lama Berkerja

Pada bagan penghasilan total didapatkan bahwa bagan memiliki skewness kanan sebesar 1.47.



Gambar 10 Bagan Umur Customer

Pada bagan penghasilan total didapatkan bahwa bagan memiliki skewness kanan sebesar 0.15.

B. Modeling and Evaluation

I. Pemodelan

Untuk melakukan klasifikasi good atau bad customer, akan digunakan model berbasis Tree, yaitu Decision Tree. Pemilihan model Decision Tree dipilih karena interpretasi yang mudah. Berdasarkan Analisis sebelumnya, fitur-fitur yang akan digunakan adalah fitur "AMT_INCOME_TOTAL", "CNT_FAM_MEMBERS", "AGE_BIRTH" (variabel dengan tipe data numerik), dan "FLAG_OWN_REALTY" "NAME_FAMILY_STATUS" (variabel dengan tipe data kategorik).

Dataset akan dibagi menjadi 20% untuk keperluan testing model sedangkan 80% bagian lainnya akan digunakan untuk melatih model. Sebelum dilakukan proses fitting, Perhatikan bahwa jumlah entri antara kelas good customer dan kelas bad customer pada data tidaklah seimbang. Oleh karena itu, akan dilakukan dua metode penyeimbangan (sehingga terdapat dua model fitting), yaitu Synthetic Minority Over-sampling Technique (SMOTE) dan undersampling. Hasil akhir dari SMOTE adalah penambahan data point sintetik kelas minoritas (bad customer), sedangkan pada undersampling dilakukan re-sampling pada kelas mayoritas sehingga ukuran kelas mayoritas dibandingkan ukuran kelas minoritas adalah cukup baik(dalam hal ini akan digunakan rasio 1:3) Tujuan dilakukan penyeimbangan adalah untuk menambah bias model kearah kelas minoritas sehingga performa model untuk kelas minoritas dapat meningkat.

Selanjutnya, untuk fitur yang bertipekan kategorik non ordinal akan dilakukan label encoding (membuat dummy variabel). Tujuan dilakukannya label encoding karena, model yang berbasis tree tidak dapat melalukan fitting jika entrinya masih terdapat string type. selain itu, parameter yang diperhatikan hanyalah kedalaman tree, dengan nilai maksimal yang kita tetapkan adalah 6.

II. Peforma Model

kedua gambar berikut merangkum performa tiap modelnya.

	precision	recall	f1-score	support
bad	0.19	0.05	0.08	79
good	0.96	0.99	0.98	1921
accuracy			0.95	2000
macro avg	0.58	0.52	0.53	2000
weighted avg	0.93	0.95	0.94	2000

Gambar 11 Classification Report Model dengan SMOTE

	precision	recall	f1-score	support
0	0.52	0.27	0.35	86
1	0.75	0.90	0.82	211
accuracy			0.72	297
macro avg	0.64	0.58	0.59	297
weighted avg	0.68	0.72	0.68	297

Gambar 12 Classification Report Model UNDERSAMPLING

Dapat dilihat bahwa model dengan SMOTE adalah model dengan tingkat akurasi baik, dengan 95% akurasi secara total. Namun, performa model dengan SMOTE tersebut sangatlah buruk ketika dihadapkan pada kelas bad customer. hal ini diekspektasikan karena ketidakseimbangan jumlah data antara dua kelas.

Dapat dilihat juga bahwa model dengan undersampling mempunyai akurasi prediksi sebesar 72% secara total. namun, akurasi prediksi pada kelas bad customer jauh lebih baik dibandingkan model dengan SMOTE, yaitu sebesar 0.35 f1-score.

Tabel dibawah ini merangkum tingkat kepentingan fitur pada model Decision Tree.

	Feature Importance
AMT_INCOME_TOTAL	0.331822
CNT_FAM_MEMBERS	0.067004
AGE_BIRTH	0.432267
FLAG_OWN_REALTY_N	0.000000
FLAG_OWN_REALTY_Y	0.108541
NAME_FAMILY_STATUS_Civil marriage	0.000000
NAME_FAMILY_STATUS_Married	0.000000
NAME_FAMILY_STATUS_Separated	0.000000
NAME_FAMILY_STATUS_Single / not married	0.016998
NAME_FAMILY_STATUS_Widow	0.043368

Gambar 13 Feature Importance

CONCLUSION AND SUGGESTION

I. Conclusion

Model SMOTE dapat memprediksi dengan baik good customer namun tidak baik untuk digunakan memprediksi bad customer sedangkan model UNDERSAMPLING mempunyai performa yang cukup untuk memprediksi good customer dan juga lebih baik daripada model SMOTE dalam memprediksi bad customer. fitur-fitur yang penting dalam proses model adalah "AMT_INCOME_TOTAL", "AGE_BIRTH" dan "FLAG_OWN_REALTY".

II. Suggestion

Untuk penelitian selanjutnya, tim 23-03-006-8 menyarankan untuk mencoba mengumpulkan data dengan proprosi yang lebih seimbang, agar dapat dibangun model dengan tingkat akurasi yang baik untuk kedua buah kelas.