Vol 2 No. 4 Desember 2025 P-ISSN : 3047-3233 E-ISSN : 3047-9819, Hal 42 – 54

JURNAL RISET TEKNIK KOMPUTER
Halaman Jurnal: https://journal.smartpublisher.id/index.php/jurtikom
Halaman UTAMA Jurnal : https://journal.smartpublisher.id

DOI:https://doi.org/10.69714/xg06wn44

PREDIKSI HOME CREDIT DATA ANALISIS PADA PT. XYZ

Annisa'ul Maesaroh [a*], Budi Suyanto [b]
[a] Sistem Informasi; 124200033@student.upnyk.ac.id, Universitas Pembangunan "Veteran"
[b] Sistem Informasi; budi.suyanto@upnyk.ac.id, Universitas Pembangunan "Veteran"
[*] Penulis Korespondensi: Annisa'ul Maesaroh

ABSTRACT

Data Analytics is one of the fields of the Merdeka Campus Certified Independent Study and Internship Program (MSIB) offered by PT. Nusantara Education Zone (Zenius Education) as a form of Zenius support in building the future of the young generation to develop and improve soft skills and hard skills in the field of Data Analytics for future industrial needs. In the Independent Campus Certified Internship and Independent Study (MSIB) program, students learn basic fundamentals first such as how to think logically well and correctly, draw correct conclusions, make critical reviews, study basic arithmetic and argumentative essays as a basis for learning before entering the field data analytics. In the field of data analytics, students will learn the basics of becoming a data analyst, such as learning the basics of data analytics, the tools used, learning SQL, Python, Pandas, Data Visualization, creating dashboards, learning looker studio and data storytelling. At the end of the lesson, a final project is held in groups as an evaluation of student learning. In this practical work, the final project discusses home credit prediction with the final result being a dashboard that can be used to make decisions.

Keywords: *Data Analyst; Data Visualization; SQL; Python*

Abstrak

Data Analytics merupakan salah satu bidang yang menjadi Program Kerja dan Studi Independen Bersertifikat (MSIB) Kampus Merdeka yang ditawarkan oleh PT XYZ (PT. Zona Edukasi Nusantara (Zenius Education)) sebagai bentuk dukungan Zenius dalam membangun masa depan generasi muda untuk berkembang meningkatkan soft skill dan hard skill dalam bidang Data Analytics untuk kebutuhan industri dimasa depan. Pada program Kerja dan Studi Independen Bersertifikat Kampus Merdeka (MSIB), mahasiswa mempelajari basic fundamental terlebih dahulu sepeti cara berfikir logika yang baik dan benar, menarik kesimpulan yang benar, membuat critical review, mempelajari aritmatika dasar serta esai argumentasi sebagai dasar pembelajaran sebelum masuk ke bidang data analytics. Pada bidang data analytics, mahasiswa akan mempelajari dasar-dasar untuk menjadi seorang data analyst seperti mempelajari dasar dasar data analytics, tools yang digunakan, mempelajari SQL, Python, Pandas, Data Visualization, membuat dashboard, mempelajari looker studio dan storytelling data. Pada akhir pembelajaran diadakan proyek akhir secara berkelompok sebagai evaluasi pembelajaran mahasiswa. Pada kerja praktik ini, proyek akhir membahas tentang prediction home credit dengan hasil akhir sebuah dashboard yang bisa digunakan untuk mengambil keputusan

Kata Kunci: Data Analyst; Data Visualization; SQL; Python

1. PENDAHULUAN

Data analytics merupakan disiplin ilmu yang mencakup pengumpulan, pembersihan, pemodelan, analisis, interpretasi dan visualisasi data untuk mendapatkan wawasan untuk pengambilan keputusan. Lingkup data analytics mencakup berbagai aspek yang mencakup data terstruktur, data tidak terstruktur, serta data semi-terstruktur dari berbagai sumber, termasuk data internal, data eksternal, dan data publik. Dalam data analytics, terdapat beberapa teknik analisis yang digunakan, seperti statistik deskriptif, analisis eksploratif,

Naskah Masuk 26 September 2025; Revisi 29 September 2025; Diterima 8 Oktober 2025; Tersedia 11 Oktober 2025

analisis regresi dan korelasi, analisis klasterisasi, serta analisis klasifikasi yang membantu dalam memahami pola, tren,dan hubungan dalam data, sehingga memungkinkan pengambilan keputusan dan informasi yang lebih bernilai. Tujuan dari data analysis adalah untuk mengekstrak informasi yang tidak mudah dijawab atau diputuskan, namun ketika dipahami akan menuntun kepada kemungkinan untuk mempelajari mekanisme sistem yang menghasilkan informasi itu, sehingga dapat memprediksi kemungkinan respons dari sistem dan evolusi kedepannya (Oktavian & Budi, 2020).

Home credit merupakan penyedia layanan pembiayaan yang aman dan bertanggung jawab (responsible financing) bagi masyarakat dengan riwayat kredit yang minim, atau bahkan bagi mereka yang tidak memiliki riwayat kredit sama sekali. Pada umumnya, bank dan lembaga keuangan meninjau riwayat kredit seseorang sebagai tolak ukur untuk menilai kemampuan membayar kembali pinjamannya. Beberapa lembaga keuangan tidak menganggap pelanggan seseorang dengan riwayat minim atau bahkan tidak memiliki riwayat kredit mampu membayar kembali pinjaman mereka. Populasi yang belum terlayani ini menghadirkan peluang bisnis tersendiri. Sebuah model yang dibentuk akan memungkinkan untuk mengembangkan bisnis dengan mengidentifikasi calon peminjam kredit yang menguntungkan atau tidak. Keuntungan ini didapatkan dari klasifikasi kelayakan seseorang dapat membayar kembali kredit atau tidak dengan indikator keuangan dan non-keuangan. Hal ini diharapkan dapat meminimalkan risiko seseorang yang tidak dapat membayar kembali pinjaman, karena akan mengganggu stabilitas keuangan serta dapat menurunkan pendapatan perusahaan. Pada masa sekarang dengan data yang diolah merupakan big data maka diperlukan penerapan *machine learning* pada bisnis ini, mengingat keefektifan dalam mengidentifikasi calon peminjam yang memiliki peluang baik dalam pengembalian kredit perlu dilakukan secara cepat dan tepat.

