FINAL PROJECT

EKSPLORASI DAN VISUALISASI DATA

KELAS SD-A2

ANALISIS PREDIKSI RISIKO GAGAL BAYAR (DEFAULT) PADA DEBITUR KREDIT

Kelompok 8

1. Nabella Yunita Sari 164231019
2. Aqila Malfa Zahira 164231036
3. Chelsea Dheirranaya Sitinjak 164231051
4. Cuthbert Young 164231052
5. Athalia Andria Loly Aruan 164231110

PROGRAM STUDI TEKNOLOGI SAINS DATA

FAKULTAS TEKNOLOGI MAJU DAN MULTIDISIPLIN

UNIVERSITAS AIRLANGGA

SURABAYA

2024

DAFTAR ISI

DAFTAR ISI..... i

DAFTAR TABEL..... ii

DAFTAR GAMBAR..... iii

BAB I..... 1

PENDAHULUAN..... 1

1.1 Latar Belakang..... 1

1.2 Rumusan Masalah..... 1

1.3 Tujuan Penelitian..... 1

BAB II..... 3

TINJAUAN PUSTAKA..... 3

2.1 Tinjauan Pustaka Statistik..... 3

2.2.1 Data Preprocessing..... 3

2.2 Tinjauan Pustaka Non Statistik..... 5

2.2.2 Faktor-Faktor yang Mempengaruhi Risiko Gagal Bayar..... 6

BAB III..... 8

METODOLOGI..... 8

3.1 Sumber Data..... 8

3.2 Metode Data Pre-processing..... 8

3.3 Metode Analisis Data..... 10

3.4 Metode Visualisasi Data..... 10

BAB IV..... 11

ANALISIS DAN PEMBAHASAN..... 11

4.1 Deskripsi Dataset..... 11

4.2 Data Pre-processing..... 12

4.2.1 Checking and Handling Outlier..... 12

4.2.2 Checking and Handling Missing Value..... 13

4.2.3 Data transforming..... 14
4.2.4 Dimensionality reduction..... 16
4.3 Hasil Analisis..... 18
4.3.1 Analisis Statistika Deskriptif..... 18
4.3.2 Uji Asumsi..... 19
4.3.3 Regresi Logistik..... 20
4.3.4 Confusion Matrix..... 23
4.3.5 Interpretasi..... 24
BAB V..... 26
KESIMPULAN DAN SARAN..... 26
5.1 Kesimpulan..... 26
5.2 Saran..... 27
DAFTAR PUSTAKA..... 28
PEMBAGIAN TUGAS..... 29
LAMPIRAN..... 30

DAFTAR TABEL

Tabel 4.1.1 Deskripsi Dataset..... 11

Tabel 4.2.1.1 Jumlah Persentase Outlier Tiap Variabel..... 12

Tabel 4.2.2.1 Parameter Handling Missing Values Teknik PMM..... 14

Tabel 4.2.2.2 Komparasi Sebelum dan Setelah Imputasi..... 14

Tabel 4.3.1.1 Analisis Statistika Deskriptif..... 19

Tabel 4.3.4.1.1 Akurasi Hasil dengan Variabel Asli.....23

Tabel 4.3.4.2.1 Akurasi Hasil dengan Dimensi PCA..... 23

Tabel 4.3.3.3.1 Akurasi Hasil dengan Forward Selection..... 24

Tabel 4.2.4.1. Hasil Forward Selection.....30

DAFTAR GAMBAR

Gambar 4.2.1.1 Boxplot setiap variabel sebelum handling outlier.....11
Gambar 4.2.1.2 Boxplot setiap variabel setelah handling outlier.....12
Gambar 4.2.2.1 Barplot missing value setiap variabel sebelum handling missing value.....12
Gambar 4.2.2.2 Visualization and Imputation of Missing Data..... 13
Gambar 4.2.3.1 Data sebelum dan setelah perlakuan min max scaling..... 14
Gambar 4.2.3.2 Data sebelum dan setelah perlakuan robust scaling..... 14
Gambar 4.2.3.3 Data sebelum dan sesudah transform polynomial..... 15
Gambar 4.2.3.4 Data sebelum dan sesudah transform logaritma..... 15
Gambar 4.2.4.1 Heatmap korelasi antar variabel..... 16
Gambar 4.2.4.1 Visualisasi PCA..... 16
Gambar 4.3.2.2 Hasil uji Multikolinearitas (VIF)..... 18
Gambar 4.3.2.2 Hasil uji Box-Tidwell..... 19
Gambar 4.3.3.1.1 Hasil Regresi Logistik dengan Variabel Asli.....20
Gambar 4.3.3.2.1 Hasil Regresi Logistik dengan Variabel hasil PCA..... 21
Gambar 4.3.3.3.1 Hasil Regresi Logistik dengan Forward Selection..... 21

BAB I
PENDAHULUAN

1.1 Latar Belakang

Pemberian kredit merupakan salah satu aktivitas utama yang dilakukan dalam dunia keuangan. Pemberian kredit bertujuan untuk memberikan akses pendanaan kepada individu maupun pelaku usaha untuk memenuhi kebutuhan konsumsi atau investasi. Di balik tujuan tersebut, ada risiko besar yaitu gagal bayar atau kerap disebut dengan *default*. Gagal bayar dapat mempengaruhi disabilitas keuangan lembaga pemberi kredit, sehingga penting untuk memiliki kemampuan untuk memprediksi risiko gagal bayar pada debitur kredit.

Gagal bayar pada debitur kredit menjadi isu yang semakin relevan di tengah kondisi ekonomi yang tidak stabil seperti adanya inflasi, ketidakpastian pasar, atau melemahnya daya beli. Dengan memanfaatkan data dan variabel yang terkait, melibatkan analisis seperti pemrosesan data hingga model prediksi dapat dimanfaatkan untuk memprediksi risiko gagal bayar gua membantu *decision making* yang lebih strategis oleh lembaga keuangan.

1.2 Rumusan Masalah

1. Bagaimana tahapan *data preprocessing* dapat meningkatkan kualitas data untuk menganalisis risiko gagal bayar?
2. Bagaimana regresi logistik dapat digunakan untuk memprediksi risiko gagal bayar pada debitur kredit?
3. Bagaimana hasil analisis prediksi risiko gagal bayar dapat divisualisasikan dan dijadikan bahan pengambilan keputusan?

1.3 Tujuan Penelitian

1. Untuk melakukan *data preprocessing*, seperti mengidentifikasi *outlier*, menangani *missing value*, transformasi data, dan reduksi dimensi guna meningkatkan kualitas data.
2. Untuk menerapkan regresi logistik sebagai metode analisis untuk memprediksi risiko gagal bayar pada debitur kredit.
3. Untuk memvisualisasikan hasil prediksi risiko gagal bayar dan menyajikannya guna mendukung pengambilan keputusan yang leih strategis.

