ISSN 2085-4579

Analisis Data Pembayaran Kredit Nasabah Bank Menggunakan Metode Data Mining

Ira Melissa, Raymond S. Oetama
Program Studi Sistem Informasi, Universitas Multimedia Nusantara, Tangerang, Indonesia
raymond.oetama@gmail.com

Diterima 1 Mei 2013
Disetujui 31 Mei 2013

Abstrak—Data mining adalah analisis atau pengamatan terhadap kumpulan data yang besar dengan tujuan untuk menemukan hubungan tak terduga dan untuk meringkas data dengan cara yang lebih mudah dimengerti dan bermanfaat bagi pemilik data. Data mining merupakan proses inti dalam Knowledge Discovery in Database (KDD). Metode data mining digunakan untuk menganalisis data pembayaran kredit peminjam pembayaran kredit. Berdasarkan pola pembayaran kredit peminjam yang dihasilkan, dapat dilihat parameter-parameter kredit yang memiliki keterkaitan dan paling berpengaruh terhadap pembayaran angsuran kredit.

Kata kunci—data mining, outlier, multikolonieritas, Anova

I. Pendahuluan

Dalam dunia perbankan, bank bertindak sebagai kreditur, di mana bank memberikan bantuan kepada nasabah yang membutuhkan pinjaman dengan memberikan kredit pinjaman kepada nasabahnya. Bank memberikan kredit pinjaman kepada nasabah yang dianggap mampu melunasi kredit setiap bulan yang besarnya sesuai dengan perjanjian antara kedua belah pihak. Bank memberikan kemudahan bagi nasabah yang ingin meminjam uang, yaitu dengan membuat program kredit pinjaman sebagai salah satu solusi keuangan.

Namun demikian, terdapat sejumlah permasalahan yang muncul dari program kredit pinjaman. Salah satu permasalahan yang sering muncul adalah adanya nasabah yang telat membayar angsuran. Penyebab yang biasa terjadi adalah adanya nasabah yang sebenarnya telah memenuhi kualifikasi peminjaman kredit tetapi nasabah tersebut memiliki potensi yang tinggi untuk terlambat membayar kredit. Misalnya orang yang meminjam tersebut memiliki banyak tanggungan, memiliki angsuran lain dan sebagainya. Analisis terhadap data kredit bank diperlukan dengan tujuan untuk meminimalisasi resiko nasabah yang terlambat membayar kredit. Analisis yang akan dilakukan adalah analisis menggunakan metode data mining untuk melihat parameter kredit dari nasabah yang melakukan pinjaman, dalam hal ini data yang akan digunakan adalah data German Credit. Hasil dari analisis ini adalah parameter kredit. Berdasarkan parameter kredit yang dihasilkan, dapat dibuat suatu penilaian terhadap status kredit pada data German Credit yaitu terlambat bayar (overdue) atau membayar tepat waktu.

Makalah ini disusun dengan urutan topik bahasan pendahuluan yang berisi latar belakang, tinjauan pustaka, metode penelitian, pembahasan hasil penelitian, dan penutup berupa kesimpulan dan saran, disertai dengan daftar pustaka.

II. Tinjauan Pustaka

Data mining adalah analisis pengamatan kumpulan data (yang besar) untuk menemukan hubungan tak terduga dan untuk meringkas data dengan cara baru yang mudah dimengerti dan bermanfaat bagi pemilik data [1]. Hubungan dan ringkasan yang diperoleh melalui pelatihan data mining sering disebut sebagai model atau pola. Data mining digunakan untuk mengekstraksi informasi prediktif tersembunyi di dalam database [7]. Data mining mengacu pada analisis data yang kuantitasnya besar dan disimpan di komputer [5]. Dalam dunia perbankan, data mining dapat digunakan untuk mendeteksi pola kecurangan bertransaksi, memodelkan pola dan perilaku nasabah, mendeteksi kejadian yang tidak wajar, dan menemukan pengetahuan. Dalam penerapan data mining pada data berskala besar diperlukan metodologi sistematis untuk menganalisis dan mempersiapkan data. Metodologi sistematis juga dibutuhkan untuk melakukan interpretasi sehingga dapat menghasilkan keputusan yang bermanfaat.

Data mining merupakan bagian dari Knowledge Discovery in Databases (KDD). Knowledge Discovery in Databases atau biasa disingkat dengan KDD berisi serangkaian proses-proses yang harus dilakukan sebelum dan sesudah menganalisis dengan metode data mining. Langkah-langkah yang dilakukan dalam KDD [3] meliputi Data cleaning, Data integration, Data selection, Data transformation, Pembentukan model, Pattern evaluation, dan Knowledge presentation.

Data Cleaning

Data cleaning adalah perbaikan terhadap data-data yang rusak, hilang atau salah (error). Pada tahapan data cleaning, hal yang harus dilakukan adalah menganalisis data untuk mendeteksi adanya data outlier. Outlier adalah kasus atau data yang memiliki karakteristik unik yang terlihat sangat berbeda jauh dari observasi-observasi lainnya dan muncul dalam bentuk nilai ekstrim baik untuk sebuah variabel tunggal atau variabel kombinasi [2].

Ada empat penyebab timbulnya data outlier, yaitu kesalahan dalam memasukan data, kegagalan dalam spesifikasi missing value ke dalam program komputer, pengambilan sample yang salah, dan sampel diambil dengan benar tetapi memiliki nilai ekstrim dan tidak terdistribusi secara normal [2].

Adapun outliers dapat dievaluasi dengan dua cara, yaitu analisis terhadap univariate outliers dan analisis terhadap multivariate outliers [3]. Multivariate outliers perlu dilakukan walaupun data yang dianalisis menunjukkan tidak ada outliers. Jarak Mahalanobis (Mahalanobis Distance) untuk tiap-tiap observasi dapat dihitung dan menunjukkan jarak sebuah observasi dari rata-rata semua variabel dalam sebuah ruang multidimensional [3].

Untuk mengetahui apakah hasil transformasi data berdistribusi normal, maka perlu dilakukan pengujian. Teknik pengujian yang digunakan adalah uji One-Sample Kolmogorov-Smimov Test. One-Sample Kolmogorov-Smimov Test adalah tes non parametik untuk probabilitas distribusi satu dimensi dan dapat digunakan untuk membandingkan sampel dengan distribusi probabilitas referensi.

Pembentukan Model Data Mining

Dalam pembentukan model data mining, dibutuhkan suatu algoritma yang sesuai dengan penelitian yang dilakukan. Algoritma-algoritma dalam data mining dapat dilihat dari dua sudut pandang yang berbeda [7]. Menurut sudut pandang statistik dan riset operasi, algoritma data mining adalah analisis cluster, regression, dan analisis diskriminan. Sedangkan, menurut sudut pandang kecerdasan buatan atau artificial intelligence, algoritma data mining adalah neural networks, rule induction dan algoritma genetika. Dari algoritma-algoritma yang telah disebutkan, algoritma-algoritma yang umum digunakan adalah algoritma regression dan decision tree.