Keberhasilan bisnis tampak ketika perusahaan mampu menjaga stabilitas keuangan dan probabilitas. Hal ini dilihat dari seberapa besar keberhasilan peminjam kredit mampu mengembalikan kredit dengan baik. Maka dari itu menjadi hal penting bagi perusahaan untuk dapat mengklasifikasikan calon peminjam yang diprediksi mampu dan tidak mampu membayar kredit.

2. METODOLOGI PENELITIAN

2.1. Analisis Permasalahan

Bisnis *home credit* kerap menghadapi tantangan yang signifikan dengan tingkat gagal bayar tinggi yang menyebabkan tingginya jumlah pelanggan yang gagal membayar kredit mereka. Ini menimbulkan masalah kritis bagi perusahaan karena secara langsung berdampak pada stabilitas keuangan dan profitabilitas. Akar penyebab permasalahan dapat dihindari dengan identifikasi yang benar dari perusahaan untuk mengetahui prediksi calon peminjam yang mampu dan tidak mampu kenal dalam membayar hutang. Permasalahan ini perlu segera diatasi untuk memitigasi risiko kerugian finansial lebih lanjut. *Dataset home credit* adalah kumpulan data yang digunakan untuk memprediksi potensi risiko kredit pelanggan berdasarkan informasi yang tersedia. Namun, dataset ini juga memiliki beberapa permasalahan yang perlu dianalisis. Pertama, terdapat ketidakseimbangan yang signifikan antara kelas pelanggan yang membayar kredit tepat waktu dan pelanggan yang mengalami keterlambatan pembayaran. Ketidakseimbangan ini dapat mengarah pada model yang cenderung memiliki kinerja yang lebih baik dalam memprediksi kelas mayoritas namun kurang efektif dalam mengidentifikasi pelanggan berisiko tinggi. Solusi seperti oversampling atau undersampling perlu dipertimbangkan untuk mengatasi masalah ini. Beberapa fitur mungkin memiliki korelasi rendah terhadap target (keterlambatan pembayaran), yang dapat mempengaruhi performa model dan memperlambat proses pelatihan. Oleh karena itu, analisis yang mendalam tentang pentingnya setiap fitur dan kemungkinan adanya fitur yang redundan perlu dilakukan untuk mengoptimalkan pemodelan dan menghindari overfitting. Diperlukan strategi pengelolaan ketidakseimbangan kelas yang tepat serta teknik pemilihan fitur yang cermat untuk membangun model yang dapat memberikan prediksi kredit yang akurat dan dapat diandalkan. Selain itu, evaluasi terus-menerus terhadap kinerja model dengan menggunakan metrik yang sesuai seperti presisi, recall, dan area di bawah kurva ROC perlu dilakukan untuk memastikan bahwa model yang dikembangkan benar-benar bermanfaat dalam membantu keputusan kredit yang lebih baik dan mengurangi risiko kredit yang tidak diinginkan.

2.2. Asumsi, Kebutuhan Data, dan Limitasi

2.2.1. Asumsi

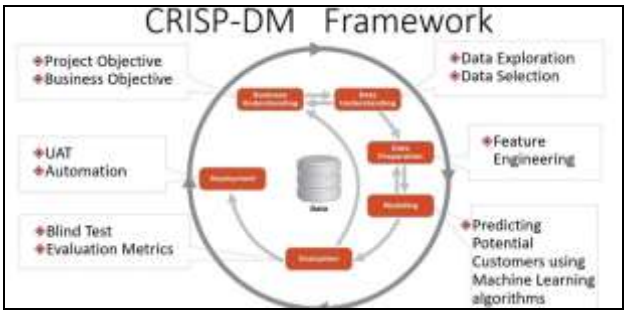
Asumsi dalam prediksi *home credit* adalah bahwa variabel "TARGET" akan dipengaruhi oleh beberapa variabel tertentu, seperti jumlah hutang, riwayat kredit, atau tingkat pendapatan pelanggan. Selain itu, variabel-variabel seperti "umur", "pekerjaan", dan "status perkawinan" , "jumlah anak", memiliki pengaruh

signifikan terhadap kelayakan kredit seseorang. Dalam *dataset home credit* ada beberapa variabel yang memiliki hubungan linear dengan variabel target, hubungan linear antara variabel nantinya akan dilakukan analisis korelasi atau visualisasi data.

2.2.2. Kebutuhan Data

Untuk melakukan prediksi orang-orang yang kemungkinan besar akan melunasi atau tidak melunasi pinjamannya menggunakan *dataset home credit*. Populasi dalam penelitian ini adalah para peminjam dengan jumlah data sebanyak 307511 baris dan 122 kolom.

Penggunaan metodologi CRISP-DM *(Cross-Industry Standard Process for Data Mining)* yang terbagi menjadi enam tahapan sebagai berikut *business understanding, data understanding, data preparation, modeling, evaluation* dan *deployment.* Dengan menerapkan CRISP-DM *Framework*, peneliti dapat mengumpulkan, membersihkan, dan mengintegrasikan data yang relevan untuk mengidentifikasi faktor risiko yang signifikan. Dengan menggunakan algoritma dan teknik analisis data yang tersedia dalam platform CRISP-DM *Framework* peneliti dapat mengembangkan model prediksi yang mampu memprediksi *home credit* dengan tingkat akurasi yang tinggi.



