TUGAS AKHIR – KS184822

ANALISIS KLASIFIKASI NASABAH TELAT BAYAR DI BANK "X" MENGGUNAKAN METODE RANDOM FOREST DAN REGRESI LOGISTIK BINER

M. KHOLILUL MUTA'AL
NRP 062116 4000 0013

Dosen Pembimbing
Dr. rer. pol. Heri Kuswanto, S.Si., M.Si.

PROGRAM STUDI SARJANA
DEPARTEMEN STATISTIKA
FAKULTAS SAINS DAN ANALITIKA DATA
INSTITUT TEKNOLOGI SEPULUH NOPEMBER
SURABAYA 2020

(Halaman ini sengaja dikosongkan)

TUGAS AKHIR – KS184822

ANALISIS KLASIFIKASI NASABAH TELAT BAYAR DI BANK "X" MENGGUNAKAN METODE RANDOM FOREST DAN REGRESI LOGISTIK BINER

M. KHOLILUL MUTA'AL
NRP 062116 4000 0013

Dosen Pembimbing
Dr. rer. pol. Heri Kuswanto, S.Si., M.Si.

PROGRAM STUDI SARJANA
DEPARTEMEN STATISTIKA
FAKULTAS SAINS DAN ANALITIKA DATA
INSTITUT TEKNOLOGI SEPULUH NOPEMBER
SURABAYA 2020

(Halaman ini sengaja dikosongkan)

FINAL PROJECT – KS184822

LATE PAYMENT CLASSIFICATION ANALYSIS OF CUSTOMER'S "X" BANK USING RANDOM FOREST METHOD AND BINARY LOGISTIC REGRESSION

M. KHOLILUL MUTA'AL
SN 062116 4000 0013

Supervisor
Dr. rer. pol. Heri Kuswanto, S.Si., M.Si.

UNDERGRADUATE PROGRAMME
DEPARTMENT OF STATISTICS
FACULTY OF SCIENCE AND DATA ANALYTICS
INSTITUT TEKNOLOGI SEPULUH NOPEMBER
SURABAYA 2020

(Halaman ini sengaja dikosongkan)

LEMBAR PENGESAHAN

ANALISIS KLASIFIKASI NASABAH TELAT BAYAR DI BANK "X" MENGGUNAKAN METODE RANDOM FOREST DAN REGRESI LOGISTIK BINER

TUGAS AKHIR
Diajukan untuk Memenuhi Salah Satu Syarat
Memperoleh Gelar Sarjana Statistika
pada
Program Studi Sarjana Departemen Statistika
Fakultas Sains dan Analitika Data
Institut Teknologi Sepuluh Nopember

Oleh:
M. Kholilul Muta'al
NRP. 062116 4000 0013

Disetujui oleh Pembimbing:
Dr. rer. pol. Heri Kuswanto, S.Si., M.Si. ( )
NIP. 19820326 200312 1 004

Mengetahui,
Kepala Departemen Statistika

Dr. Dra. Kartika Fithriasari, M.Si
NIP. 19691212 199303 2 002

SURABAYA, Agustus 2020

(Halaman ini sengaja dikosongkan)

ANALISIS KLASIFIKASI NASABAH TELAT BAYAR DI BANK "X" MENGGUNAKAN METODE RANDOM FOREST DAN REGRESI LOGISTIK BINER

Nama Mahasiswa : M. Kholilul Muta'al
NRP : 062116 4000 0013
Departemen : Statistika-FSAD-ITS
Dosen Pembimbing : Dr.rer.pol. Heri Kuswanto, S.Si., M.Si.