Regression atau regresi merupakan metode pokok untuk menganalisis statistik hubungan karakter antara satu variabel dependen dengan satu atau lebih variabel independen. Model regresi dapat digunakan untuk berbagai tujuan misalnya untuk prediksi dan penjelasan. Model regresi terdiri dari linear regression dan logistic regression. Algoritma linear regression digunakan pada data yang non linear, sedangkan algoritma logistic regression digunakan pada data yang linier.

Decision tree atau pohon keputusan pada data mining merujuk pada aturan stuktur pohon. Algoritma ini secara otomatis akan menentukan variabel mana yang paling penting berdasarkan kemampuan mengurutkan data ke dalam kategori keluaran yang benar.

Pattern Evaluation

Pattern evaluation adalah analisis pola-pola untuk menemukan pola terbaik yang disesuaikan atau diperbandingkan dengan tujuan penelitian. Hasil analisis ini pada umumnya terdiri dari banyak pola. Namun, tidak semua pola itu merupakan pola yang diperlukan. Oleh karena itu, diperlukan suatu pengembangan teknik yang dapat menilai ketertarikan pola yang ditemukan. Proses ini diperlukan untuk memilih pola terbaik sebagai solusi atas jawaban dari tujuan penelitian.

Knowledge Presentation

Knowledge presentation adalah langkah-langkah yang dilakukan untuk mengubah data menjadi pengetahuan, mengubah pengetahuan eksplisit menjadi pengetahuan implisit, dan mengolah data mentah menjadi informasi yang dibutuhkan. Pada tahap knowledge presentation, konsep dasar manajemen pengetahuan diterapkan. Hasil dari knowledge presentation harus disajikan dalam bahasa yang mudah dimengerti, menggunakan representasi visual, atau bentuk lain yang lazim bagi pengguna. Hal ini dimaksudkan agar pengetahuan yang dibagikan mudah dipahami oleh orang lain. Pengetahuan dapat dibagikan ke orang lain melalui berbagai cara, yaitu melalui buku, pelatihan, seminar, diskusi, dan lain-lain.

Analysis of Variance (Anova)

Analysis of Variance (Anova) adalah suatu metode untuk menguji hubungan antara satu variabel dependen dengan satu atau lebih variabel independen. Terdapat beberapa jenis Anova, yaitu One Way Anova, Two Ways Anova dan Three Ways Anova. One Way Anova adalah hubungan antara satu variabel dependen dengan satu variabel independen, sedangkan hubungan antara satu variabel dependen metrik dengan dua variabel independen kategorikal disebut Two Ways Anova. Hubungan antara satu variabel dependen metrik dengan tiga variabel independen kategorikal disebut Three Ways Anova. Fungsi dari Anova adalah untuk mengetahui pengaruh utama dan pengaruh interaksi dari variabel independen kategorikal terhadap variabel dependen metrik. Pengaruh utama adalah pengaruh langsung variabel independen terhadap variabel

dependen, sedangkan pengaruh interaksi dapat diartikan pengaruh bersama dua atau lebih variabel independen terhadap variabel dependen.

Algoritma Logistic Regression

Logistic regression adalah sebuah algoritma untuk menguji probabilitas terjadinya variabel dependen dapat diprediksi dengan variabel independennya. Algoritma logistic regression digunakan karena analisis dengan logistic regression tidak memerlukan asumsi normalitas data pada variabel bebasnya. Hal ini juga berarti bahwa algoritma logistic regression pada umumnya dipakai jika asumsi multivariate normal distribution tidak dipenuhi.

Pada analisis kredit nasabah Bank, terdapat dua kasus, yaitu pembayaran kredit yang macet atau bad (B) dan pembayaran kredit yang lancer atau good (G). Maka, probabilitas pembayaran kredit macet dan baik dilambangkan dengan P(M|B). Probabilitas sering dinyatakan dalam istilah odds. Odds dan probabilitas berisikan informasi yang sama dalam bentuk yang berbeda. Oleh karena itu, odds dapat diubah menjadi probabilitas atau sebaliknya. Dalam kasus pembayaran kredit nasabah bank, hubungan antara odds dan probabilitas dapat dihitung dengan :

$$P(B|G) = \frac{Odds\ (B|G)}{1 + Odds(B|G)}$$

Setelah melakukan perhitungan odds, selanjutnya dapat dihitung nilai log naturalnya (LN). Persamaan logistic regression untuk k variabel bebas dinyatakan sebagai berikut :

$$Ln[odds(S|X1, X2, \dots, X3)] = b0 + b1X1 + b2X2 + \cdots + bkXk$$

Metode Oversampling dan Undersampling

Metode oversampling dan undersampling adalah metode yang digunakan untuk mengatasi masalah imbalanced data. Imbalanced data atau data yang tak berimbang merupakan suatu kondisi dimana pada sebuah subset yang terdapat dalam suatu dataset memiliki jumlah instance yang jauh lebih kecil bila dibandingkan dengan subset lainnya lainnya. Metode oversampling dilakukan dengan cara menambahkan data di kelas minor dengan tujuan untuk menyeimbangkan jumlah distribusi data. Sedangkan metode undersampling dilakukan dengan mengeluarkan data kelas mayoritas agar distribusi data menjadi seimbang. Adapun jumlah data yang dihapus disesuaikan dengan persentase metode undersampling yang dipilih.

Ten Fold Validation

Ten fold validation adalah suatu metode validasi yang umum digunakan dalam penelitian. Ten fold validation dilakukan untuk memastikan perbandingan yang adil antara model statistik[6]. Dataset dibagi ke dalam sepuluh subset yang mana setiap subset mewakili kategori dari target penelitian. Sembilan dari sepuluh subset tersebut dijadikan data training sedangkan satu subset sisanya dijadikan data testing. Dengan kata lain, 90% data dari dataset dijadikan data training dan 10% data dari dataset dijadikan data testing. Proses validasi data dilakukan sebanyak sepuluh kali dengan catatan setiap proses validasi harus mengganti subset yang menjadi data training dan data testing. Istilah ini dikenal dengan nama Leave One Out Cross Validation (LOOCV).

III. METODE PENELITIAN

Sebelum melakukan suatu penelitian, hal yang harus dilakukan adalah menentukan langkah-langkah penelitian agar penelitian terarah sesuai dengan apa yang diharapkan.

Framework

Langkah-langkah dalam penelitian ini dapat dijabarkan sebagai berikut :

Pertama-tama, data terlebih dahulu dibersihkan dengan cara mendeteksi data outlier pada dataset tersebut. Apabila terdapat data yang termasuk data outlier, data tersebut dikeluarkan dari dataset . Dataset tersebut dinamakan sebagai data outlier, sedangkan data yang tidak dikeluarkan dari dataset disimpan sebagai data non outlier. Identifikasi data outlier dilakukan dengan menggunakan software SPSS Statistics. Apabila nilai outlier lebih besar dari 3, maka baris data tersebut merupakan data outlier [1]. Teknik menghilangkan data outlier yang digunakan dalam penelitian ini adalah menghlangkan baris-baris data yang terdeteksi sebagai data outlier dari dalam dataset.