BAB II
TINJAUAN PUSTAKA

2.1 Tinjauan Pustaka Statistik

2.2.1 *Data Preprocessing*

Data Preprocessing merupakan suatu teknik atau proses penting dalam suatu analisis data yang bertujuan untuk membersihkan, mengubah format, dan mempersiapkan data agar lebih mudah diinterpretasikan dan lebih akurat dalam proses analisis lanjutan (Daniswara, 2023). *Data preprocessing* sendiri memiliki beberapa teknik yang bisa digunakan, namun penggunaannya kembali lagi kepada dataset yang digunakan serta tujuan analisis yang ingin dicapai. Beberapa teknik umum dalam data preprocessing meliputi:

1. *Checking & handling outlier*

   Outlier adalah data atau nilai yang secara signifikan berbeda dari sebagian besar data dalam sebuah dataset. Outlier dapat berupa nilai yang sangat besar atau sangat kecil dibandingkan dengan nilai lainnya dan sering kali dianggap sebagai pecilan. Untuk mengidentifikasi adanya outlier, terdapat 2 teknik yaitu melalui metode visualisasi dengan melakukan pengamatan titik yang ada di luar garis whisker pada boxplot dan metode statistik seperti Z-Score yang menghitung sejauh mana suatu nilai menyimpang dari rata-rata dalam satuan standar deviasi, dengan nilai di atas 3 atau di bawah -3 dianggap sebagai outlier. Selain itu, metode Interquartile Range (IQR) juga populer digunakan, di mana nilai di bawah $Q1 - 1.5IQR$ atau di atas $Q3 + 1.5IQR$ dianggap sebagai outlier. Dalam menangani outlier ini, terdapat beberapa teknik, seperti drop outlier, transformasi data, dan lain sebagainya

2. *Checking & handling missing value*

   Missing values merujuk pada ketidakhadiran data dalam suatu dataset, yang dapat disebabkan oleh berbagai faktor seperti kesalahan pengumpulan data atau tidak relevansinya informasi untuk beberapa entitas. Ada beberapa tipe missing values, yaitu MCAR (*Missing Completely at Random*), di mana data hilang secara acak tanpa adanya

pola tertentu, MNAR (*Missing Not at Random*), di mana hilangnya data bergantung pada nilai yang hilang itu sendiri, dan MAR (*Missing at Random*), di mana hilangnya data bergantung pada variabel lain yang tidak hilang. Untuk menangani missing values, terdapat beberapa teknik, seperti imputasi menggunakan mean atau median, menggunakan teknik *Predictive mean Matching* (PMM), ataupun teknik-teknik pada modelling seperti KNN, regresi, dan lain sebagainya.

3. *Transforming the data*

   Transformasi data adalah proses mengubah atau memodifikasi data agar lebih sesuai untuk analisis atau pemodelan dengan tujuan meningkatkan performa model atau memastikan asumsi-asumsi model terpenuhi. Beberapa teknik transformasi yang umum digunakan meliputi normalisasi yaitu mengubah data menjadi skala yang lebih kecil, **s**tandarisasi yaitu mengubah data menjadi distribusi dengan rata-rata 0 dan deviasi standar 1, dan lain sebagainya. Transformasi ini membantu algoritma pemodelan untuk bekerja lebih efisien dan menghasilkan prediksi yang lebih baik.

4. *Dimensionality reduction*

   Dimensionality reduction adalah teknik untuk mengurangi jumlah variabel dalam dataset yang sering digunakan ketika dataset memiliki variabel yang tidak relevan secara berlebihan. Tujuannya adalah untuk menyederhanakan model, mengurangi waktu komputasi, dan meningkatkan interpretabilitas tanpa mengurangi informasi penting. Beberapa metode populer dalam *dimensionality reduction* adalah *Principal Component Analysis* (PCA) yang mengubah data ke dalam ruang baru dengan mengidentifikasi komponen utama yang menjelaskan variasi terbesar dalam data. Dengan mengurangi jumlah dimensi, dimensionality reduction dapat membantu dalam menghindari overfitting dan mempermudah visualisasi data.

2.2.2 Regresi Logistik

Regresi logistik merupakan metode analisis statistik yang sering digunakan dalam pemodelan data biner, yaitu ketika variabel dependen memiliki dua kemungkinan nilai, seperti gagal bayar (default) atau tidak gagal bayar (non-default) pada kartu kredit. Regresi logistik ini menggunakan fungsi logit yang memodelkan probabilitas kejadian dengan rentang nilai antara 0 dan 1. Fungsi logit ini direpresentasikan melalui persamaan:

$$P(y = 1 \mid X) = \frac{1}{1 + e^{-(\beta 0 + \beta 1 x 1 + \beta 2 x 2 + \ldots + \beta k x k)}}$$

di mana P(y=1 | X) adalah probabilitas bahwa seorang nasabah akan gagal bayar (y =1) β0 adalah intercept, β1, β2,....βk adalah koefisien variabel independen $x1$, $x2$,...xk, dan e adalah bilangan eksponensial. Koefisien dalam model diestimasi menggunakan metode Maximum Likelihood Estimation (MLE), yang bertujuan untuk menemukan parameter yang memaksimalkan kemungkinan pengamatan data.

Regresi logistik juga memiliki kemampuan untuk mengukur dampak relatif dari setiap variabel independen terhadap probabilitas gagal bayar. Selain itu, model ini memungkinkan pembuatan matrik evaluasi seperti AUC-ROC, precision, recall, dan confusion matrix untuk mengukur performa prediksi. Dalam konteks keuangan, penggunaan regresi logistik membantu lembaga keuangan menilai risiko dengan lebih akurat, memitigasi potensi kerugian, serta meningkatkan strategi pengelolaan kredit berdasarkan prediksi default yang dihasilkan.

2.2 Tinjauan Pustaka Non Statistik

2.2.1 Kegiatan Kredit

Kredit adalah suatu bentuk pinjaman uang yang diberikan oleh lembaga keuangan, seperti bank, kepada individu atau perusahaan dengan syarat pengembalian dalam jangka waktu tertentu beserta bunga yang telah disepakati. Penerima kredit atau yang sering disebut sebagai debitur, dapat menggunakan dana tersebut untuk berbagai tujuan, seperti membeli barang, membiayai pendidikan, modal usaha, atau investasi. Pihak yang memberikan kredit, atau yang sering disebut kreditur, akan memastikan adanya kemampuan debitur

untuk membayar kembali melalui analisis kelayakan kredit sebelum memberikan pinjaman. Pembayaran kembali dilakukan dalam cicilan atau pembayaran penuh sesuai dengan jadwal yang telah ditetapkan. Kredit dapat dibagi menjadi berbagai jenis, seperti kredit konsumsi (untuk keperluan pribadi), kredit usaha (untuk modal usaha), dan kredit perumahan (untuk membeli rumah). Ketika debitur gagal memenuhi kewajibannya, hal tersebut dapat berakibat pada denda, bunga tambahan, atau bahkan eksekusi aset yang dijadikan jaminan (collateral). Sebagai salah satu instrumen keuangan, kredit memainkan peran penting dalam perekonomian dan berdampak pada perputaran uang serta mendukung berbagai aktivitas ekonomi.