Gambar 1. Metode CRISP-DM Framework

Berikut langkah-langkah dalam metode *CRISP-DM Framework* :

1. Business Understanding

Tahapan pemahaman bisnis *(Business Understanding)* merupakan tahapan yang berfokus pada pemahaman tujuan kebutuhan berdasarkan penilaian bisnis (Fadillah, 2015). Serta menterjemahkan tujuan atau batasan untuk dijadikan formula dari permasalahan dari *data mining*, menyiapkan strategi awal untuk tercapainya tujuan. *Business Understanding* digunakan untuk mendapatkan gambaran yang jelas tentang apa yang sebenarnya ingin dicapai oleh bisnis atau organisasi. *Business Understanding* juga membantu mengarahkan proses analisis, memastikan bahwa solusi yang dihasilkan akan relevan dan bermanfaat bagi pengambilan keputusan bisnis.

2. Data Understanding

Tahapan pemahaman data *(Data Understanding)* adalah langkah dalam proses analisis data yang melibatkan eksplorasi dan pemahaman mendalam tentang data yang akan digunakan dalam proyek. Tujuan dari *data understanding* adalah untuk mengidentifikasi karakteristik, struktur, dan kualitas data yang ada, serta memahami bagaimana data tersebut berkaitan dengan tujuan bisnis yang ingin dicapai. Langkah ini melibatkan pengumpulan informasi tentang variabel, atribut, nilai yang mungkin ada dalam dataset, serta potensi masalah atau anomali dalam data seperti *missing values, outliers*, dan lainnya. Hal ini dilakukan untuk mulai melakukan pengumpulan data, menggunakan analisis penyelidikan data untuk menggali lebih lanjut data dan pencarian pengetahuan awal serta mengevaluasi kualitas data. *Data Understanding* membantu memastikan bahwa data yang digunakan dalam analisis memiliki kualitas yang cukup dan sesuai untuk mencapai tujuan bisnis yang telah ditetapkan.

3. Data Preparation

Tahapan persiapan (*Data Preparation*) untuk persiapan data yaitu mulai disiapkan dari data awal, mengumpulkan data yang akan digunakan secara keseluruhan pada fase berikutnya. Memilih kasus dan variable yang ingin di analisis dan yang sesuai analisa yang akan dilakukan. Pada tahap data preparation ini adalah proses pengumpulan, penggabungan, penataan, dan pengorganisasian data sehingga dapat digunakan dalam proses prediksi. Proses ini meliputi tahap-tahap seperti mengumpulkan data mentah, membersihkan data dari kesalahan atau duplikasi, dan memformat data agar sesuai dengan kebutuhan analisis. Dengan adanya tahap ini, akan menghasilkan *output* data yang sudah siap digunakan dalam

modeling. Beberapa hal yang bisa dilakukan dalam tahapan ini yaitu cleansing (pembersihan data), transformasi, *feature engineering*, pemilihan fitur, penghapusan duplikat, pengelompokan dan agregasi, dan sampling (jika dataset terlalu besar, mungkin diperlukan pengambilan sampel untuk mempercepat proses analisis).

4. Modeling
Tahapan *modeling* yaitu tahap dalam proses analisis data di mana data yang telah dipersiapkan digunakan untuk membangun model statistik atau algoritma prediktif. Tujuan dari tahap ini adalah untuk mengembangkan model yang dapat menggambarkan hubungan antara variabel *input* (fitur) dan variabel *output* (target) serta dapat digunakan untuk membuat prediksi atau mengambil keputusan berdasarkan data baru. Teknik dapat digunakan pada permasalahan data mining yang sama. Selanjutnya dapat kembali ke tahap data preparation jika perlu untuk menjadikan data sesuai kebutuhan (Navisa et al., 2021). Langkah-langkah dalam tahap modeling meliputi pemilihan model, pembagian data, pelatihan model, pengujian model, penyetelan model, validasi silang, dan evaluasi kinerja.

5. Evaluation
Melakukan evaluasi pada satu model atau lebih yang digunakan pada fase pemodelan, agar mendapat efektivitas sebelum menerapkannya di lapangan. Mengoreksi sekaligus menetapkan apakah model yang digunakan telah sesuai dengan kebutuhan kasus. Kemudian menentukan apakah ada permasalahan yang belum ditangan, selanjutnya mengambil keputusan yang terkait dengan hasil dari proses data mining. Setelah melakukan proses olah data pada dataset dengan beberapa teknik pemodelan, dilakukan proses evaluasi yaitu pengujian dari hasil proses data mining. Tujuan dari tahap ini adalah untuk mengukur seberapa baik model atau hasil analisis dapat mengatasi masalah yang ada atau memenuhi tujuan bisnis yang telah ditetapkan. Evaluasi melibatkan penggunaan metrik yang sesuai untuk mengukur kinerja model atau analisis dengan cara yang dapat diinterpretasikan secara praktis.

6. Deployment
Tahap terakhir dari CRISP-DM yang mempresentasikan hasil dari model yang telah digunakan pada proses *data mining*. Penyajian dari hasil proses, mulai dari pengetahuan yang didapat selama proses sehingga dapat dipahami oleh pengguna. Dalam tahap ini, model atau solusi yang telah berhasil dievaluasi dan divalidasi akan diimplementasikan untuk digunakan secara aktif dalam situasi dunia nyata. Tujuan utama dari tahap *deployment* adalah untuk mengintegrasikan solusi ke dalam alur kerja bisnis atau sistem yang relevan sehingga dapat memberikan manfaat langsung.