ABSTRAK

Abstrak

Bank merupakan sebuah lembaga keuangan yang menghimpun dana dari masyarakat dalam bentuk simpanan dan menyalurkan kepada masyarakat dalam bentuk kredit dana atau bentuk-bentuk lain dengan tujuan untuk meningkatkan taraf hidup orang banyak. Salah satu produk bank yang memberikan pemasukan yang besar adalah kredit. Dari produk bank tersebut tentunya bisa memunculkan masalah bagi bank, salah satunya adalah resiko nasabah membayar angsuran yang telat atau tidak sesuai waktu jatuh tempo. Analisa mengenai telat bayar ini guna untuk mengukur resiko nasabah mengalami telat bayar menggunakan metode klasifikasi. Metode yang digunakan dalam penelitian adalah Random Forest dan Regresi Logistik Biner. Penelitian ini nantinya ingin melakukan perbandingan dari kedua metode tersebut untuk mencari metode yang terbaik dalam menganalisis telat bayar nasabah. Data yang digunakan dalam penelitian ini merupakan data pembiayaan nasabah di Bank "x" pada tahun 2015-2018. Nasabah akan diklasifikasikan dalam dua kelas, yaitu nasabah yang telat membayar angsuran dan nasabah yang tidak telat membayar angsuran. Hasil penelitian menunjukkan bahwa metode Random Forest didapatkan nilai akurasi sebesar 70,63% dan AUC sebesar 72,75% sedangkan pada metode Regresi Logistik Biner didapatkan nilai akurasi sebesar 57,67% dan AUC sebesar 59,98%. Hal tersebut menunjukan bahwa metode Random Forest lebih baik dibandingkan dengan metode Regresi Logistik Biner.

Kata Kunci : AUC, Klasifikasi, Nasabah Telat Bayar, Random Forest, Regresi Logistik Biner

(Halaman ini sengaja dikosongkan)

LATE PAYMENT CLASSIFICATION ANALYSIS OF CUSTOMER'S "X" BANK USING RANDOM FOREST METHOD AND BINARY LOGISTIC REGRESSION

Name : **M. Kholilul Muta'al**
Student Number : **062116 4000 0013**
Department : **Statistics**
Supervisor : **Dr.rer.pol. Heri Kuswanto, S.Si., M.Si.**

Abstract

A bank is a financial institution that accepts deposits from the public and distributes to the public in the form of credit funds or other forms to improve the lives of many people. One of the bank products that provide a large income is credit. That bank's products can certainly bring up problems for banks, one of the risks is the customer is late in paying installments or not paying according to the due date. This analysis of late payment is to measure the risk of customer experiencing late payment using the classification method. The methods that being used in this research are Random Forest and Binary Logistic Regression. This research's aim is to find the most suitable method in analyzing the customer's late payment by comparing the two methods. The data used in this study is customer financing data at "X" Bank in 2015-2018. Customers will be classified into two classes, namely customers who are late in paying installments and customers who are not late in paying installments. The results showed that the Random Forest method obtained an accuracy value of 70.63% and AUC of 72.75% while the Binary Logistic Regression method obtained an accuracy value of 57.67% and AUC of 59.98%. This result shows that the Random Forest method is better than the Binary Logistic Regression method.

Keywords: AUC, Binary Logistic Regression, Classification, Late Payment, Random Forest

(Halaman ini sengaja dikosongkan)

KATA PENGANTAR

Puji Syukur penulis panjatkan atas berkat, rahmat, dan karunia yang telah diberikan Allah SWT sehingga penulis dapat menyelesaikan Laporan Tugas Akhir ini yang berjudul **"Analisis Klasifikasi Nasabah Telat Bayar di Bank "X" Menggunakan Metode Random Forest dan Regresi Logistik Biner "** dengan tepat waktu.

Penulis menyadari dalam penyusunan Tugas Akhir ini tidak akan selesai tanpa bantuan maupun dukungan dari berbagai pihak. Pada kesempatan ini penulis menyampaikan terima kasih kepada:

1. Kedua orang tua dan keluarga penulis yang selalu memberikan doa dan dukungan kepada penulis selama penyusunan Tugas Akhir.
2. Dr. Kartika Fithriasari, M.Si. selaku Kepala Departemen Statistika dan Dr. Santi Wulan Purnami, S.Si., M.Si. selaku Sekretaris Departemen I Bidang Akademik dan Kemahasiswaan yang telah memberikan fasilitas, sarana dan prasarana.
3. Dr. Dra. Agnes Tuti Rumiati, M.Sc. selaku dosen yang menjadi dosen wali selama masa studi yang telah membantu dan arahan dalam proses belajar di Departemen Statistika
4. Dr. rer. pol. Heri Kuswanto, S.Si., M.Si. selaku dosen pembimbing yang telah memberikan bimbingan, saran, serta motivasi selama penyusunan Tugas Akhir berlangsung.
5. Santi Puteri Rahayu, S.Si., M.Si., Ph.D dan Dr. Ir. Setiawan, M.S. selaku dosen penguji yang telah memberikan masukan dan bantuan dalam menyelesaikan Tugas Akhir.
6. Seluruh dosen Statistika ITS yang telah memberikan ilmu dan pengetahuan yang tak ternilai harganya, serta segenap karyawan Departemen Statistika ITS.
7. Thalia, Abid, Icha Tirhiss, Fitria yang telah membantu dalam memberikan saran dan arahan dalam mengerjakan Tugas Akhir agar senantiasa lancar.
8. Partner terdekat Friska yang selalu memberikan semangat dan menjadi tempat curhat sekaligus keluh kesah penulis selama proses pengerjaan Tugas Akhir.

9. Sahabat seperjuangan yaitu Rachel, Inan, Bima, Widya yang selalu memberikan support dan semangat serta membantu penulis dalam proses penyusunan Tugas Akhir.
10. Teman-teman TR16GER yang selalu memberikan semangat kepada penulis dalam penyusunan Tugas Akhir dan sudah memberikan banyak pelajaran kepada penulis.
11. Seluruh pihak yang turut membantu dalam penyelesaian laporan Tugas Akhir ini baik secara langsung maupun tidak langsung yang tidak bisa penulis sebutkan namanya satu persatu.

Penulis menyadari masih banyak kekurangan dalam pembuatan laporan Tugas Akhir ini. Penulis berharap semoga laporan Tugas Akhir ini dapat bermanfaat dan menambah wawasan bagi pembaca. Kritik dan saran sangat diperlukan untuk perbaikan di masa yang akan datang.

Surabaya, Agustus 2020

Penulis

DAFTAR ISI

DAFTAR GAMBAR

(Halaman ini sengaja dikosongkan)

DAFTAR TABEL

(Halaman ini sengaja dikosongkan)

DAFTAR LAMPIRAN

(Halaman ini sengaja dikosongkan)

BAB I
PENDAHULUAN

1.1 Latar Belakang

Indonesia saat ini menjadi salah satu negara berkembang diantara negara-negara lain di dunia, hal tersebut ditunjukan berdasarkan kondisi perekonomian dan penduduk Indonesia. Ciri-ciri negara berkembang antara lain adalah rendahnya pendapatan perkapita dan pertumbuhan penduduk yang tinggi (Todaro, 1994). Indonesia mengalami peningkatan pendapatan perkapita dari tahun ke tahun, hal tersebut menunjukan bahwa Indonesia yang semakin lama semakin berkembang dengan baik. Dalam menumbuhkan perkembangan ekonomi di Indonesia membutuhkan modal. Modal yang digunakan untuk menciptakan pertumbuhan ekonomi Indonesia dapat bersumber dari dalam negeri dan bersumber dari luar negeri. Dari dalam negeri sendiri terdapat beberapa sumber pembiayaan (modal) salah satunya adalah tabungan dari masyarakat. Tabungan dari masyarakat memiliki potensi untuk andil dalam menumbuhkan perekonomian Indonesia, dimana modal tersebut harus diputar kembali atau dikembalikan lagi ke masyarakat yang membutuhkan modal untuk usaha atau keperluan lain.