2.2.2 Faktor-Faktor yang Mempengaruhi Risiko Gagal Bayar

Dalam menganalisis faktor-faktor yang mempengaruhi gagal bayar (default) dalam pinjaman, berbagai variabel baik yang bersifat pribadi maupun finansial dapat mempengaruhi kemampuan seorang peminjam untuk memenuhi kewajibannya. Faktor-faktor ini tidak hanya mencakup aspek angka dan data, tetapi juga berhubungan dengan stabilitas keuangan dan karakteristik sosial-ekonomi peminjam. Beberapa faktor yang mempengaruhi kemungkinan gagal bayar antara lain :

1. Umur : Umur dapat berpengaruh signifikan terhadap kemampuan seseorang untuk mengelola keuangan. Pemohon yang lebih tua cenderung memiliki pengalaman lebih banyak dalam mengatur keuangan
2. Tingkat pendidikan : Tingkat pendidikan sering dikaitkan dengan pemahaman yang lebih baik tentang manajemen keuangan. Individu dengan tingkat pendidikan yang lebih tinggi mungkin lebih mampu mengelola utang
3. Pengalaman kerja : Peminjam yang memiliki pengalaman kerja yang lebih lama cenderung memiliki penghasilan yang lebih stabil dan lebih besar
4. Lama tinggal di alamat sekarang : . Peminjam yang telah tinggal lebih lama di alamat saat ini kemungkinan memiliki tingkat kestabilan yang lebih tinggi dalam kehidupan pribadi mereka, yang berpotensi mencerminkan kemampuan untuk mengelola keuangan

5. Pendapatan tahunan : Semakin tinggi pendapatan tahunan, semakin besar kemungkinan peminjam dapat memenuhi kewajibannya.
6. Rasio utang terhadap pendapatan : Rasio utang terhadap pendapatan yang tinggi mengindikasikan bahwa peminjam mungkin kesulitan untuk memenuhi kewajiban utang
7. Jumlah utang Kartu Kredit (Credit_Debt) dan Utang Lainnya (Other_Debt) : Semakin banyak utang yang dimiliki seseorang, semakin besar kemungkinan mereka mengalami kesulitan dalam memenuhi kewajiban finansial.

BAB III
METODOLOGI

3.1 Sumber Data

Data yang digunakan pada penelitian ini merupakan data sekunder yang diunduh dari situs kaggle. Dataset berjudul *"Loans and Liability"* ini dikumpulkan oleh pihak sebelumnya untuk analisis finansial. Pada penelitian ini digunakan untuk menganalisis pengaruh berbagai metrik keuangan terhadap persetujuan peminjaman dengan memprediksi kemungkinan gagal bayar (default). Informasi yang tersedia dalam dataset mencakup variabel penting yang menggambarkan profil finansial individu, seperti pendapatan tahunan individu, jumlah utang, rasio utang terhadap pendapatan, umur, tingkat pendidikan, pengalaman kerja (lama bekerja), dan lama tinggal di alamat saat ini. Selain itu, dataset ini juga menyediakan rincian lebih spesifik terkait kewajiban finansial, termasuk jumlah utang kartu kredit (*credit debt*) dan utang lainnya (*other debt*). Variabel target dalam dataset ini adalah 'default', digunakan untuk mengidentifikasi apakah seseorang mengalami gagal bayar pada pinjamannya.

3.2 Metode Data *Pre-processing*

Data *pre-processing* adalah sebuah langkah awal yang dilakukan agar kualitas data menjadi lebih baik. Terdapat empat tahapan *pre-processing* data yang kami lakukan dalam penelitian ini. Tahapan tersebut adalah *Checking and Handling Outlier*, *Checking and Handling Missing Value*, *Data transformation*, dan *Dimensionality reduction.*

1. ***Checking and Handling Outlier***

   Langkah ini bertujuan untuk mengidentifikasi dan menangani nilai-nilai ekstrem dalam data yang berpotensi mengganggu analisis. Pertama, identifikasi outlier dilakukan menggunakan metode visualisasi yaitu boxplot dan IQR. Setelah itu, handling outlier dilakukan dengan pendekatan yang sesuai, seperti *Winsorizing*, yaitu mengganti outlier dengan nilai batas tertentu (seperti Q1 dan Q3 berdasarkan IQR) untuk mengurangi pengaruh outlier terhadap hasil analisis.

2. ***Checking and Handling Missing Value***

   Tahap ini bertujuan untuk memastikan data lengkap dan tidak memiliki celah yang dapat memengaruhi model. Langkah pertama adalah memeriksa jumlah dan pola missing value menggunakan visualisasi seperti histogram, heatmap atau bar chart. Setelah itu, handling dilakukan menggunakan teknik

imputasi yang sesuai yaitu *Predictive Mean Matching* (PMM), yang mampu mengisi nilai yang hilang dengan mempertimbangkan distribusi data. Untuk memastikan hasil imputasi yang akurat dan konsisten digunakan parameter seperti jumlah imputasi (m=5) dan iterasi maksimum (maxit=50).

3. Data transformation

Transformasi data dilakukan untuk meningkatkan performa model dengan menyesuaikan distribusi variabel sesuai kebutuhan analisis. Variabel creddebt ditransformasikan menggunakan Min-Max Scaling karena memiliki koefisien yang signifikan dan cukup besar dibandingkan variabel lain, sehingga normalisasi diperlukan untuk menjaga proporsi skala antarvariabel. Untuk variabel income, diterapkan Robust Scaling menggunakan median dan IQR untuk mengatasi rentang nilai yang besar dan sensitivitas terhadap outlier. Variabel age awalnya ditransformasikan menggunakan pangkat dua (polynomial transformation) karena distribusinya sudah cukup normal, namun hubungan dengan target kemungkinan bersifat non-linear. Setelah itu, variabel age dinormalisasi kembali dengan Min-Max Scaling karena rentang nilainya menjadi sangat besar. Sementara itu, variabel othdebt dianalisis dan menunjukkan distribusi *right skewed*, sehingga dilakukan transformasi log untuk mendekati distribusi normal dan mengurangi pengaruh outlier ekstrem. Transformasi ini bertujuan memastikan variabel berada pada skala yang lebih seimbang dan distribusi yang lebih sesuai untuk model.

4. Dimensionality reduction

Pengurangan dimensi dilakukan untuk menyederhanakan dataset dengan mengurangi jumlah variabel sambil tetap mempertahankan informasi penting. Dua pendekatan utama yang digunakan adalah extraction dan selection. Teknik extraction, seperti Principal Component Analysis (PCA), merangkum variabel berkorelasi tinggi menjadi beberapa komponen utama yang tidak berkorelasi, sehingga mengurangi dimensi tanpa kehilangan sebagian besar variansi data. Selain itu dalam penelitian ini juga menggunakan forward selection, sebuah pendekatan selection yang memilih subset variabel paling relevan secara iteratif. Proses ini dimulai dengan model kosong, kemudian menambahkan variabel satu per satu berdasarkan kontribusinya yang signifikan terhadap model prediktif.

Pemilihan variabel dilakukan dengan mengevaluasi peningkatan kualitas model menggunakan metrik tertentu, seperti AIC (Akaike Information Criterion).

3.3 Metode Analisis Data

Untuk menganalisis dan memprediksi variabel respon "default" (gagal bayar/tidak), penelitian ini menggunakan model regresi logistik. Regresi logistik dipilih karena mampu menangani variabel respon yang bersifat kategorikal biner (default: ya/tidak) dengan efektif. Model ini bekerja dengan mengestimasi probabilitas kejadian berdasarkan kombinasi linier dari variabel prediktor, menggunakan fungsi logit sebagai hubungan antara probabilitas kejadian dan variabel independen. Dengan metode ini, dapat diidentifikasi faktor-faktor signifikan yang memengaruhi kemungkinan gagal bayar sekaligus memberikan prediksi yang akurat.