3. HASIL DAN PEMBAHASAN

Pada penelitian prediction home credit menggunakan proses *Cross Industry Standard Process for Data Mining* (CRISP-DM). CRISP-DM merupakan salah satu standar untuk proses data mining yang terdapat beberapa tahapan dari pemahaman bisnis hingga evaluasi. Dengan adanya beberapa tahapan tersebut diharapkan dapat memperoleh pemodelan yang sesuai saat proses *data mining* melalui pemahaman data yang telah disiapkan sehingga menghasilkan informasi yang dituju. Berikut tahapan CRISP-DM pada penelitian ini dijelaskan sebagai berikut:

1. Business Understanding
Home Credit adalah perusahaan yang fokus pada pemberian pinjaman kepada konsumen dengan riwayat kredit yang terbatas atau tidak ada. Tujuan utama dari penggunaan dataset ini adalah untuk mengembangkan model prediksi risiko kredit yang akurat, yang dapat membantu *Home Credit* dalam mengambil keputusan yang lebih baik terkait persetujuan atau penolakan pinjaman. Pemahaman bisnis yang lebih dalam mencakup penilaian risiko kredit yang lebih baik, peningkatan akses ke pembiayaan, pengambilan keputusan cermat. Penggunaan *dataset Home Credit* bertujuan untuk meningkatkan kualitas penilaian risiko kredit, memberikan akses keuangan kepada sektor masyarakat yang lebih luas, dan mengoptimalkan pengambilan keputusan terkait penawaran pinjaman.

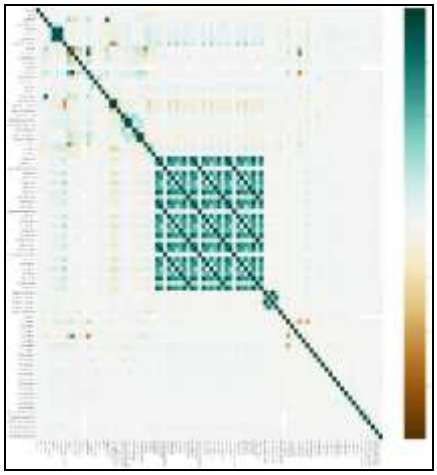
2. Data Understanding
Dataset Home Credit memiliki 6 data csv yang terdiri dari :
a. *Application_train.csv* : data pengajuan pinjaman yang memiliki kolom target dalam sampel data. Data inilah yang akan digunakan dalam pengerjaan proyek ini.
b. *Application_test.csv* : data pengajuan pinjaman yang tidak memiliki kolom target

c. *Previous_application.csv*: data pengajuan pinjaman sebelumnya yang dilakukan oleh klien yang memiliki pinjaman dalam sampel data.
d. *POS_CASH_balance.csv*: snapshot saldo bulanan dari pinjaman POS dan tunai sebelumnya yang dimiliki oleh pemohon dengan home credit.
e. *IIstallments_payments.csv*: riwayat pembayaran untuk kredit yang telah diberikan sebelumnya oleh home credit terkait dengan pinjaman dalam sampel data.
f. *Credit_card_balance.csv*: snapshot saldo bulanan dari kartu kredit sebelumnya yang dimiliki oleh pemohon dengan home credit.

Berikut describing data (atribut) yang ada pada data application_train.csv :
a. *SK_ID_CURR (int64):* nomor ID pengajuan.
b. *TARGET (int64):* 0 berarti calon debitur tidak mengalami kesulitan pembayaran dan 1 berarti calon debitur mengalami kesulitan pembayaran (gagal bayar)
c. *NAME_CONTRACT_TYPE (object*) : tipe pinjaman (cash atau revolving)
d. *CODE_GENDER (object)* : jenis kelamin calon debitur.
e. *FLAG_OWN_CAR (object)* : status kepemilikan mobil pribadi.
f. *FLAG_OWN_REALTY (object)* : status pemilikan rumah atau apartemen.
g. *CNT_CHILDREN (int64)* : jumlah anak.
h. *AMT_INCOME_TOTAL (float64)* : total pendapatan calon debitur.
i. *AMT_CREDIT (float64)* : total kredit yang dipinjamkan.
j. *AMT_ANNUITY (float64)* : besar angsuran tiap periode
k. *AMT_GOODS_PRICE (float64)* : harga barang tujuan pinjaman
l. *NAME_TYPE_SUITE (object)* : orang yang menemani peminjam (teman, suami, anak, dll)
m. *NAME_INCOME_TYPE (object)* : tipe sumber pendapatan calon debitur
n. *NAME_EDUCATION_TYPE (object)* : pendidikan terakhir/teratas calon debitur.
o. *NAME_FAMILY_STATUS (object)* : status pernikahan.
p. *NAME_HOUSING_TYPE (object)* : status tempat tinggal.
q. *REGION_POPULATION_RELATIVE (float64)* : jumlah populasi di sekitar peminjam yang dinormalisasi.
r. *DAYS_BIRTH (int64)* : umur peminjam dalam hari.
s. *DAYS_EMPLOYED (int64)* : lama peminjam memiliki pekerjaan terakhir dalam hari.
t. *DAYS_REGISTRATION (float64) :* lama peminjam menempati tempat tinggal terakhir dalam hari.
u. *DAYS_ID_PUBLISH (int64)* : lama peminjam memiliki identitas terakhir dalam hari.
v. *OWN_CAR_AGE (float64)* : usia mobil peminjam.
w. *FLAG_MOBIL (int64)*: status kepemilikan handphone.

3. Data Preparation
Pada penelitian ini menggunakan dataset yang digunakan adalah "dataset application_train.csv". Beberapa tahapan yang dilakukan dalam *data preparation* sebagai berikut:

a. Mengecek *missing value* secara keseluruhan terlebih dahulu.
Berikut adalah hubungan antar variabel dalam bentuk *heatmap*, semakin gelap warnanya, maka semakin kuat hubungannya (dapat dilihat pada gambar 2.1)



Gambar 2.1 *Heatmap*

b. Melakukan *handling missing values* pada data numerik. *Handling missing values* pada data numerik dilakukan menggunakan model *linear regression* pada seluruh data numerik untuk mengisi nilai yang kosong atau hilang. Kolom numerik akan diisi dengan regresi linear dengan kolom-kolom lain yang memiliki relasi sebagai input. Sebelum melakukan *handling numerical* data pada suatu atribut, dilakukan pengecekan total data yang hilang *(missing values)* pada setiap atribut (dapat dilihat pada gambar 2.2).