Wewenang untuk mengatur atau menyalurkan dana kembali kepada masyarakat tersebut adalah wewenang dari lembaga keuangan. Menurut keputusan SK Menkeu RI no. 792 tahun 1990, menyatakan bahwa lembaga keuangan adalah seluruh badan usaha yang bergerak di bidang keuangan adalah menghimpun dana dan menyalurkan kembali kepada masyarakat atau nasabah untuk biaya investasi. Lembaga keuangan sendiri terdiri dari dua jenis, yaitu lembaga keuangan bukan bank dan lembaga keuangan bank. Lembaga keuangan memiliki tugas untuk mengatur, menghimpun dan menyalurkan dana dari masyarakat kemudian untuk disalurkan kembali kepada masyarakat. Pada dasarnya bank merupakan lembaga keuangan utama yang menyediakan fasilitas kepada nasabah untuk menyimpan uang.

Menurut Undang-Undang RI No. 10 tahun 1998 tentang perbankan, menyatakan bahwa bank merupakan sebuah badan usaha yang menghimpun dana dari masyarakat dalam bentuk simpanan dan

menyalurkan kepada masyarakat dalam bentuk kredit dana atau bentuk-bentuk lain dengan tujuan untuk meningkatkan taraf hidup orang banyak. Bank menjadi tempat atau lembaga yang dipercaya masyarakat untuk menyimpan dana dan menyalurkan dana tersebut kepada nasabah yang membutuhkan untuk modal. Perbankan merupakan sektor yang sangat penting dalam perekonomian nasional, dimana bank merupakan tempat terjadinya aliran dana yang mendukung untuk kegiatan ekonomi nasional. Bank memiliki produk produk yang disediakan untuk nasabah, menurut Sinungan (2000) bahwa bank memiliki produk-produk yang yaitu tabungan, giro, deposito dan kredit. Dari beberapa produk tersebut salah satu yang menyumbang pemasukan bank adalah dari dana kredit.

Kata kredit berasal dari bahasa yunani "Credere" yang berarti bahwa kepercayaan akan kebenaran. Menurut Undang-Undang Nomor 10 tahun 1998, kredit adalah penyediaan uang atau tagihan yang dapat dipersamakan dengan itu, berdasarkan persetujuan atau kesepakatan pinjam meminjam antara bank atau instansi keuangan dengan pihak lain yang mewajibkan pihak peminjam untuk melunasi utangnya setelah jangka waktu tertentu dengan jumlah bunga. Kredit atau pembiayaan merupakan dana yang dapat dipinjamkan kepada masyarakat yang membutuhkan untuk kebutuhan investasi atau modal usaha, dimana terdapat angunan atau jaminan yang harus diberikan kepada pihak bank untuk menjamin dana yang akan dipinjam sesuai dengan jaminan yang diberikan.

Kredit sebagai salah satu produk bank, tentunya juga memiliki masalah atau resiko yang mungkin bisa timbul dari nasabah terhadap bank. Resiko yang berpotensi terjadi adalah telat membayar angsuran. Telat bayar atau gagal bayar dapat diartikan ketidakmampuan seseorang atau instansi untuk membayar kewajiban keuangan yang telah jatuh tempo akibat tidak memiliki sumber daya keuangan yang memadai (Laitinen, 2006). Resiko telat bayar dari nasabah memungkinkan pihak bank untuk merugi, apabila kondisi dari nasabah banyak yang telat membayar angsuran maka memungkinkan bank akan tutup atau bangkrut. Perlu adanya analisa mengenai kemungkinan nasabah mengalami telat membayar angsuran pada

bank, hal ini diperlukan bank untuk mengetahui seberapa besar resiko nasabah mengalami telat membayar angsuran.