Setelah tahap data outlier, selanjutnya dataset tersebut diuji apakah data tersebut memiliki multikolonieritas. Apabila ada data yang terdeteksi multikolonieritas, data tersebut harus dikeluarkan dari dataset. Dataset tersebut sedangkan data yang tidak dikeluarkan dari dataset disimpan sebagai data non kolonier. Tahapan yang sama juga dilakukan untuk data non outlier. Uji multikolonieritas dilakukan dengan menggunakan software SPSS Statistics. Menurut Ghozali[1], jika pada uji multikolonieritas ditemukan nilai Condition Index (CI) lebih besar dari 30, maka terdapat multikolonieritas yang sangat kuat. Jika nilai CI antara 10 dan 30, maka terdapat multikolonieritas yang cukup kuat. Jika nilai CI lebih kecil dari 10, maka multikolonieritasnya lemah.

Selanjutnya data yang multikolonier akan dianalisis dengan metode Anova. Berdasarkan Anova akan dilihat hubungan antara dua variabel independen. Two Ways Anova digunakan untuk menganalisis hubungan

antara satu variabel dependen dengan dua variabel independen. Variabel dependen adalah status_kredit sedangkan variabel lainnya yang diperlukan dalam proses Two Ways Anova ini diambil dari variabel-variabel german credit lainnya. Dalam melakukan Anova, hal yang harus diperhatikan adalah nilai level of significance atau yang lebih dikenal sebagai α. Besarnya nilai α yang digunakan dalam penelitian ini adalah sebesar 5% atau 0.05. Jika nilai alpha lebih kecil dari 0.05, maka terdapat ketergantungan antar variabel independen di dalam dataset [1].

Pada proses pendekatan data, semua dataset akan dianalisis dan dihitung nilai persentase dari hitrates dan accuracy. Nilai persentase hitrates dan accuracy dari setiap dataset akan dibandingkan setelah itu akan dilihat dataset mana yang memiliki nilai persentase hitrates terbesar dan nilai persentase accuracy terbesar. Pendekatan data digunakan untuk menganalisis nilai hitrates, accuracy dan status kredit yang baik dan buruk dalam suatu dataset. Adapun nilai hitrates adalah nilai persentase dari status kredit yang dideteksi buruk dan pada data sebenarnya juga terdeteksi buruk terhadap data status kredit yang terdeteksi buruk. Nilai accuracy adalah nilai persentase hasil prediksi yang benar dari status kredit yang baik dan yang buruk terhadap jumlah seluruh data di dalam dataset.

Hardware dan Software

Pada penelitian ini dibutuhkan hardware dan software dengan spesifikasi tertentu yang akan digunakan untuk mendukung penelitian. Hardware yang digunakan dalam penelitian ini adalah sebuah notebook Intel Pentium Dual Core Processor T2130 (1.86 Ghz 1MB L2 Cache 533 MHZ FSB), 1.5 GB DDR2 RAM, 120 GB HDD 5400RPM, dan Intel Graphics Media Accelerator (GMA) 950. Sedangkan software yang digunakan dalam penelitian ini adalah IBM SPSS Statistics Desktop dan WEKA. IBM SPSS versi 19.0.0 adalah suatu software yang memiliki fungsi statistik dasar seperti statistik deskriptif (tabulasi silang, frekuensi, statistik deskriptif rasio), statistik bivariat (means, t-test, Anova, korelasi , test non parametik), prediksi untuk mengidentifikasi kelompok (analisis faktor, analisis klaster, analisis diskriminan). Sedangkan software WEKA versi 3.6.4 adalah aplikasi data mining open source berbasis Java yang terdiri dari koleksi algoritma yang dapat digunakan untuk melakukan generalisasi atau formulasi dari sekumpulan data sampling. WEKA saat ini sudah cukup banyak mendukung algoritma untuk pemodelan data atau biasa disebut classifier, diantaranya adalah logistic regression, linear regression, naïve bayes dan lain-lain.

Data

Data yang digunakan dalam penelitian ini adalah data german credit. Data ini berisi atribut numerik yang dikonversi dari dataset asli yang diberikan penelitian. Data ini diproduksi oleh Strathclyde University, berisi atribut numerik yang dikonversi dari dataset asli yang diberikan oleh Prof. Dr. Hans Hofmann (Institut Statistika dan Ekonometrika Universitas Hamburg). Data german credit dipilih sebagai data yang digunakan dalam penelitian ini dikarenakan beberapa hal sebagai berikut : Data german credit merupakan data yang sering digunakan sebagai contoh kasus dalam penelitian termasuk juga dalam penulisan karya ilmiah. Contohnya karya ilmiah berjudul "Bayesian network classifiers for the German credit data" yang ditulis oleh Scott A. Zonneveldt, Kevin B. Korb dan Ann E. Nicholson dari Monash University tahun 2007, dan karya ilmiah berjudul "Credit scoring with a data mining approach based on support vector machines" yang ditulis Cheng-Lung Huang, Mu-Chen Chen dan Chieh-Jen Wang dari Taiwan pada tahun 2007 [4]. Data german credit juga berisi informasi yang sangat diperlukan dalam penelitian yaitu status_kredit sebagai variabel dependen yang akan dipakai sebagai target estimasi dari variabel yang lain. Kemudian, data German credit terdiri atas 1000 baris data sehingga data german credit dikatakan cukup untuk dianalisis. Data German credit dapat diimplementasikan di Indonesia karena pada data german credit terdapat parameter-parameter kredit yang pada umumnya terdapat pada data kredit di Indonesia, misalnya parameter pekerjaan, usia, dan sebagainya.

Variable

Variable memegang peranan penting dalam proses eksperimen. Variabel yang dimaksud adalah variabel parameter kredit. Parameter kredit adalah atribut-atribut yang terdapat dalam suatu dataset. Setiap atribut tersebut mewakili suatu kriteria yang diklasifikasikan dengan tujuan menghasilkan nilai pembobotan dari kriteria tersebut. Data german credit terdiri terdiri dari 7 atribut numerik dan 13 atribut kategorikal.

Algoritma

Ada banyak algoritma yang dapat digunakan untuk menganalisis data, namun penelitian ini menggunakan algoritma logistic regression dengan alasan sebagai berikut : Pertama, karakteristik dataset german credit adalah multivariate (data yang terdiri dari banyak variabel dan antar variabel saling berkorelasi) dan karakteristik atribut data german credit adalah kategorikal dan numerik. Oleh sebab itu diperlukan algoritma yang sesuai untuk menganalisis dataset dengan karakteristik tersebut. Kedua, data german credit memiliki satu variabel dependen dan lebih dari satu variabel independen. Variabel dependen yang terdapat dalam dataset german credit terdiri dari dua kategori yaitu good dan bad. Algoritma-algoritma

yang dapat menganalisis variabel dependen dengan satu non metrik dua kategori adalah logistic regression dan analisis diskriminan. Data german credit terdiri dari variabel-variabel independen yang memiliki ketergantungan satu sama lain. Algoritma analisis diskriminan tidak dapat menganalisis data yang seperti itu, maka algoritma logistic regression dipilih sebagai algoritma yang digunakan dalam penelitian karena algoritma logistic regression memiliki kemampuan untuk menganalisis data yang memiliki ketergantungan antar variabel independen.