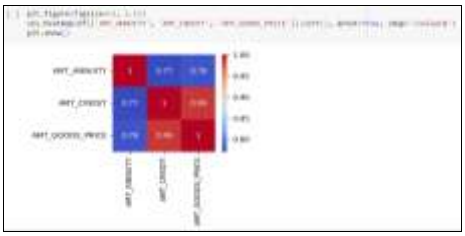
```
print('Total Missing Values')
for i in num_cols:
    print('  '+i+': '+str(df[i].isnull().sum()))
```

```
Total Missing Values
  SK_ID_CURR: 0
  TARGET: 0
  CNT_CHILDREN: 0
  AMT_INCOME_TOTAL: 0
  AMT_CREDIT: 0
  AMT_ANNUITY: 12
  AMT_GOODS_PRICE: 278
  REGION_POPULATION_RELATIVE: 0
  DAYS_BIRTH: 0
  DAYS_EMPLOYED: 0
  DAYS_REGISTRATION: 0
  DAYS_ID_PUBLISH: 0
  FLAG_MOBIL: 0
  FLAG_EMP_PHONE: 0
  FLAG_WORK_PHONE: 0
  FLAG_CONT_MOBILE: 0
  FLAG_PHONE: 0
  FLAG_EMAIL: 0
  CNT_FAM_MEMBERS: 2
  REGION_RATING_CLIENT: 0
  REGION_RATING_CLIENT_W_CITY: 0
  HOUR_APPR_PROCESS_START: 0
  REG_REGION_NOT_LIVE_REGION: 0
  TOTALAREA_MODE: 148431
  OBS_30_CNT_SOCIAL_CIRCLE: 1021
  DEF_30_CNT_SOCIAL_CIRCLE: 1021
  OBS_60_CNT_SOCIAL_CIRCLE: 1021
  DEF_60_CNT_SOCIAL_CIRCLE: 1021
  DAYS_LAST_PHONE_CHANGE: 1
  FLAG_DOCUMENT_2: 0
  FLAG_DOCUMENT_3: 0
  FLAG_DOCUMENT_4: 0
  FLAG_DOCUMENT_5: 0
  FLAG_DOCUMENT_6: 0
  FLAG_DOCUMENT_7: 0
  FLAG_DOCUMENT_8: 0
  FLAG_DOCUMENT_9: 0
  FLAG_DOCUMENT_10: 0
  FLAG_DOCUMENT_11: 0
  FLAG_DOCUMENT_12: 0
  FLAG_DOCUMENT_13: 0
  FLAG_DOCUMENT_14: 0
  FLAG_DOCUMENT_15: 0
  FLAG_DOCUMENT_16: 0
  FLAG_DOCUMENT_17: 0
  FLAG_DOCUMENT_18: 0
  FLAG_DOCUMENT_19: 0
  FLAG_DOCUMENT_20: 0
  FLAG_DOCUMENT_21: 0
  AMT_REQ_CREDIT_BUREAU_HOUR: 41519
  AMT_REQ_CREDIT_BUREAU_DAY: 41519
  AMT_REQ_CREDIT_BUREAU_WEEK: 41519
  AMT_REQ_CREDIT_BUREAU_MON: 41519
  AMT_REQ_CREDIT_BUREAU_QRT: 41519
  AMT_REQ_CREDIT_BUREAU_YEAR: 41519
```

Gambar 2.2 Pengecekan *missing values*

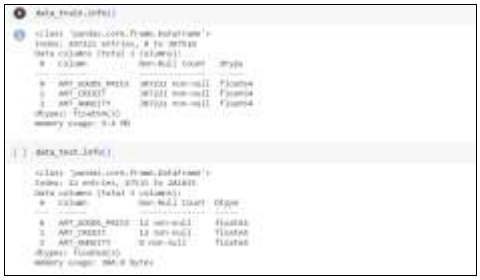
Berikut contoh salah satu *handling numerical* data pada atribut 'AMT_ANNUITY' (atribut besar angsuran tiap periode) :

1. Melakukan pengecekan matriks korelasi antara tiga atribut dalam *data frame*, yaitu 'AMT_ANNUITY', 'AMT_CREDIT' dan 'AMT_GOODS_PRICE' menggunakan sebuah *heatmap* (dapat dilihat pada gambar 2.3). Hasil yang diperoleh yaitu atribut 'AMT_ANNUITY' (besar angsuran tiap periode) memiliki korelasi yang kuat dengan atribut 'AMT_CREDIT' (total kredit yang dipinjamkan).



Gambar 2.3 Korelasi atribut AMT_ANNUITY

2. Mengecek nilai yang hilang pada ketiga atribut yaitu 'AMT_ANNUITY', 'AMT_CREDIT' dan 'AMT_GOODS_PRICE' dalam *dataset train* dan *dataset test*. Dapat dilihat pada gambar 2.4 untuk *dataset train* ketiga atribut lengkap tidak ada nilai yang hilang, sedangkan pada *dataset test* atribut 'AMT_ANNUITY' ada nilai yang hilang yang harus di *handling* agar dapat dilakukan tahap selanjutnya.