3.4 Metode Visualisasi Data

Dalam penelitian ini, kami menggunakan R Shiny sebagai platform untuk memvisualisasikan data secara interaktif, dengan aplikasi web yang terdiri dari enam tab utama: Deskripsi Data, Data Frame Mentah, Pre-processing, Dataframe (Cleaned), EDA, dan Hasil Analisis. Tab Deskripsi Data memberikan gambaran umum dataset, sedangkan Data Frame Mentah menampilkan data asli sebelum proses pembersihan. Tab Pre-processing menjelaskan tahapan pembersihan data, seperti penanganan missing values, outlier, transformasi, dan reduksi dimensi, dengan tab Dataframe (Cleaned) menampilkan hasil data yang sudah siap digunakan. EDA menyediakan eksplorasi data melalui visualisasi interaktif untuk memahami pola dan hubungan antar variabel, sementara Hasil Analisis menyajikan hasil model, metrik evaluasi, dan visualisasi pendukung secara komprehensif.

BAB IV

ANALISIS DAN PEMBAHASAN

4.1 Deskripsi Dataset

Jenis data yang digunakan dalam analisis ini adalah data sekunder yang diunduh dari Kaggle. Dataset ini berisi informasi terkait profil dan status keuangan peminjam, yang dapat digunakan untuk analisis prediktif, seperti menentukan kemungkinan gagal bayar. Data ini telah dikumpulkan sebelumnya oleh pihak lain dengan tujuan penelitian finansial.

Tabel 4.1.1 Deskripsi Dataset

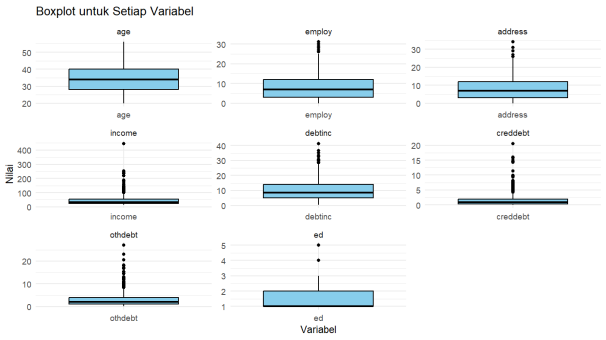
Variabel	Keterangan	Variabel	Tipe Data
age	Usia peminjam (dalam tahun)	Independen	Numerik
ed	Tingkat pendidikan peminjam (1 = SD, 2 = SMA, 3 = SARJANA, 4=PASCASARJANA)	Independen	Kategorikal (Ordinal)
employ	Lama pengalaman kerja peminjam (dalam tahun)	Independen	Numerik
address	Lama tinggal di alamat saat ini (dalam tahun)	Independen	Numerik
income	Pendapatan tahunan peminjam (dalam ribuan dolar)	Independen	Numerik
debtinc	Rasio total hutang terhadap pendapatan peminjam (dalam format persentase)	Independen	Numerik
creddebt	Total utang kartu kredit yang dimiliki peminjam (dalam ribuan dolar)	Independen	Numerik
othdebt	Total utang lainnya yang dimiliki peminjam (dalam ribuan dolar)	Independen	Numerik
default	Status gagal bayar (1=Gagal bayar, 0=Tidak gagal bayar)	Dependen	Kategorikal (Nominal)

4.2 Data Pre-processing

4.2.1 Checking and Handling Outlier

Pada penelitian ini, *checking outlier* dilakukan menggunakan boxplot. Titik titik yang berada di luar garis whisker dari setiap variabel maka dianggap sebagai outlier.

Berikut merupakan boxplot dari setiap variabel :



Gambar 4.2.1.1 *Boxplot* setiap variabel sebelum *handling outlier*

Jumlah persentase outlier tiap variabel (menggunakan metode IQR) :

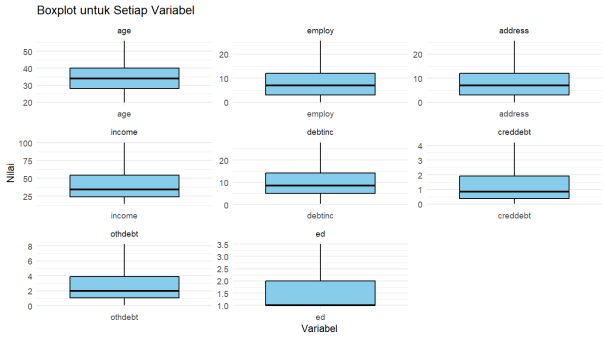
Tabel 4.2.1.1 Jumlah Persentase Outlier Tiap Variabel

Nama Variabel	Jumlah Persentase Outlier
age	2.85%
employ	1.42%
address	2%
income	11.42%
debtinc	2%
creddebt	7.85%
othdebt	6.85%
education	8.71%

Setelah dilakukan perhitungan persentase outlier per variabel, ternyata didapatkan bahwa jumlah outlier per variabel tidak terlalu signifikan, yang mana masih dibawah rentang 15%. Oleh karena itu, handling outlier dilakukan dengan cara sederhana yakni dengan metode winsorize.

Metode winsorize yakni teknik untuk menangani outlier dengan mengganti nilai ekstrim (outlier) menjadi nilai yang lebih dekat dengan data utama. Nilai outlier diganti dengan batas bawah atau batas atas yang ditentukan berdasarkan Interquartile Range (IQR).

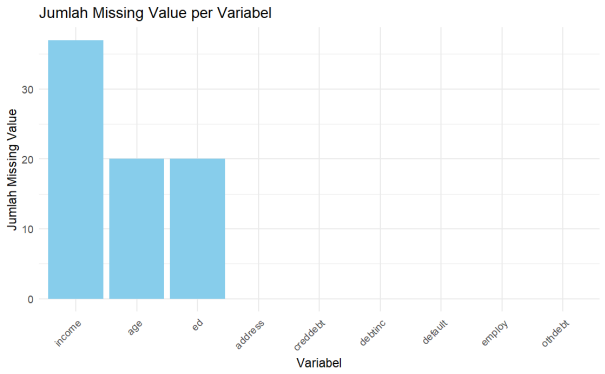
Berikut merupakan visualisasi boxplot setelah handling outlier. Terlihat pada semua variabel sudah tidak memiliki outlier lagi.



Gambar 4.2.1.2 *Boxplot* setiap variabel setelah *handling outlier*

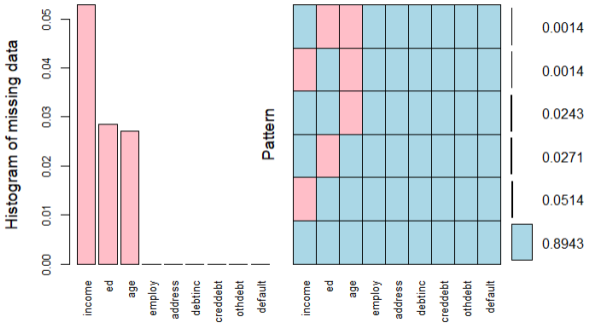
4.2.2 Checking and Handling Missing Value

Pada dataset, masih terdapat nilai yang hilang atau missing values yang tentunya perlu ditangani. Berikut merupakan visualisasi barchart terkait banyaknya missing values dari setiap variabel



Gambar 4.2.2.1 *Barplot missing value* setiap variabel sebelum *handling missing value*

Dari barchart tersebut, dapat diinterpretasikan bahwa terdapat 3 variabel yang memiliki missing values yakni *income, age* dan juga *education* dengan masing-masing jumlahnya yakni 37, 20 dan 20. Sedangkan variabel lainnya tidak memiliki missing values. Selain itu, pola missing value dapat dianalisis lebih lanjut melalui VIM (Visualization and Imputation of Missing Data) berikut.