Penelitian mengenai resiko kredit pernah dilakukan oleh Belaid dkk., (2017). Penelitian tersebut menganalisis mengenai resiko kredit bank di Tunisia pada periode 2001-2012 yang berdampak terhadap intensitas dan hubungan pinjaman perusahan pada bank. Hasil penelitian menunjukan bahwa perusahaan yang memiliki hubungan yang baik dengan pihak bank, kecil kemungkinannya untuk mengalami telat membayar angsuran. Penelitian mengenai resiko kredit juga pernah dilakukan oleh Abdelmoula (2015). Berdasarkan penelitian tersebut, peneliti ingin melakukan prediksi mengenai gagalnya pengembalian dana peminjaman dari nasabah dalam jangka pendek di bank komersial Tunisia. Dalam penelitian tersebut menggunakan metode klasifikikasi *K-Nearest Neighbor* (KNN), dari hasil penelitian tersebut menunjukan bahwa nilai akurasi klasifikasi sebesar 88,63% dan kebaikan model berdasarkan nilai AUC adalah sebesar 87,4%. Penelitian mengenai analisis resiko kredit menggunakan metode *Random Forest* pernah dilakukan oleh Ghatasheh (2014) dimana penelitian tersebut membahas mengenai evaluasi ketepatan *machine learning* dalam memprediksi resiko kredit untuk analisa bisnis dengan difokuskan dengan metode *Random Forest*. Hasil penelitian menunjukan bahwa metode *Random Forest* merupakan metode yang menjanjikan untuk analisa bisnis dalam memprediksi resiko kredit. Keuntungan dari menggunakan metode *Random Forest* memiliki akurasi yang bagus dan sederhana sehingga memudahkan peneliti dalam menentukan keputusan.

Random forest merupakan metode *machine learning* yang populer digunakan dalam berbagai riset (Speiser dkk., 2019). Menurut Liu dkk., (2013), dalam penelitian tersebut membandingkan metode *Random Forest*, *Support Vector Machine* dan *Back Propagation Neural Network* untuk klasifikasi data lidah elektrik dalam pengenalan minuman jeruk dan cuka Cina. Berdasarkan hasil penelitian didapatkan bahwa tingkat kebaikan model untuk metode BPNN sebesara 86,68%, metode SVM sebesar 66,45% dan metode *Random Forest* sebesar 99,07%. Berdasarkan hasil tersebut bahwa metode *Random Forest* terbukti lebih baik dibandingkan metode BPNN dan SVM, juga memiliki beberapa keuntungan untuk kasus-kasus seperti

unbalanced data, multiclass, dan sampel data kecil. Menurut Ghatasheh (2014), penelitian tersebut melakukan prediksi status kredit individu dengan object data yang diteliti merupakan *The German Credit dataset* yang didapatkan dari (UCI) *Machine Learning Repository*, dimana penelitian ini bertujuan untuk melihat metode *machine learning* yang paling baik untuk prediksi resiko kredit dengan berfokus pada metode *Random Forest*. Hasil penelitian tersebut menunjukan bahwa metode *Random Forest* terbukti memiliki akurasi yang paling baik, dimana nilai akurasinya terbesar dibandingkan metode *machine learning* yang lain. Nilai akurasi yang didapatkan sebesar 0,784 dan nilai AUC sebesar 0,800. Selain metode *Random Forest*, metode Regresi logistik juga banyak digunakan untuk klasifikasi dalam berbagai penelitian.

Regresi logistik menjadi salah satu metode statistik yang paling banyak digunakan oleh ahli statistik dan peneliti, untuk analisis yang memiliki data respon kategorik (Hilbe, 2009). Yu dkk., (2010) melakukan perbandingan beberapa metode *machine learning* untuk mengevaluasi resiko kredit individu pada *German commercial personal credit dataset* dan *Australian personal credit dataset*, dimana metode metode yang dibandingkan adalah *Logistic Regresion*, *Decision Tree*, *Support Vector Machine* (SVM) dan *Neural Networks* (NN). Berdasarkan penelitian tersebut didapatkan bahwa *Logistic Regression* dan *Support Vector Machine* (SVM) menghasilkan akurasi klasifikasi terbaik dibandingkan metode yang lainnya. Sedangkan pada penelitian Kruppa dkk., (2013) melakukan prediksi resiko kredit pada perusahaan produksi alat rumah tangga yang menggunakan pembayaran secara kredit. Pada penelitian ini metode machine learning yang dibandingkan adalah *Logistic Regression, Random Forest*, kNN dan bNN. Berdasarkan hasil penelitian didapatkan bahwa metode yang terbaik adalah *Random Forest* pada data uji kredit angsuran jangka pendek. Berdasarkan penelitian-penelitian sebelumnya dan latar belakang diatas, maka akan dilakukan analisis klasifikasi nasabah telat bayar di Bank "x" dengan melakukan perbandingan metode *Random Forest* dan Regresi logistik berdasarkan nilai AUC. Nasabah yang sudah termasuk golongan "nasabah telat bayar" apabila nasabah mengalami mengalami telat membayar angsuran untuk yang kedua kalinya, jadi apabila nasabah