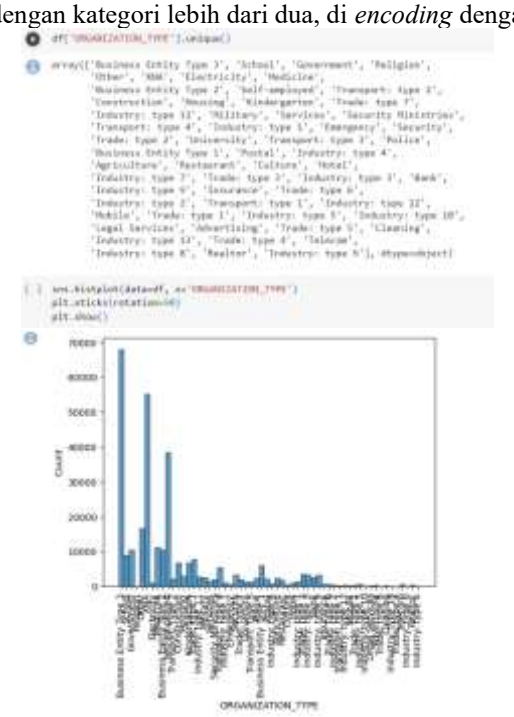


Gambar 2.4 *Data train* dan *data test numerical*

3. Melakukan regresi linear dengan kolom-kolom lain yang memiliki relasi yang akan digunakan sebagai isi dari input kolom numerik. Perhitungan dan pengisian handling missing values (nilai yang hilang) atribut 'AMT_ANNUITY' menggunakan regresi linear bisa dilihat pada gambar 2.5.

Gambar 2.5 *Regresi linear* kolom numerik

b. Melakukan *handling missing values* pada data *categorical*. Menangani *missing values* dengan median/modus kategori yang ada atau pengisian random dengan mempertahankan besar perbandingan jumlah data antar kategori.
Berikut adalah proses dalam *handling categorical* data:
1. Pengecekan *unique* pada salah satu atribut yaitu'ORGANIZATION_TYPE' dan melihat distribusinya menggunakan diagram hisplot (dapat dilihat pada gambar 2.6). Pengubahan bentuk data kategori ke numerik dengan *encoding*. Kolom dengan dua kategori (ya, tidak atau *true, false*) di *encoding* dengan *label encoding*. Sedangkan kolom dengan kategori lebih dari dua, di *encoding* dengan *metode one hot encoding*.



Gambar 2.6 Cek *unique* atribut *ORGANIZATION_TYPE*

2. Melakukan *label encoding* pada kolom atau atribut yang memiliki dua kategori (ya, tidak atau *true, false*) (dapat dilihat pada gambar 2.7). Pada proses ini mengubah nilai-nilai dalam kolom 'FLAG_OWN_REALTY', 'FLAG_OWN_CAR', dan 'EMERGENCYSTATE_MODE' menjadi kode-kode bilangan bulat lalu menyimpan nilai-nilai yang telah dikodekan pada kolom baru dan menghapus kolom asli untuk menggantinya dengan kolom yang telah dikodekan. Tujuan dari proses ini adalah untuk mengubah data kategorikal menjadi format yang dapat diterima oleh algoritma machine learning, karena banyak algoritma memerlukan input dalam bentuk angka.

Gambar 2.7 Label encoding

3. Melakukan *one hot encoding* pada kategori yang memiliki lebih dari dua nilai (dapat dilihat pada gambar 2.8). Hasilnya nilai dari atribut kategori akan berubah menjadi angka (0 dan 1).

Gambar 2.8 *One hot encoding kategorical*

4. Modeling

Pada tahap modeling akan dilakukan pemrosesan data, *feature engineering*, dan rekayasa fitur. Untuk menentukan metode analisis data yang tepat untuk memecahkan permasalahan yang ada. Percobaan dilakukan sebanyak tiga kali dengan metode yang berbeda, percobaan pertama hanya pemodelan menggunakan *models implementation and show confusion matrix*, percobaan kedua sebelum pemodelan diterapkan *sampling* menggunakan *SMOTETomek Sampling* dan percobaan ketiga menggunakan *Random Over Sampler*. Berikut model yang digunakan sebagai perbandingan dalam penelitian ini :

a. Logistic Regression

Logistic Regression adalah salah satu metode dalam analisis regresi yang digunakan untuk memodelkan hubungan antara variabel dependen yang memiliki dua kategori dengan satu atau lebih variabel independen. Metode ini menggunakan fungsi logistik untuk mengestimasi probabilitas hasil yang diinginkan berdasarkan nilai - nilai variabel independen. Mengestimasi probabilitas hasil yang diinginkan berdasarkan nilai- nilai variabel independen.

b. Decission Tree

Decission tree merupakan salah satu metode klasifikasi yang menggunakan representasi struktur pohon *(tree)* di mana setiap node merepresentasikan atribut, cabangnya merepresentasikan nilai dari atribut, dan daun merepresentasikan kelas. *Node* yang paling atas dari *decision tree* disebut sebagai *root* (Saifullah et al., 2017).

c. Random Forest

Random Forest adalah metode *ensemble learning* yang menggabungkan beberapa pohon keputusan untuk meningkatkan kinerja prediksi. Setiap pohon dalam random forest diberi bobot secara acak pada *subset* data yang berbeda. Prediksi akhir dihasilkan dengan menggabungkan hasil prediksi dari setiap pohon. *Random Forest* efektif dalam mengatasi *overfitting* dan menghasilkan model yang stabil dan akurat.

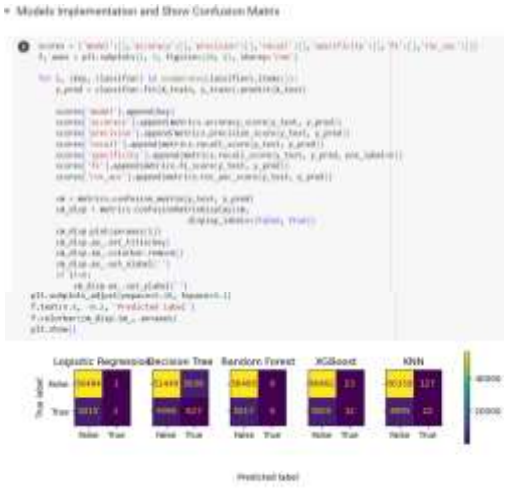
d. XGBoost

XGBoost (eXtreme Gradient Boosting) adalah metode *ensemble learning* yang menggunakan pendekatan *boosting* untuk meningkatkan kinerja prediksi. *XGBoost* menggabungkan banyak pohon keputusan lemah dan memperkuatnya secara bertahap, dengan memberikan bobot lebih pada data yang sulit diprediksi. Hal ini memungkinkan *XGBoost* untuk menghasilkan model yang akurat dan memiliki kemampuan penanganan data yang besar.