Gambar 4.2.2.2 *Visualization and Imputation of Missing Data*

Berdasarkan visualisasi diatas pada pola data yang hilang (kanan), kotak merah muda mewakili nilai yang hilang, sementara kotak biru muda menunjukkan nilai yang tersedia. Pola ini mengindikasikan bahwa missing data pada variabel "income," "ed," dan "age" saling terkait, namun tidak ada variabel lain yang terpengaruh oleh pola tersebut. Hal ini menunjukkan bahwa pendekatan imputasi perlu mempertimbangkan hubungan antar variabel tersebut untuk menghasilkan nilai yang lebih akurat.

Untuk handling missing values, penelitian ini menggunakan teknik PMM (Predictive Mean Matching) dengan parameter sebagai berikut :

Tabel 4.2.2.1 Parameter *Handling Missing Values* Teknik PMM

Parameter	Nilai
m	5
maxit	50
seed	123

Berikut merupakan komparasi antara sebelum dan sesudah diberi perlakuan imputasi :

Tabel 4.2.2.2 Komparasi Sebelum dan Setelah Imputasi

Missing Values	*age*	*employ*	*address*	*income*	*debtinc*	*creddebt*	*othdebt*	*ed*
Sebelum Imputasi	20	0	0	37	0	0	0	20
Setelah Imputasi	0	0	0	0	0	0	0	0

4.2.3 Data transforming

1. Transformasi Min Max Scaling

Metode normalisasi min-max menggunakan nilai maksimum dan minimum untuk mendapatkan hasil normalisasi berupa nilai baru antara 0 sampai 1. Cara kerjanya setiap nilai pada sebuah fitur dikurangi dengan nilai minimum fitur tersebut, kemudian dibagi dengan rentang nilai atau nilai maksimum dikurangi nilai minimum dari fitur tersebut. Berikut contoh data sebelum dan sesudah diberi perlakuan.

creddebt_before <dbl>	creddebt <dbl>
4.201299	1.00000000
2.658720	0.63153182
1.787436	0.42341249
0.392700	0.09025876
1.358348	0.32091835
3.929600	0.93510068

Gambar 4.2.3.1 Data sebelum dan setelah perlakuan *min max scaling*

2. Transform Robust Scaling

Metode Robust Scaling menggunakan nilai median dan rentang interkuartil (IQR) untuk menskalakan data. Cara kerjanya adalah setiap nilai pada fitur dikurangi dengan nilai median fitur tersebut, kemudian dibagi dengan rentang interkuartil (IQR), yaitu selisih antara kuartil ketiga dan kuartil pertama. Berikut merupakan contoh data sebelum dan sesudah diberi perlakuan

Before_Scaling <dbl>	After_Scaling <dbl>
26	-0.30000000
24	-0.36666667
23	-0.40000000
28	-0.23333333
75	1.33333333
41	0.20000000
21	-0.46666667
21	-0.46666667
36	0.03333333
81	1.53333333

Gambar 4.2.3.2 Data sebelum dan setelah perlakuan robust scaling

3. Transform Polynomial

Metode Transformasi polynomial merupakan metode yang memiliki cara kerja mengkuadratkan data asli. Berikut merupakan contoh data sebelum dan sesudah diberi perlakuan

Before_Transformation <dbl>	After_Transformation <dbl>
33	1089
24	576
36	1296
43	1849
40	1600
42	1764
23	529
23	529
29	841
47	2209

Gambar 4.2.3.3 Data sebelum dan sesudah transform polynomial

4. Transform Logaritma

Metode Transformasi polynomial merupakan metode yang memiliki cara kerja mengambil logaritma dari setiap nilai asli. Berikut merupakan contoh data sebelum dan sesudah diberi perlakuan

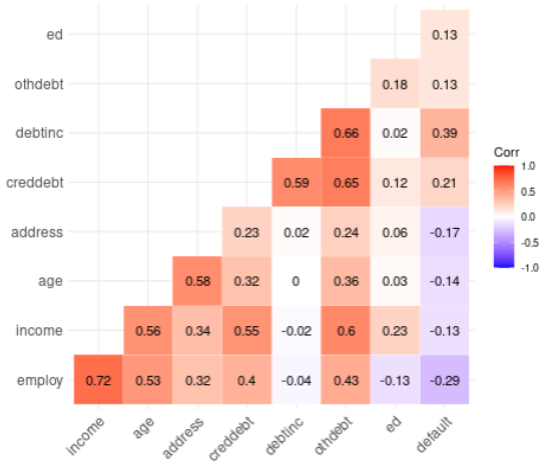
Before_Transformation <dbl>	After_Transformation <dbl>
0.775372	0.5740100
0.818400	0.5979570
2.421854	1.2301825
1.237444	0.8053341
0.540075	0.4318311
1.038530	0.7122290
0.553896	0.4407653
1.618344	0.9625421
0.376992	0.3199014
2.949210	1.3735156

Gambar 4.2.3.4 Data sebelum dan sesudah transform logaritma

4.2.4 *Dimensionality reduction*

1. PCA

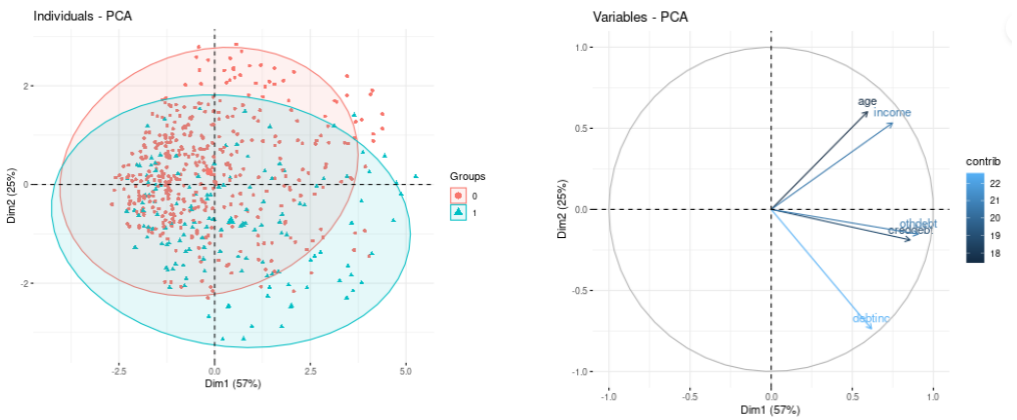
Reduksi dimensi dilakukan untuk mengurangi jumlah variabel yang tinggi, terutama ketika dataset memiliki banyak fitur. Sebelum memutuskan apakah reduksi dimensi diperlukan, langkah pertama yang perlu dilakukan adalah memeriksa tingkat korelasi antara variabel-variabel dalam data yang digunakan.



Gambar 4.2.4.1 *Heatmap* korelasi antar variabel

Heatmap korelasi antar variabel di atas menunjukkan adanya hubungan linear antar beberapa variabel, dengan nilai korelasi berkisar dari -1 hingga 1. Terdapat korelasi paling kuat yaitu antara income dan employ (0.72). Selain itu terdapat korelasi positif yang cukup kuat antara variabel *income* dan *othdebt* (0.60). Berdasarkan pola korelasi ini, terdapat redundansi informasi antar beberapa variabel, sehingga analisis PCA (Principal Component Analysis) sangat relevan untuk mengurangi dimensi data dan mengekstrak komponen utama yang mewakili informasi paling signifikan.