IV. PEMBAHASAN

Pada tahapan pembersihan data, terdapat tujuh variabel yang memiliki nilai diatas 3 yaitu satu record pada variabel duration, lima record pada variabel credit_amount, dan satu record pada variabel age. Untuk data-data yang terdeteksi outlier ini, dikeluarkan dari dalam dataset dan sisanya disimpan sebagai hasil proses identifikasi outlier.

Tabel 1. Nilai skor outlier

SQR DURATION	SQR CREDIT AMOUNT	SQR AGE	ZSQR DURATION	ZSQR CREDIT AMOUNT	ZSQR AGE
6.93	120.09	5.00	2.00362	3.12228	-0.97828
7.35	126.27	7.62	2.33585	3.41023	1.89055
2.45	37.07	8.66	-1.53695	-0.74216	3.03608
7.75	125.11	4.58	2.65009	3.35616	1.43608
8.49	74.80	4.90	3.23454	1.01421	-1.08907
6.24	119.08	5.48	1.46352	3.07518	-0.45488
2.45	122.05	8.25	-1.53695	3.21359	2.58198

Setelah mendeteksi adanya data outlier pada dataset, penelitian dilanjutkan dengan menguji multikolonieritas atau biasa disebut multicollinearity.

Diagnostic untuk melihat ada tidaknya korelasi antara variabel independen. Pada bagian correlations dapat terdapat nilai -0,680 pada variabel duration dan credit_amount. Hal ini berarti bahwa terdapat korelasi sebesar 68% pada variabel duration dan credit_amount. Karena hasil persentase korelasi kedua variabel lebih dari 50%, maka dapat dipastikan bahwa variabel duration dan credit_amount memiliki hubungan yang kuat satu sama lain. Pada tabel collinearity diagnostics terdapat nilai Condition Index (CI) sebesar 17,194 pada dimensi ke 8. Nilai CI berada diantara 10 dan 30 yang berarti bahwa terdapat multikolonieritas yang cukup kuat pada variabel independen yang digunakan. Pada bagian correlations di tabel coefficient correlations terdapat nilai -0,676 pada variabel duration dan credit_amount. Hal ini berarti bahwa terdapat korelasi sebesar 67,6% pada variabel duration dan credit_amount. Hasil persentase korelasi kedua variabel lebih dari 50%, yang berarti bahwa variabel duration dan credit_amount memiliki hubungan yang kuat satu sama lain. Pada tabel collinearity diagnostics terdapat nilai Condition Index (CI). sebesar 17,167 pada dimensi ke 8 yang berarti bahwa terdapat multikolonieritas yang cukup kuat di antara variabel-variabel independen. Berdasarkan hasil uji multikolonieritas pada dataset outlier dan dataset non outlier dapat disimpulkan bahwa terdapat multikolonieritas yang cukup kuat pada variabel duration dan credit_amount.

Setelah mengetahui variabel-variabel apa saja yang memiliki multikolonieritas, langkah selanjutnya ialah menguji coba dataset german credit dengan menggunakan metode Two Ways Anova untuk memeriksa hubungan joint effect dua atau lebih variabel terhadap variabel dependen. Hubungan joint effect yang dimaksud adalah hubungan ketergantungan antara satu variabel independen dengan variabel independen lainnya. Pada dataset german credit terdapat 1 variabel dependen dan 20 variabel independen. Setiap variabel independen tersebut dibandingkan satu dengan yang lainnya terhadap variabel dependen dengan melihat nilai tingkat signifikasinya.

Hasil uji coba Anova pada dataset OU0CL0 adalah tingkat signifikasi variabel status_of*duration adalah 0.242. Nilai sebesar 0.242 lebih besar dari 0.05. Angka 0.05 digunakan sebagai batasan untuk menentukan tingkat signifikasi yang mana jika lebih besar dari 0.05 maka tidak signifikan dan jika nilainya lebih kecil dari 0.05 maka dinyatakan signifikan. Tingkat signifikasi variabel status_of*property adalah 0.002. Hal ini berarti bahwa terdapat hubungan yang signifikan antara status_of*property. Hasil lengkapnya dapat dilihat pada Tabel 2 di bawah ini.

Tabel 2. Hasil Ujicoba ANOVA pada OU0CL0

Hasil Uji Anova Dataset OU0CL0 yang Signifikan		
1	status_of*property	0.002
2	status_of*age	0.044
3	status_of*job	0.002
4	property*present_emp	0.050
5	job*foreign_worker	0.049
6	present_emp*residence	0.001
7	present_emp*housing	0.040
8	residence*housing	0.041
9	housing*credit_amount	0.053
10	housing*installment_rate	0.036
11	housing*installment_plans	0.025
12	housing*number_of	0.024
13	duration*credit_amount	0.021
14	credit_hist*installment_plans	0.002
15	purpose*installment_plans	0.013
16	age*number_of	0.045
17	number_of*installment_rate	0.051
18	number_of*telephone	0.030
19	telephone*installment_rate	0.001

Berdasarkan hasil tersebut dibuat sebuah matriks pemetaan hubungan setiap variabel. Setelah itu dilihat variabel mana yang paling banyak memiliki relasi

dengan variabel lain. Variabel tersebut kemudian dihapus, begitu pula dengan variabel-variabel lain yang masih memiliki relasi dengan variabel lain. Variabel-variabel tersebut dihapus mulai dari variabel yang paling banyak memiliki relasi dengan variabel lain hingga ke variabel yang paling sedikit memiliki relasi dengan variabel lain. Terdapat 8 variabel yang telah selesai diuji Anova, yaitu variabel property, job, residence, duration, credit_hist, purpose, age, dan installment_rate. Untuk dataset dengan variabel yang telah dihapus, akan disimpan dengan nama file OU0CL0AN1.dat, sedangkan untuk dataset yang tidak mengalami proses penghapusan variabel, akan disimpan sebagai OU0CL0AN0.dat.

Hasil uji coba Anova pada dataset OU0CL1 untuk perkalian antar variabel independen yang dinyatakan signifikan dapat dilihat pada Tabel 3. Pada dataset OU0CL1 terdapat 17 perkalian variabel yang signifikan. Kemudian dibuat matriks pemetaan sesuai dengan hubungan antar variabelnya. Terdapat 7 variabel yang dapat digunakan dan disimpan sebagai tab delimited file yang baru dengan nama OU0CL1AN1.dat. Variabel tersebut adalah property, job, residence, credit_hist, purpose, age, dan installment_rate. Untuk dataset OU0CL1 yang tidak mengalami penghapusan variabel akan disimpan sebagai tab delimited file yang baru dengan nama OU0CL1AN0.dat.