e. KNN (K-Nearest Neighbors)

KNN (K-Nearest Neighbors) adalah metode klasifikasi dan regresi *non-parametrik* yang menggunakan prinsip "tetangga terdekat" untuk memprediksi nilai atau kelas dari data yang baru berdasarkan data latihan yang ada. Dalam KNN, prediksi dilakukan dengan mencari K titik data terdekat dalam ruang fitur, dan hasil prediksi diambil berdasarkan mayoritas kelas dari tetangga terdekat tersebut. KNN juga di gunakan dalam *clustering* untuk mengelompokkan data berdasarkan kedekatannya dalam fitur. Pemodelan dilakukan sebanyak tiga kali sebagai berikut:

1. Percobaan pertama hanya menampilkan model saja tidak menggunakan *sampling* (dapat dilihat pada gambar 2.9). *Scoring ROC-AUC* didapatkan dari percobaan pertama sebagai berikut:
a. Logisitic Regression : 0.5001904703965
b. Decision Tree : 0.5179093357524666
c. Random Forest : 0.5
d. XGBoost : 0.5009923399505151
e. KNN: 0.5010683530823915



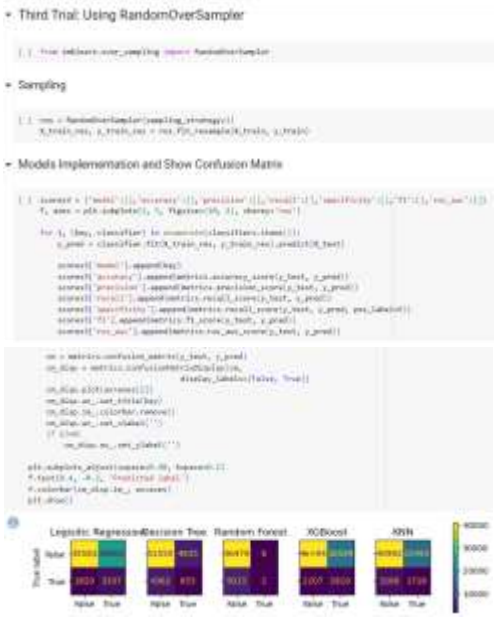
Gambar 2.9 *First trial*

2. Percobaan kedua sebelum pemodelan diterapkan *sampling* menggunakan *SMOTETomek. Scoring ROC-AUC* didapatkan dari percobaan kedua *(SMOTETomek Sampling)* sebagai berikut :
a. *Logisitic Regression* : 0.62582416385157
b. *Decision Tree* : 0.5193928533576263
c. *Random Forest* : 0.5002281681955602
d. *XGBoost* : 0.5016545603847512
e. KNN: 0.5451670763592508

Gambar 2.10 *Second trial*

3. Percobaan ketiga menggunakan *Random Over Sampler*, mendapatkan *scoring ROC-AUC* sebagai berikut :
a. Logisitic Regression : 0.6335941308447

b. Decision Tree : 0.5215938906769828
c. Random Forest : 0.5001462108586493
d. XGBoost : 0.6332035424074391
e. KNN : 0.53507186656807



Gambar 2.1 *Third trial*

3. Evaluation
Setelah dilakukan tiga kali percobaan dengan metode yang berbeda dan lima model yang dilakukan *(Logisitic Regression, Decision Tree, Random Forest, XGBoost, KNN)* didapatkan hasil sebagai berikut :

1. Scoring ROC-AUC didapatkan dari percobaan pertama sebagai berikut:
a. Logisitic Regression : 0.5001904703965
b. Decision Tree : 0.5179093357524666
c. Random Forest : 0.5
d. XGBoost : 0.5009923399505151
e. KNN: 0.5010683530823915

2. Scoring ROC-AUC di dapatkan dari percobaan kedua *(SMOTETomek Sampling)* sebagai berikut:
a. Logisitic Regression : 0.6258241638515796
b. Decision Tree : 0.5193928533576263
c. Random Forest : 0.5002281681955602
d. XGBoost : 0.5016545603847512
e. KNN: 0.5451670763592508

3. Scoring ROC-AUC didapatkan dari percobaan ketiga (Random Over Sampler) sebagai berikut:
a. Logisitic Regression : 0.6335941308447
b. Decision Tree : 0.521593890676982
c. Random Forest : 0.50014621085864
d. XGBoost : 0.6332035424074391
e. KNN : 0.53507186656807

Jika skor ROC-AUC mendekati 1, itu menunjukkan bahwa performa model memiliki kemampuan yang sangat baik dalam membedakan antara kelas positif dan kelas negatif. Ini berarti model mampu dengan sangat baik memisahkan data dari kedua kelas dan memiliki sedikit kesalahan dalam mengklasifikasikan. Untuk lebih detail terkait interpretasi dari skor ROC-AUC *(Receiver Operating Characteristic - Area Under the Curve)* sebagai berikut :

a. AUC > 0.5, Kemampuan prediksi dari model lebih baik daripada prediksi acak, semakin dekat dengan '1' maka semakin baik kinerjanya.
b. AUC = 0.5, Prediksi dari model tidak memberikan informasi yang bermakna.

c. AUC < 0.5, Prediksi dari model cenderung berbanding terbalik (salah) terhadap kelas yang sebenarnya, kinerjanya buruk.