Berikut merupakan Visualisasi hasil PCA untuk melihat distribusi data setelah reduksi dimensi.



Gambar 4.2.4.1 Visualisasi PCA

Plot individu dan variabel dalam ruang PCA memberikan visualisasi yang intuitif untuk memahami struktur data yang kompleks. Plot individu menunjukkan distribusi data dalam ruang yang tereduksi, membantu mengidentifikasi pola, klaster, atau outlier, sementara plot variabel membantu memahami kontribusi dan hubungan variabel terhadap masing-masing komponen utama.

2. Metode Subset Selection

Pada penelitian ini, awalnya kami melakukan reduksi data menggunakan Principal Component Analysis (PCA) untuk menyederhanakan struktur data dengan mempertahankan sebagian besar varians. Namun, setelah dilakukan uji kecocokan menggunakan Kaiser-Meyer-Olkin (KMO) test, hasilnya menunjukkan nilai sebesar 0.58. Nilai ini berada di bawah ambang batas minimum 0.6, yang merupakan syarat kecocokan untuk menerapkan PCA. Hasil ini menunjukkan bahwa korelasi antar variabel dalam dataset tidak cukup kuat untuk mendukung analisis PCA. Oleh karena itu, sebagai alternatif, kami menggunakan metode subset selection untuk melakukan reduksi data. Metode yang dipilih dalam penelitian ini adalah forward selection, yaitu salah satu teknik subset selection yang dimulai dengan model kosong, lalu secara bertahap menambahkan variabel satu per satu berdasarkan kontribusi signifikan terhadap model.

Berdasarkan tabel 0. pada lampiran diketahui iterasi terakhir (iterasi ke-8) menghasilkan nilai AIC terendah, yaitu 476.18. Ini menunjukkan bahwa model yang menggunakan variabel debtinc, employ, creddebt, address, age, ed, dan othdebt adalah model terbaik di antara semua iterasi. Model ini dianggap terbaik karena nilai AIC yang lebih rendah menunjukkan bahwa model tersebut lebih efisien dalam menyeimbangkan kompleksitas dan goodness-of-fit. Dengan kata lain, model ini memberikan penyesuaian terbaik terhadap data dengan memasukkan variabel yang signifikan tanpa terlalu kompleks. Namun, perlu diperhatikan bahwa meskipun variabel othdebt dimasukkan dalam model akhir, nilai signifikansinya (p-value = 0.101451) lebih tinggi dari tingkat signifikansi konvensional 0.05, sehingga kontribusinya terhadap model mungkin tidak terlalu kuat.

4.3 Hasil Analisis

4.3.1 Analisis Statistika Deskriptif

Tabel 4.3.1.1 Analisis Statistika Deskriptif

Variab	Mean	Media	Q1	Q3	Varian	SD	Range	Sum	Min	Max

el		n			ce					
age	34.9442	34.0000	29.0000	41.0000	65.2830	8.07979	38.0000	24461.0	20.0000	58.0000
ed	1.67428	1.00000	1.00000	2.00000	0.68846	0.82973	2.50000	1172.00	1.00000	3.50000
employ	8.33857	7.00000	3.00000	12.0000	42.3923	6.51094	25.5000	5837.00	0.00000	25.5000
address	8.22428	7.00000	3.00000	12.0000	44.7843	6.69211	25.5000	5757.00	0.00000	25.5000
income	42.4032	34.0000	24.0000	55.0000	583.967	24.1654	86.2500	29682.2	14.0000	100.250
debtinc	10.1731	8.60000	5.00000	14.1250	42.9061	6.55028	27.4125	7121.17	0.40000	27.8125
creddebt	1.32356	0.85486	0.36905	1.90195	1.53685	1.23970	4.18960	926.493	0.40000	4.20129
othdebt	2.7688	1.9875	1.0441	3.9230	5.3494	2.3128	8.1958	1938.2	0.0455	8.2413
default	0.2614	0.0000	0.0000	1.0000	0.1933	0.4397	1.0000	183.00	0.0000	1.0000

Berdasarkan hasil dari analisis statistika deskriptif di atas, analisis ini dapat memberikan gambaran umum mengenai distribusi variabel dalam dataset. Secara keseluruhan, nilai mean, median, kuartil, dan rentang menunjukkan variasi data yang signifikan pada beberapa variabel. Hal tersebut mencerminkan karakteristik responden. Standar deviasi dan varians menunjukkan tingkat penyebaran data dengan beberapa variabel memiliki variasi yang lebih tinggi dibandingkan yang lain.

4.3.2 Uji Asumsi

- **Uji Multikolinearitas (VIF)**

```
     age       ed   employ  address   income  debtinc creddebt  othdebt
1.790911 1.310972 2.603896 1.517256 7.468069 6.308401 3.982662 7.528932
```

Gambar 4.3.2.2 Hasil uji Multikolinearitas (VIF)

Seluruh nilai hasil uji VIF berada di bawah 10, yang menunjukkan bahwa tidak ada multikolinearitas tinggi di antara variabel

independen dalam model sehingga seluruh variabel dapat dipertimbangkan untuk tetap digunakan dalam model regresi.

- **UJi Linearitas**

```
         MLE of lambda Score Statistic (t) Pr(>|t|)
age            8.90880              1.0368  0.30019
ed            -1.27358             -0.7414  0.45873
employ         0.63318              2.3198  0.02065 *
address        0.21537              1.0124  0.31170
income         0.72997             -0.3444  0.73065
debtinc        0.95928              0.0763  0.93923
creddebt      24.08640              0.7467  0.45550
othdebt        0.39686              1.3495  0.17763
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

iterations =  26

Score test for null hypothesis that all lambdas = 1:
F = 2.3359, df = 8 and 683, Pr(>F) = 0.01767
```

Gambar 4.3.2.2 Hasil uji Box-Tidwell

Hasil uji Box-Tidwell menunjukkan bahwa hubungan antara variabel independen dan dependen linear kecuali pada variabel *employ* dengan p-value sebesar 0.02065 yang menunjukkan hubungan tidak linear. H0 dari uji ini yaitu semua variabel memiliki hubungan linear dengan variabel *default*. HIpotesis ini ditolak pada tingkat signifikansi p=0.01767 yang menunjukkan bahwa setidaknya ada satu variabel yang tidak linear yaitu variabel *employ.*

4.3.3 Regresi Logistik

Regresi logistik digunakan untuk memodelkan hubungan antara faktor-faktor (variabel independen) dengan risiko gagal bayar (default) pada debitur kredit. Pada analisis ini, dilakukan dua pendekatan:

- Model dengan Variabel Asli: Menggunakan variabel seperti age, employ, debtinc, dll.
- Model dengan Dimensi PCA: Menggunakan kombinasi variabel asli yang telah direduksi menjadi Dim.1 hingga Dim.5 melalui PCA.
- Model dengan Forward Selection : Menggunakan beberapa variabel dengan pemilihan fitur yang paling berpengaruh/signifikan