Tabel 3. Hasil Ujicoba ANOVA pada OU0CL1

Hasil Uji Anova Tabel OU0CL1 yang Signifikan		
1	status_of*property	0.002
2	status_of*age	0.044
3	status_of*job	0.002
4	property*present_emp	0.050
5	job*foreign_worker	0.049
6	present_emp*residence	0.001
7	present_emp*housing	0.040
8	residence*housing	0.041
9	housing*installment_rate	0.036
10	housing*installment_plans	0.025
11	housing*number_of	0.024
12	credit_hist*installment_plans	0.002
13	purpose*installment_plans	0.013
14	age*number_of	0.045
15	number_of*installment_rate	0.051
16	number_of*telephone	0.030
17	telephone*installment_rate	0.001

Tabel 4. Hasil Ujicoba Anova terhadap UO1CL0

Hasil Uji Anova Tabel OU1CL0 yang Signifikan		
1	status_of*property	0.002
2	status_of*job	0.002
3	job*foreign_worker	0.048
4	property*present_emp	0.050
5	present_emp*residence	0.002
6	present_emp*housing	0.032
7	residence*installment_plans	0.054
8	residence*housing	0.046
9	housing*installment_rate	0.051
10	housing*installment_plans	0.025
11	housing*number_of	0.026
12	duration*credit_amount	0.040
13	credit_hist*installment_plans	0.002
14	purpose*installment_plans	0.020
15	age*number_of	0.045
16	number_of*telephone	0.032
17	telephone*installment_rate	0.003

Hasil uji coba Anova pada dataset OU1CL0 untuk perkalian antar variabel independen yang dinyatakan signifikan dapat dilihat pada Tabel 4. Tabel ini terdiri dari 17 hasil perkalian variabel yang nilainya dibawah 0.05. Selanjutnya dilakukan proses pemetaan terhadap variabel-variabel tersebut. Setelah dilakukan proses pemetaan matriks perkalian variabel, didapatkan 8 variabel yang tersisa, yaitu variabel property, job, residence, duration, credit_hist, purpose, age, dan telephone. Variabel-variabel yang terdapat pada dataset OU1CL0 dihapus kecuali 8 variabel yang terdapat pada Gambar 4.7. Dataset yang telah dihapus variabelnya akan disimpan dengan nama file OU1CL0AN1.dat, sedangkan dataset yang tidak melalui proses penghapusan variabel, akan disimpan sebagai OU1CL0AN0.dat.

Yang terakhir adalah hasil uji coba Anova pada dataset OU1CL1 untuk perkalian antar variabel independen yang dinyatakan signifikan. Hasil analisisnya dapat dilihat pada Tabel 5 yang terdiri dari 16 hasil perkalian variabel yang nilainya dibawah 0.05. Selanjutnya dilakukan proses pemetaan terhadap 16 variabel tersebut. Hasil pemetaan matriks yang signifikan pada dataset OU1CL1 dapat dilihat pada Tabel 5. Setelah dilakukan proses pemetaan dan penghapusan variabel, maka jumlah variabel yang tersisa menjadi 6 variabel. Variabel yang tersisa adalah property, residence, credit_hist, foreign_worker, dan installment_rate. Keenam variabel tersebut disimpan ke dalam sebuah tab delimited file dengan nama OU1CL1AN1.dat. Dataset yang tidak dihapus variabelnya akan disimpan dengan nama file OU1CL1AN0.dat

Tabel 5. Hasil Uji Anova pada OU1CL1

Hasil Uji Anova Tabel OU1CL1 yang Signifikan		
1	status_of*property	0.002
2	status_of*job	0.002
3	property*present_emp	0.050
4	job*foreign_worker	0.048
5	present_emp*residence	0.002
6	present_emp*housing	0.032
7	residence*installment_plans	0.054
8	residence*housing	0.046
9	housing*number_of	0.026
10	housing*installment_rate	0.051
11	housing*installment_plans	0.025
12	credit_hist*installment_plans	0.002
13	purpose*installment_plans	0.020
14	age*number_of	0.045
15	number_of*telephone	0.032
16	telephone*installment_rate	0.003

Yang didapatkan setelah melakukan uji Anova terhadap dataset OU0CL0, OU0CL1, OU1CL0, dan OU1CL1 adalah dalam melakukan uji Anova tidak perlu dilakukan semuanya. Hal ini dikarenakan hasil uji Anova pada dataset OU0CL0 sama dengan hasil uji Anova pada dataset OU0CL1. Begitu pula dengan dataset OU1CL0 hasil uji Anovanya sama dengan dataset OU1CL1. Perbedaannya, tidak ada perkalian terhadap variabel duration dan credit_amount pada dataset OU0CL1 dan OU1CL1. Selebihnya nilai uji Anovanya memberikan hasil yang sama.

Penelitian dilanjutkan dengan tahap yang terakhir, yaitu tahap pendekatan data. Pada tahap terakhir ini akan dibandingkan perhitungan nilai accuracy dan hitrates dari setiap dataset yang telah diproses sebelumnya. Terdapat 8 dataset german credit yang telah diproses sebelumnya, yaitu OU0CL0AN0, OU0CL0AN1, OU0CL1AN0, OU0CL1AN1 ,OU1CL0AN0, OU1CL0AN1, OU1CL1AN0, dan OU1CL1AN1. Setiap dataset ini akan dianalisis dengan menggunakan bantuan software WEKA untuk perhitungan statistik dengan algoritma logistic regression dan proses validasi 10-fold validation.

Beberapa istilah hasil confusion matrix pada WEKA dapat diterangkan sbb: GG adalah data yang benar, dimana data prediksi menyatakan bahwa pembayaran kredit nasabah baik dan dibandingkan dengan data sebenarnya yang menyatakan bahwa pembayaran kredit nasabah baik. GB adalah data yang salah diprediksi, dimana data prediksi menyatakan bahwa pembayaran kredit nasabah baik dan dibandingkan dengan data sebenarnya yang menyatakan bahwa pembayaran kredit nasabah kurang baik / macet. BB adalah data yang benar, dimana data prediksi menyatakan bahwa pembayaran kredit nasabah kurang baik/ macet dan dibandingkan dengan data sebenarnya yang menyatakan bahwa pembayaran kredit nasabah kurang baik/ macet. BG adalah data yang salah diprediksi, dimana data prediksi menyatakan bahwa pembayaran kredit nasabah kurang baik/ macet dan dibandingkan dengan data sebenarnya yang menyatakan bahwa pembayaran kredit nasabah baik.

Dataset OU0CL0AN0 memiliki nilai accuracy tertinggi yaitu 75.2%. Berdasarkan nilai tersebut dapat disimpulkan bahwa langkah-langkah analisis data yang dilakukan untuk mendapatkan dataset OU0CL0AN0 adalah langkah-langkah yang paling sesuai dan akurat dalam menganalisis data pembayaran kredit pada data german credit. Dengan kata lain, dataset yang memiliki akurasi dan hitrates paling tinggi terdapat dalam data asli german credit atau yang disimpan sebagai dataset OU0CL0AN0.dat. Sedangkan pada Tabel 6 dapat dilihat bahwa nilai hitrates terbesar juga terdapat di data asli. Proses-proses yang dilakukan terhadap data seperti proses data outlier, uji multikolonieritas dan Anova ternyata tidak mampu menaikkan nilai hitrates. Proses-proses yang telah dilakukan bukan merupakan proses yang gagal, akan tetapi perlu dilakukan analisis lebih lanjut terhadap data.

Untuk menyelesaikan proses data outlier, cara yang digunakan dalam penelitian ini dengan menghapus data yang terdeteksi sebagai outlier. Penghapusan data bukan merupakan solusi yang terbaik untuk permasalahan data outlier dan dalam melakukan penghapusan data diperlukan suatu studi terhadap data german credit karena suatu data yang terdeteksi sebagai data outlier bisa saja merupakan data yang memiliki informasi yang penting. Solusi untuk menghilangkan data outlier adalah dengan memperbaiki data. Namun solusi tersebut tidak dapat dilakukan dalam penelitian ini mengingat terdapat keterbatasan waktu dalam penelitian.