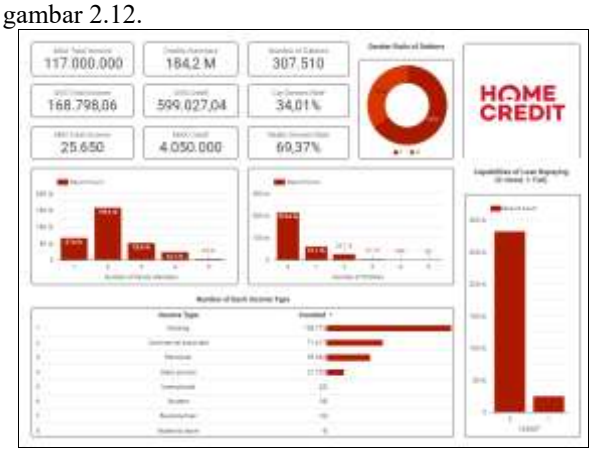
Selain itu, nilai FN (prediksi bisa membayar namun pada aktualnya tidak) dari *confusion matrix* juga menjadi pertimbangan. Sesuai dengan *bussiness case*, bank semaksimal mungkin menghindari untuk menyita aset kreditur, sehingga nilai FN dari model yang akan digunakan harus yang paling rendah (menghindari kerugian). Dari percobaan ketiga, didapatkan nilai FN terendah yaitu pada *Model Logistic Regression* : '1820' daripada keempat model lainnya pada percobaan yang sama.

Berdasarkan kecepatan pemrosesan, model yang dapat diproses dengan waktu singkat (cepat) hingga membutuhkan beberapa waktu dapat diurutkan sebagai berikut. (cepat hingga lambat) . *Logistic Regression - Decision Tree - Random Forest - XGBoost - KNN*.

Dari ketiga percobaan tersebut dengan beberapa model yang diimplementasikan, Skor *ROC-AUC* tertinggi yaitu pada angka '0.6335941308447446' dimana menggunakan model *'Logisitic Regression'* dan metode sampling *'RandomOverSampler'*.

6. Deployment

Tahap *deployment* merupakan penerapan dan implementasi hasil analisis data ke dalam pengambilan keputusan dan operasionalisasi dalam lingkungan produksi. Salah satu implementasi dalam tahap *deployment* adalah pembuatan *dashboard* untuk stakeholder. Hasil dashboard yang telah dibuat dapat dilihat pada gambar 2.12.



Gambar 2.12 *Dashboard*

Berdasarkan *dashboard* diatas dapat disimpulkan hasil sebagai berikut :

1. Total *income* (pendapatan) peminjam yang melakukan pinjaman paling besar yaitu Rp 117.000.000.
2. Paling banyak peminjam yang melakukan peminjaman berstatus sebagai pekerja dan *commercial associate*.
3. Peminjam paling banyak dilakukan oleh perempuan yaitu sekitar 65,8%, sisanya sekitar 34,2% laki – laki yang melakukan peminjaman.
4. Sekitar 34,01% peminjam memiliki mobil.
5. Jumlah peminjam yang memiliki *realty* (properti) sekitar 69,37%.
6. Sebagian besar peminjam tidak mengalami kesulitan pembayaran kredit.

4. KESIMPULAN DAN SARAN

Berdasarkan proyek *prediction home credit* maka terdapat beberapa hal yang dapat disimpulkan sebagai berikut :

a. Pada proyek kali ini menggunakan metode *CRISP-DM (Cross-Industry Standard Process for Data Mining)*.
b. Model terbaik yang dihasilkan yaitu menggunakan model *'Logisitic Regression'* dan metode sampling *'RandomOverSampler'* dengan mendapatkan skor keakuratan ROC-AUC tertinggi yaitu pada angka '0.6335941308447446'

c. Dilihat dari hasil prediksi pada atribut 'target', calon debitur lebih banyak tidak mengalami kesulitan pembayaran, dengan mayoritas belum memiliki tanggungan anak, dan paling banyak perempuan yang melakukan kredit.

Keberlanjutan

Untuk keberlanjutan hasil penelitian tentang *prediction home credit* dapat dilakukan pengembangan, penelitian lebih lanjut dengan penambahan beberapa prediksi, pengembangan pengujian untuk menentukan atribut apa saja yang sangat mempengaruhi dengan atribut peminjam bisa berhasil atau gagal dalam pembayaran kredit, dan pengembangan pembuatan *dashboard* supaya lebih mudah dimengerti dan memiliki banyak *insight* yang bisa didapatkan.

DAFTAR PUSTAKA

[1] Allaam, A. (n.d.). Prediksi Churn Konsumen Menggunakan Algoritma Random Forest dengan Fuzzy C-Means untuk Meningkatkan Produktivitas Penjualan Bisnis. Fakultas Sains dan Teknologi UIN Syarif Hidayatullah Jakarta.

[2] Fadillah, A. P. (2015). Penerapan Metode CRISP-DM untuk Prediksi Kelulusan Studi Mahasiswa Menempuh Mata Kuliah (Studi Kasus Universitas XYZ). Jurnal Teknik Informatika Dan Sistem Informasi, 1(3).

[3] Navisa, S., Hakim, L., & Nabilah, A. (2021). Komparasi Algoritma Klasifikasi Genre Musik pada Spotify Menggunakan CRISP-DM. Jurnal Sistem Cerdas, 4(2), 114–125.

[4] Oktavian, R. S., & Budi, S. (2020). Analisis Dataset Google Playstore Menggunakan Metode Exploratory Data Analysis. Jurnal STRATEGI-Jurnal Maranatha, 2(2), 636–649.

[5] Saifullah, S., Zarlis, M., Zakaria, Z., & Sembiring, R. W. (2017). Analisa Terhadap Perbandingan Algoritma Decision Tree Dengan Algoritma Random Tree Untuk Pre-Processing Data. J-SAKTI (Jurnal Sains Komputer Dan Informatika), 1(2), 180–185.