4.3.3.1 Model Dengan Variabel Asli

```
Call:
glm(formula = default ~ age + ed + employ + address + income +
    debtinc + creddebt + othdebt, family = binomial, data = train_data)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.058314   0.777470  -3.934 8.37e-05 ***
age          0.039076   0.020209   1.934 0.053167 .
ed           0.251371   0.149211   1.685 0.092053 .
employ      -0.190831   0.034035  -5.607 2.06e-08 ***
address     -0.088847   0.024844  -3.576 0.000349 ***
income       0.001854   0.014796   0.125 0.900285
debtinc      0.136713   0.045115   3.030 0.002443 **
creddebt     0.514725   0.178972   2.876 0.004027 **
othdebt     -0.061770   0.137542  -0.449 0.653359
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 640.57  on 559  degrees of freedom
Residual deviance: 463.35  on 551  degrees of freedom
AIC: 481.35
```

Gambar 4.3.3.1.1 Hasil Regresi Logistik dengan Variabel Asli

Berdasarkan hasil regresi logistik, didapatkan persamaan model sebagai berikut:

$$\text{logit(default)} = -3.058314 + 0.039076 \cdot \text{age} + 0.251371 \cdot \text{ed} - 0.190831 \cdot \text{employ} + 0.088847 \cdot \text{address} + 0.001854 \cdot \text{income} + 0.206123 \cdot \text{debtinc} + 0.514725 \cdot \text{creddebt} - 0.061770 \cdot \text{othdebt}$$

Pada model dengan variabel asli, variabel seperti employ dan address memiliki efek negatif signifikan terhadap default, artinya semakin lama bekerja atau tinggal di alamat tertentu, semakin rendah risiko default. Sebaliknya, debtinc dan creddebt memiliki efek positif signifikan, menunjukkan bahwa rasio utang terhadap penghasilan dan utang kartu kredit yang tinggi meningkatkan risiko default.

4.3.3.2 Model Dengan Dimensi PCA

```
Call:
glm(formula = default ~ Dim.1 + Dim.2 + Dim.3 + Dim.4 + Dim.5,
    family = binomial, data = train_data)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.29230    0.11861 -10.895  < 2e-16 ***
Dim.1        0.17428    0.06377   2.733  0.00627 **
Dim.2       -0.93124    0.11605  -8.025 1.02e-15 ***
Dim.3        0.22486    0.15815   1.422  0.15507
Dim.4        0.45909    0.18138   2.531  0.01137 *
Dim.5        0.38689    0.50491   0.766  0.44353
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 640.57  on 559  degrees of freedom
Residual deviance: 527.13  on 554  degrees of freedom
AIC: 539.13
```

Gambar 4.3.3.2.1 Hasil Regresi Logistik dengan Variabel hasil PCA

Berdasarkan hasil regresi logistik, didapatkan persamaan model sebagai berikut:

$\text{logit(default)} = -1.29230 + 0.17428 \cdot \text{Dim.1} - 0.93124 \cdot \text{Dim.2} + 0.22486 \cdot \text{Dim.3} + 0.45909 \cdot \text{Dim.4} + 0.38689 \cdot \text{Dim.5}$

Model dengan dimensi PCA menggunakan kombinasi variabel asli yang dirangkum menjadi dimensi utama seperti Dim.1 dan Dim.2. Dim.1 memiliki efek positif signifikan terhadap risiko tidak gagal bayar, sedangkan Dim.2 memiliki efek negatif signifikan terhadap risiko gagal bayar. Meskipun efisien dan mampu menangkap pola utama, model ini sulit diinterpretasi secara langsung karena setiap dimensi adalah kombinasi abstrak dari variabel asli. Oleh karena itu, model ini lebih cocok digunakan untuk tujuan prediksi daripada untuk analisis faktor.

4.3.3.3 Model dengan Forward Selection

```
Call:
glm(formula = default ~ age + ed + employ + address + debtinc +
    creddebt, family = binomial, data = train_data)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.00069    0.39085  -5.119 3.07e-07 ***
age          1.57384    0.72743   2.164 0.030499 *
ed           0.20585    0.13746   1.498 0.134260
employ      -0.19950    0.02975  -6.706 2.00e-11 ***
address     -0.09084    0.02454  -3.702 0.000214 ***
debtinc      0.12168    0.02241   5.429 5.66e-08 ***
creddebt     2.20833    0.58870   3.751 0.000176 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 640.57  on 559  degrees of freedom
Residual deviance: 463.09  on 553  degrees of freedom
AIC: 477.09

Number of Fisher Scoring iterations: 5
```

Gambar 4.3.3.3.1 Hasil Regresi Logistik dengan Forward Selection

Berdasarkan hasil regresi logistik, didapatkan persamaan model sebagai berikut:

$\text{logit(default)} = -2.00069 + 1.57384 \cdot \text{age} + 0.20585 \cdot \text{ed} - 0.19950 \cdot \text{employ} - 0.09084 \cdot \text{address} + 0.12168 \cdot \text{debtinc} + 2.20833 \cdot \text{creddebt}$

Model dengan forward selection, secara sistematis memilih variabel-variabel independen yang paling berkontribusi terhadap variasi variabel dependen. Metode ini dimulai dengan memasukkan variabel yang memiliki korelasi paling kuat dengan variabel dependen, lalu secara bertahap menambahkan variabel lain yang signifikan secara statistik. Variabel yang kurang berpengaruh seperti income dan otherdebt tidak diikutkan dalam permodelan. Terlihat juga bahwa signifikasi per variabel bertambah setelah menggunakan metode ini.

4.3.4 Confusion Matrix

Confusion matrix digunakan untuk mengevaluasi performa model regresi logistik berdasarkan prediksi dan nilai aktual.

4.3.4.1 Hasil dengan Variabel Asli

Tabel 4.3.4.1.1 Akurasi Hasil dengan Variabel Asli

Predicted	Actual: 0	Actual: 1
0	94	23
1	8	15

Akurasi = $\frac{94 + 15}{140}$ = 77.86 %

Model ini lebih baik dalam mendeteksi debitur yang gagal bayar (lebih banyak True Positives) meskipun terdapat beberapa False Negatives. Model ini cocok untuk skenario di mana mendeteksi debitur yang benar-benar berisiko tinggi (default) menjadi prioritas utama.

4.3.4.2 Hasil dengan Dimensi PCA

Tabel 4.3.4.2.1 Akurasi Hasil dengan Dimensi PCA

Predicted	Actual: 0	Actual: 1
0	98	28
1	4	10

Akurasi = $\frac{98 + 10}{140}$ = 77.14 %

Model ini memiliki tingkat kesalahan False Positives yang lebih rendah dibandingkan model dengan variabel asli, sehingga lebih konservatif dalam memprediksi risiko default. Namun, model ini kurang

efektif dalam mendeteksi debitur yang benar-benar gagal bayar (lebih sedikit True Positives) dibandingkan model variabel asli.

4.3.4.3 Hasil dengan Forward Selection

Tabel 4.3.3.3.1 Akurasi Hasil dengan Forward Selection

Predicted	Actual: 0	Actual: 1
0	94	20
1	8	18

Akurasi = $\frac{94 + 18}{140}$ = 80 %

4.3.5 Interpretasi

- Variasi Data

  Analisis statistik deskriptif menunjukkan adanya variasi signifikan pada beberapa variabel utama, seperti usia (*age*), pendapatan (*income*), dan rasio hutang terhadap penghasilan (*debtinc*). Penyebaran data ini mencerminkan keberagaman karakteristik responden, memberikan dasar yang kuat untuk memahami hubungan antara variabel dan risiko gagal bayar.