Tabel 6. Perbandingan Hitrates antar Dataset

Dataset	Hitrates
OU0CL0AN0	49%
OU0CL0AN1	37.2%
OU0CL1AN0	39.9%
OU0CL1AN1	32.3%
OU1CL0AN0	48.9%
OU1CL0AN1	34.6%
OU1CL1AN0	44.2%
OU1CL1AN1	30.9%

Pada proses uji multikolonier, cara yang digunakan untuk menghilangkan multikolonier pada data german credit adalah dengan menghapus kolom yang memiliki hubungan multikolonier dengan variabel dependen. Cara penghapusan seperti ini bukanlah merupakan solusi yang terbaik. Cara yang lebih baik adalah dengan memperbaiki data-data yang terdeteksi multikolonier, bukan dengan menghapus kolom. Untuk melakukan perbaikan terhadap data-data yang terdeteksi multikolonier dibutuhkan waktu penelitian

yang lebih panjang. Oleh karena itu, solusi perbaikan data multikolonier tidak dapat dilakukan dalam penelitian ini. Pada proses Anova, cara yang digunakan untuk menghilangkan korelasi antara variabel independen yang satu dengan variabel independen lainnya adalah dengan melakukan pemetaan terhadap setiap variabel independen dan menghapus variabel yang memiliki korelasi paling banyak. Penghapusan kolom yang dilakukan dalam penelitian ini juga bukan merupakan solusi yang terbaik. Pada proses Anova, hal yang terlebih dahulu harus dilakukan adalah mencari hubungan korelasi antara dua variabel independen yang berkorelasi, kemudian penghapusan dilakukan bukan pada kolom-kolom variabel yang berinteraksi satu sama lain melainkan pada data yang menyebabkan kedua kolom tersebut berinteraksi. Untuk dapat mengetahui baris data yang memiliki masalah diperlukan studi tersendiri. Karena adanya keterbatasan waktu, hal tersebut tidak dapat dilakukan.

Dari semua dataset yang ada diketahui bahwa nilai hitrates masih belum cukup tinggi. Oleh karena itu, penelitian dilanjutkan dengan melakukan analisis uji coba dengan menggunakan metode sampling. Metode sampling yang digunakan adalah oversampling / uppersampling dan undersampling. Kedua metode sampling ini digunakan untuk menganalisis dataset yang memiliki imbalanced data. Tujuannya adalah agar hasil analisis menjadi lebih baik. Penjelasan mengenai tahap ujicoba dengan oversampling dan undersampling akan dibahas pada sub bab selanjutnya.

Metode oversampling digunakan untuk menganalisis dataset yang tidak seimbang (*imbalanced data*). Metode oversampling dilakukan dengan menyeimbangkan jumlah distribusi data dengan meningkatkan jumlah data kelas minor (yang lebih sedikit jumlahnya). Pada metode random oversampling, data kelas minor dipilih untuk diduplikasi. Akan tetapi, hasil dari random oversampling tidak selalu meningkatkan prediksi kelas minor. Apabila data kelas minor diduplikasi dalam jumlah yang besar, akan mempersulit proses identifikasi pada data yang memiliki kemiripan karakteristik namun berada di kelas yang berbeda. Sebelum melakukan oversampling, dataset terlebih dahulu dibagi menjadi 10 kelompok yang mana isi dari setiap kelompok tersebut berupa data yang dipilih secara acak. Jumlah data dari setiap kelompok harus sama, misalnya jumlah data dalam suatu dataset adalah 1000, maka setiap kelompok terdiri atas 100 data yang dipilih secara acak. Dari 10 kelompok data tersebut, dipilih 9 kelompok sebagai data training sedangkan 1 kelompok sisanya menjadi data testing. Dataset OU0CL0AN0 terdiri dari 1000 baris data. Dataset tersebut dikelompokkan menjadi 10 kelompok. Kelompok data tersebut selanjutnya akan dianalisis dengan metode oversampling. Langkah-langkah uji coba dengan menggunakan metode oversampling dapat diuraikan sbb:

Tabel 7. Confusion Matrix pada OU0CL0AN0 dengan metode oversampling

Data Testing	Data Training	GG	GB	BG	BB
OU0CL0AN0PE1	OU0CL0AN0PR1	33	20	22	25
OU0CL0AN0PE2	OU0CL0AN0PR2	60	9	11	20
OU0CL0AN0PE3	OU0CL0AN0PR3	61	9	11	19
OU0CL0AN0PE4	OU0CL0AN0PR4	69	5	16	10
OU0CL0AN0PE5	OU0CL0AN0PR5	64	14	9	13
OU0CL0AN0PE6	OU0CL0AN0PR6	59	11	16	14
OU0CL0AN0PE7	OU0CL0AN0PR7	57	19	8	16
OU0CL0AN0PE8	OU0CL0AN0PR8	57	20	5	18
OU0CL0AN0PE9	OU0CL0AN0PR9	43	28	9	20
OU0CL0AN0PEX	OU0CL0AN0PRX	40	22	9	29
Total		543	157	116	184

Berdasarkan perhitungan tersebut dapat disimpulkan bahwa nilai hitrates mengalami peningkatan dari 49% menjadi 61.3%, namun nilai accuracy mengalami penurunan dari 75.2% menjadi 72.7%. Hal ini berarti bahwa persentase nilai status kredit yang bad terhadap seluruh data yang terdeteksi bad mengalami peningkatan sebanyak 12.3%. Sehubungan dengan tujuan penelitian yang difokuskan hanya untuk menganalisis data pembayaran kredit yang buruk, maka peningkatan hitrates dinilai lebih menunjukkan keberhasilan estimasi kredit yang buruk, sedangkan penurunan nilai accuracy tidak relevan dalam menggambarkan hasil estimasi dikarenakan accuracy merupakan estimasi nasabah yang buruk dan yang baik.

Berbeda dengan metode oversampling, metode undersampling dilakukan dengan menghapus data kelas mayoritas agar dataset menjadi seimbang. Penghapusan data kelas mayoritas disesuaikan berdasarkan persentase data yang ingin dihapus. Misalnya menghapus data yang memiliki nilai status_ kredit good sebanyak 10% dari jumlah baris pada dataset.

Setelah melakukan analisis data dengan metode oversampling, penelitian dilanjutkan dengan analisis data menggunakan metode undersampling. Metode undersampling memiliki kemiripan dengan metode oversampling, yang mana dataset terlebih dahulu harus dibagi ke dalam 10 kelompok dan isi dari setiap kelompok dipilih secara acak. Analisis dengan metode undersampling pada dataset OU0CL0AN0 dilakukan sebanyak 5 kali. Yang pertama adalah undersampling dengan 90% good. Maksudnya data yang memiliki nilai status_kredit good akan dihapus sebanyak 10% dari total dataset training. Berarti sebanyak 90 baris data pada data training dihapus. Kemudian undersampling dilanjutkan dengan persentase 80%, 70%, 60% dan 50%. Jumlah data yang dihapus disesuaikan dengan sisa persentasenya. Nilai hitrates dengan metode undersampling 90% diketahui sebesar 57% dan nilai accuracy sebesar 74.5% seperti terlihat pada Tabel 8. Nilai hitrates dengan metode undersampling 80% adalah sebesar 64.6% dan nilai accuracy sebesar

71.1%.