- Uji Asumsi

  Tidak ditemukan multikolinearitas signifikan di antara variabel independen (VIF < 10), sehingga semua variabel dapat digunakan dalam model. Uji linearitas menunjukkan sebagian besar variabel memiliki hubungan linear dengan risiko gagal bayar, kecuali variabel *employ*, yang telah diakomodasi untuk menjaga keakuratan analisis.

- Regresi Logistik dan Pemilihan Model

  Model dengan variabel asli menunjukkan bahwa lama bekerja (*employ*) dan lama tinggal di satu alamat (*address*) berkontribusi signifikan dalam menurunkan risiko gagal bayar. Sebaliknya, rasio utang terhadap penghasilan (*debtinc*) dan utang kartu kredit (*creddebt*) secara signifikan meningkatkan risiko. Dengan akurasi sebesar 77.86%, model ini cukup andal dalam mendeteksi individu berisiko tinggi (*True*

Positives) sekaligus memberikan interpretasi langsung terhadap faktor risiko.

Pada model berbasis PCA menggunakan kombinasi variabel menjadi dimensi abstrak seperti *Dim.1* dan *Dim.2*. Model ini lebih konservatif dengan tingkat *False Positives* lebih rendah, tetapi kurang efektif dalam mendeteksi individu yang benar-benar berisiko gagal bayar. Oleh karena itu, model dengan variabel asli dipilih sebagai model akhir karena memberikan interpretasi yang lebih mendalam dan relevan untuk analisis faktor risiko gagal bayar.

Sedangkan, pada model regresi logistik dengan forward selection mengidentifikasi lima variabel signifikan: age, employ, address, debtinc, dan creddebt. Hasil menunjukkan bahwa usia (age) dan tekanan keuangan seperti debtinc dan creddebt meningkatkan risiko gagal bayar, sementara stabilitas pekerjaan (employ) dan tempat tinggal (address) mengurangi risiko. Dengan residual deviance 463.09, AIC 477.09, dan akurasi sebesar 80%, model ini memiliki kinerja yang baik dalam memprediksi risiko gagal bayar sekaligus memberikan wawasan yang jelas dan relevan terhadap faktor-faktor risiko utama.

BAB V
KESIMPULAN DAN SARAN

5.1 Kesimpulan

Berdasarkan hasil eksplorasi dan visualisasi data yang dilakukan, dapat disimpulkan sebagai berikut:

1. Tahapan Data Preprocessing

   Tahapan data preprocessing yang meliputi deteksi dan penanganan outlier, penanganan missing values, transformasi data, serta reduksi dimensi terbukti meningkatkan kualitas data untuk analisis. Teknik winsorize efektif dalam menangani outlier, sementara metode Predictive Mean Matching (PMM) berhasil mengatasi missing values tanpa mengurangi kualitas data.
2. Regresi Logistik untuk Prediksi Risiko Gagal Bayar
   - Model dengan variabel asli menunjukkan bahwa lama bekerja (employ) dan lama tinggal di satu alamat (address) memiliki efek negatif signifikan terhadap risiko gagal bayar, sedangkan rasio utang terhadap penghasilan (debtinc) dan utang kartu kredit (creddebt) memiliki efek positif signifikan.
   - Model berbasis PCA menawarkan efisiensi namun kurang memberikan interpretasi langsung dibandingkan model dengan variabel asli.
   - Model dengan Forward Selection menggunakan beberapa variabel dengan pemilihan fitur yang paling berpengaruh/signifikan. Variabel yang kurang berpengaruh seperti income dan otherdebt tidak diikutkan dalam pemodelan.
3. Visualisasi Hasil Analisis

   Visualisasi melalui confusion matrix menunjukkan bahwa model dengan Forward Selection memiliki akurasi yang lebih baik dalam mendeteksi debitur yang berisiko tinggi gagal bayar dibandingkan model dengan variabel asli dan berbasis PCA. Hal ini menjadikan model dengan Forward Selection lebih cocok untuk pengambilan keputusan strategis oleh lembaga keuangan.
4. Hasil Akhir

   Model regresi logistik berbasis Forward Selection dipilih sebagai model

akhir karena memiliki akurasi yang memadai (80%) serta interpretasi yang lebih jelas terkait faktor-faktor risiko gagal bayar.

5.2 Saran

1. Penggunaan Model

   Lembaga keuangan dapat menggunakan model regresi logistik berbasis variabel asli untuk memprediksi risiko gagal bayar debitur secara lebih akurat. Faktor seperti lama bekerja dan rasio utang terhadap penghasilan perlu menjadi perhatian utama dalam analisis kredit.
2. Perbaikan Dataset

   Untuk penelitian selanjutnya, data dengan cakupan lebih luas dan variabel tambahan seperti skor kredit atau histori pembayaran dapat digunakan untuk meningkatkan akurasi prediksi.
3. Pengembangan Metode

   Selain regresi logistik, pendekatan lain seperti machine learning berbasis pohon keputusan atau random forest dapat dieksplorasi untuk membandingkan performa model.
4. Penggunaan Visualisasi

   Visualisasi data yang lebih interaktif, seperti dashboard, dapat membantu pengambil keputusan memahami pola risiko gagal bayar secara lebih mendalam dan efisien.

Semoga kesimpulan dan saran ini dapat memberikan kontribusi untuk pengembangan sistem prediksi risiko gagal bayar di sektor keuangan.

DAFTAR PUSTAKA

Daniswara, A. A. A., & Nuryana, I. K. D. (2023). Data preprocessing pola pada penilaian mahasiswa program profesi guru. *Journal of Informatics and Computer Science (JINACS)*, *5*(1), 97. ISSN: 2686-2220.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. New York: Wiley-Interscience.

Jolliffe, I. T., & Cadima, J. (2016). Principal Component Analysis: A Review and Recent Developments. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374(2065), 20150202.

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied Logistic Regression*. Wiley.

Rose, P. S., & Hudgins, S. C. (2013). *Bank Management & Financial Services*. McGraw-Hill Education.

PEMBAGIAN TUGAS

Nama	NIM	Pembagian Tugas
Nabella Yunita Sari	164231019	- Sumber Data - Metode Data Pre-processing - Metode Analisis Data - Metode Visualisasi Data - Data Pre-processing
Aqila Malfa Zahira	164231036	- Tinjauan Pustaka Statistik - Tinjauan Pustaka Non-Statistik - Data Pre-processing
Chelsea Dheirranaya Sitinjak	164231051	- Hasil Analisis - Kesimpulan - Saran
Cuthbert Young	164231052	- Deskripsi Data - Visualisasi Data - R-Shiny
Athalia Andria Loly Aruan	164231110	- Latar Belakang - Rumusan Masalah - Tujuan Penelitian - Hasil Analisis

LAMPIRAN

Lampiran I.

Tabel Hasil Forward Selection

Tabel 4.2.4.1. Hasil *Forward Selection*

Iterasi	Variabel yang Digunakan	AIC
1	- (model awal tanpa variabel)	642.57
2	debtinc	554.12
3	debtinc, employ	501.28
4	debtinc, employ, creddebt	487.07
5	debtinc, employ, creddebt, address	479.91
6	debtinc, employ, creddebt, address, age	477.65
7	debtinc, employ, creddebt, address, age, ed	476.92
8	debtinc, employ, creddebt, address, age, ed, othdebt	476.18

Lampiran 2 *Link R-Shiny*

https://cuthbert.shinyapps.io/KELOMPOK8/

Lampiran 3 *Link Google Drive Code, Data,* dan *Output*

https://bit.ly/ProjekUASEVDKelompok8Tahun2024