Tabel 8 Confusion Matrix pada OU0CL0AN0 dengan metode undersampling 90%

Data Testing	Data Training	GG	GB	BG	BB
OU0CL0AN0NE90_1	OU0CL0AN0NR90_1	70	1	20	9
OU0CL0AN0NE90_2	OU0CL0AN0NR90_2	37	24	15	24
OU0CL0AN0NE90_3	OU0CL0AN0NR90_3	41	17	10	32
OU0CL0AN0NE90_4	OU0CL0AN0NR90_4	51	19	9	21
OU0CL0AN0NE90_5	OU0CL0AN0NR90_5	48	17	13	22
OU0CL0AN0NE90_6	OU0CL0AN0NR90_6	54	16	12	18
OU0CL0AN0NE90_7	OU0CL0AN0NR90_7	70	6	13	11
OU0CL0AN0NE90_8	OU0CL0AN0NR90_8	67	12	13	8
OU0CL0AN0NE90_9	OU0CL0AN0NR90_9	63	8	13	16
OU0CL0AN0NE90_X	OU0CL0AN0NR90_X	73	6	11	10
Total		574	126	129	171

Tabel 9. Confusion Matrix pada OU0CL0AN0 dengan metode undersampling 80%

Data Testing	Data Training	GG	GB	BG	BB
OU0CL0AN0NE80_1	OU0CL0AN0NR80_1	71	0	25	4
OU0CL0AN0NE80_2	OU0CL0AN0NR80_2	29	32	5	34
OU0CL0AN0NE80_3	OU0CL0AN0NR80_3	28	30	7	35
OU0CL0AN0NE80_4	OU0CL0AN0NR80_4	40	30	5	25
OU0CL0AN0NE80_5	OU0CL0AN0NR80_5	38	27	10	25
OU0CL0AN0NE80_6	OU0CL0AN0NR80_6	48	22	8	22
OU0CL0AN0NE80_7	OU0CL0AN0NR80_7	66	10	12	12
OU0CL0AN0NE80_8	OU0CL0AN0NR80_8	65	14	10	11
OU0CL0AN0NE80_9	OU0CL0AN0NR80_9	59	12	14	15
OU0CL0AN0NE80_X	OU0CL0AN0NR80_X	73	6	10	11
Total		517	183	106	194

Selanjutnya nilai hitrates dengan metode undersampling 70% adalah sebesar 71.6% dan nilai accuracy sebesar 66.1% (Tabel 10). dengan metode undersampling 60% adalah 80% dan nilai accuracy sebesar 59.7% (Tabel 11). Nilai hitrates dengan metode undersampling 50% adalah sebesar 83.3% dan nilai accuracy sebesar 48.8% (Tabel 12).

Setelah melakukan perhitungan hitrates dan accuracy sesuai dengan nilai dari setiap variabel, maka dapat disimpulkan bahwa nilai hitrates terbesar terdapat pada analisis undersampling 50%, yang mana memiliki nilai hitrates sebesar 83.3%. Sedangkan, nilai hitrates terkecil terdapat pada analisis undersampling 90% sebesar 57%. Adapun nilai hitrates yang terkecil masih lebih besar dibandingkan dengan nilai hitrates pada data asli german credit yang besarnya 49%. Nilai accuracy terbesar terdapat pada analisis undersampling 90%, yang mana memiliki nilai accuracy sebesar 74.5%. Sedangkan, nilai accuracy terkecil terdapat pada analisis undersampling 50% sebesar 48.8%. Adapun nilai accuracy terbesar masih lebih kecil dibandingkan dengan nilai accuracy pada data asli german credit yang besarnya 75.2%.

Tabel 10. Confusion Matrix pada OU0CL0AN0 dengan metode undersampling 70%

Data Testing	Data Training	GG	GB	BG	BB
OU0CL0AN0NE70_1	OU0CL0AN0NR70_1	71	0	27	2
OU0CL0AN0NE70_2	OU0CL0AN0NR70_2	18	43	3	36
OU0CL0AN0NE70_3	OU0CL0AN0NR70_3	19	39	2	40
OU0CL0AN0NE70_4	OU0CL0AN0NR70_4	26	44	3	27
OU0CL0AN0NE70_5	OU0CL0AN0NR70_5	28	37	8	27
OU0CL0AN0NE70_6	OU0CL0AN0NR70_6	42	28	7	23
OU0CL0AN0NE70_7	OU0CL0AN0NR70_7	56	20	7	17
OU0CL0AN0NE70_8	OU0CL0AN0NR70_8	61	18	7	14
OU0CL0AN0NE70_9	OU0CL0AN0NR70_9	54	17	10	19
OU0CL0AN0NE70_X	OU0CL0AN0NR70_X	71	8	11	10
Total		446	254	85	215

Tabel 11. Confusion Matrix pada OU0CL0AN0 dengan metode undersampling 60%

Data Testing	Data Training	GG	GB	BG	BB
OU0CL0AN0NE60_1	OU0CL0AN0NR60_1	71	0	25	4
OU0CL0AN0NE60_2	OU0CL0AN0NR60_2	4	57	1	38
OU0CL0AN0NE60_3	OU0CL0AN0NR60_3	8	50	1	41
OU0CL0AN0NE60_4	OU0CL0AN0NR60_4	15	55	1	29
OU0CL0AN0NE60_5	OU0CL0AN0NR60_5	14	51	2	33
OU0CL0AN0NE60_6	OU0CL0AN0NR60_6	33	37	3	27
OU0CL0AN0NE60_7	OU0CL0AN0NR60_7	41	35	3	21
OU0CL0AN0NE60_8	OU0CL0AN0NR60_8	56	23	3	18
OU0CL0AN0NE60_9	OU0CL0AN0NR60_9	51	20	10	19
OU0CL0AN0NE60_X	OU0CL0AN0NR60_X	64	15	11	10
Total		357	343	60	240

Tabel 12. Confusion Matrix pada OU0CL0AN0 dengan metode undersampling 50%

Data Testing	Data Training	GG	GB	BG	BB
OU0CL0AN0NE50_1	OU0CL0AN0NR50_1	69	2	27	2
OU0CL0AN0NE50_2	OU0CL0AN0NR50_2	0	61	1	38
OU0CL0AN0NE50_3	OU0CL0AN0NR50_3	1	57	0	42
OU0CL0AN0NE50_4	OU0CL0AN0NR50_4	3	67	0	30
OU0CL0AN0NE50_5	OU0CL0AN0NR50_5	1	64	0	35
OU0CL0AN0NE50_6	OU0CL0AN0NR50_6	15	55	1	29
OU0CL0AN0NE50_7	OU0CL0AN0NR50_7	19	57	1	23
OU0CL0AN0NE50_8	OU0CL0AN0NR50_8	27	52	3	18
OU0CL0AN0NE50_9	OU0CL0AN0NR50_9	41	30	7	22
OU0CL0AN0NE50_X	OU0CL0AN0NR50_X	62	17	10	11
Total		238	462	50	250

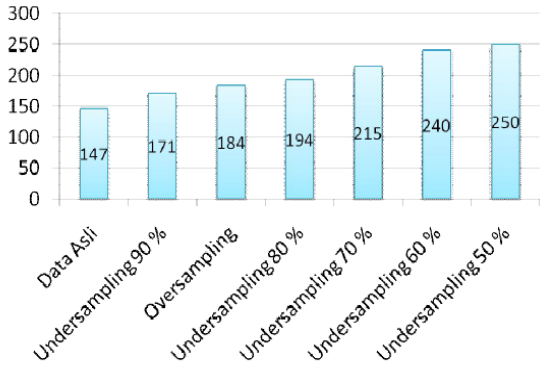
Yang didapatkan setelah melakukan analisis data dengan metode oversampling dan undersampling adalah proses analisis data dengan menggunakan metode oversampling dan undersampling dapat meningkatkan nilai hitrates pada kasus data german credit, yang mana peningkatan nilai hitrates yang paling tinggi terdapat pada proses undersampling 50%. Akan tetapi, proses analisis data dengan menggunakan metode oversampling dan undersampling ternyata tidak dapat meningkatkan nilai accuracy pada kasus data german credit dan kedua metode tersebut ternyata menurunkan nilai accuracy. Perbandingan nilai hitrates dan accuracy pada data asli, data oversampling dan data undersampling dapat dilihat pada Tabel 13.

Tabel 13. Perbandingan HitRates dan Accuracy antar Dataset

Dataset	Hitrates	Accuracy
Data asli (OU0CL0AN0.dat)	49%	75.2%
Data oversampling	61.3%	72.7%
Data undersampling 90%	57%	74.5%
Data undersampling 80%	64.6%	71.1%
Data undersampling 70%	71.6%	66.1%
Data undersampling 60%	80%	59.7%
Data undersampling 50%	83.3%	48.8%

Pada penelitian ini, nilai hitrates dipilih untuk dijadikan nilai tolak ukur dalam penelitian. Nilai persentase pada hitrates menunjukkan persentase nasabah dengan status kredit yang benar-benar *bad* (data yang diprediksi *bad* dan data sebenarnya juga

bad). Sedangkan, nilai pada accuracy menunjukan nilai persentase nasabah dengan status kredit yang benar-benar *bad* ditambahkan dengan yang benar-benar *good* sehingga terjadi keambiguan apabila nasabah dengan status kredit yang benar-benar *bad* dijadikan fokus. Oleh sebab tujuan dari penelitian ini adalah menganalisis data pembayaran kredit yang *bad*, maka nilai hitrates lebih mewakili tujuan dari penelitian ini dibandingkan dengan nilai accuracy.



Gambar 1. Jumlah Nasabah dengan kredit buruk yang terdeteksi oleh setiap metode

Gambar 1 menggambarkan jumlah nasabah dengan status kredit bad yang dapat dideteksi oleh setiap proses. Proses undersampling 50% paling banyak menebak jumlah nasabah dengan status kredit bad. Maka dapat disimpulkan metode yang paling baik untuk mendeteksi jumlah nasabah dengan status kredit bad pada dataset german credit adalah metode undersampling 50%.

V. PENUTUP

Kesimpulan yang bisa ditarik dari penelitian ini adalah: Data german credit merupakan data well preprocessed yang berarti bahwa data german credit sudah melalui proses data cleaning jadi tidak perlu dilakukan proses pembersihan data. Metode undersampling dan oversampling pada data german credit mampu menaikkan nilai hitrates menjadi lebih tinggi dibandingkan nilai hitrates pada data asli yang berarti bahwa kedua metode tersebut mampu mendeteksi nasabah dengan status kredit bad secara lebih optimal. Jumlah nasabah dengan status kredit bad yang dapat dideteksi oleh metode yang paling optimal (metode undersampling 50%) adalah 250. Cara menemukan model atau pola terbaik untuk pembayaran kredit nasabah pada data german credit adalah dengan tidak melakukan proses data cleaning karena data german credit dapat langsung digunakan, perlu melakukan analisis undersampling untuk mendapatkan hasil yang terbaik dan menggunakan algoritma logistic regression karena algoritma logistic regression merupakan algoritma yang sesuai untuk menganalisis data german credit. Hubungan parameter kredit dan status kredit terlihat dalam model logistic regression dengan dimensi multivariate, yang mana parameter kredit merupakan variabel independen dan status kredit merupakan variabel dependen. Parameter-parameter yang diperlukan dalam pola pembayaran kredit pada data german credit adalah semua parameter kredit karena setiap parameter memiliki informasi yang dibutuhkan. Parameter-parameter tersebut adalah 20 variabel independen pada data german credit. Parameter yang paling berpengaruh dalam mengidentifikasi nasabah yang tidak lancar membayar kredit pada data german credit adalah semua parameter karena setiap parameter memiliki hubungan satu sama lain dan memiliki informasi yang dibutuhkan dalam menganalisis data german credit. Parameter-parameter yang dimaksud adalah 20 variabel independen pada data german credit. Proses data outlier, uji multikolonieritas dan Anova tidak dapat mendeteksi data pembayaran kredit yang buruk pada data german credit secara signifikan.

Cara mengoptimalkan hasil analisis pembayaran kredit terhadap status kredit yang buruk atau bad pada data german credit adalah dengan melakukan analisis undersampling 50%.

Saran yang dapat diberikan apabila ingin melanjutkan penelitian ini adalah menambahkan tahap segmentasi untuk persentase status kredit yang bad dan good. Tahap segmentasi dapat dilakukan pada nilai status kredit yang presentase badnya dari 55%, kemudian 60%, sampai dengan 100%. Saran berikutnya adalah penelitian selanjutnya sebaiknya dilakukan dengan menggunakan data yang berasal dari perusahaan atau organisasi tertentu dengan harapan dapat menemukan hasil analisis yang lebih bervariasi.

DAFTAR PUSTAKA

[1] Ghozali, Imam, Aplikasi Analisis Multivariate dengan Program SPSS, Semarang: Badan Penerbit Universitas Diponegoro, 2001.

[2] Hair, Joseph, et al, Multivariate Data Analysis with Readings, Englewood Cliffs: Prentice Hall, 1995.

[3] Hand, David J, dan Smyth P., Principles of Data Mining, Cambridge: MIT Press, 2001.

[4] Huang, Y.-M., Hung, C.-M., & Jiau, H. C. (2006). Evaluation of neural networks and data mining methods on a credit assessment task for class imbalance problem. Nonlinear Analysis: Real World Applications, 7(4), 720-747.

[5] Olson, David dan Yong Shi, Introduction to Business Data Mining, Boston:McGraw-Hill, 2007.

[6] Rhodes, Daniel, New Topics in Environmental Research, New York: Nova Science Publishers, 2006.

[7] Sulianta, Feri dan Dominikus Juju, Data Mining Meramalkan Bisnis Perusahaan, Jakarta: Elex Media Komputindo, 2010.