telat membayar angsuran untuk pertama kali, nasabah tersebut akan dikenakan sistem EWS yang dimiliki oleh Bank "x", kemudian apabila nasabah pada angsurannya selanjutnya telat membayar angsuran sudah bisa dikatakan "nasabah telat bayar".

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah diuraikan di atas, diperlukan adanya sebuah analisis untuk mencari metode yang baik untuk menganalisis kasus nasabah telat bayar di Bank "x", apabila kasus nasabah yang telat bayar ini banyak terjadi, maka bisa saja Bank "x" akan mengalami kebangkrutan. Selain itu, apabila dalam data terjadi kondisi *imbalance* atau ketidakseimbangan data pada variabel respon, dapat memberikan hasil klasifikasi yang tidak sesuai. Pada penelitian ini digunakan dua metode *machine learning*. Metode *machine learning* yang digunakan pada penelitian ini adalah *Random Forest* dan Regresi logistik biner. Dimana, apabila terdapat *imbalance* data dapat diatasi dengan metode SMOTE untuk mendapatkan hasil klasifikasi yang sesuai. Dari hasil analisis yang dilakukan, dapat diketahui hasil prediksi dari setiap metode, kemudian nanti dapat diketahui metode yang baik untuk kasus nasabah telat bayar di Bank "x", sehingga nantinya dari metode yang baik dapat dijadikan metode untuk menganalisis mengenai nasabah telat bayar di Bank "x" kedepannya.

1.3 Tujuan

Berdasarkan rumusan masalah, adapun tujuan yang ingin dicapai pada penelitian ini adalah sebagai berikut.

1. Memperoleh hasil deskripsi karakteristik nasabah telat bayar di Bank "x"
2. Mendapatkan hasil klasifikasi menggunakan metode *Random Forest* dan Regresi logistik biner data nasabah telat bayar di Bank "x"
3. Mendapatkan hasil klasifikasi menggunakan metode *Random Forest* dan Regresi logistik biner data nasabah telat bayar di Bank "x" dengan SMOTE.
4. Memperoleh metode terbaik untuk klasifikasi data nasabah telat bayar di Bank "x"

1.4 Manfaat Penelitian

Manfaat yang diharapkan pada penelitian ini adalah sebagai berikut :

1. Bagi Keilmuan Statistika
   Dapat digunakan untuk penelitian berikutnya mengenai analisa telat bayar dengan pendekatan metode statistika. Menjadikan wawasan dan pengetahuan mengenai metode *Random Forest* dan Regresi logistik biner
2. Bagi Bank "x"
   Metode terbaik yang didapatkan dari hasil perbandingan kedua metode, dapat menjadikan metode tersebut untuk menyelesaikan kasus klasifikasi mengenai nasabah telat bayar, dari metode tersebut juga dapat digunakan untuk prediksi nasabah telat bayar di waktu yang akan datang.

1.5 Batasan Masalah

Pada penelitian ini terdapat batasan masalah, dimana batasan masalah pada penelitian ini adalah data yang digunakan pada penelitian ini adalah data nasabah pembiayaan Bank "x" dari tahun 2015-2018. Pada metode Regresi logistik biner tidak dilakukan pencarian *best model*.

BAB II
TINJAUAN PUSTAKA

Tinjauan pustaka berisi landasan teori yang dipakai pada penelitian ini.Teori yang digunakan pada penelitian ini berasal dari buku, jurnal ilmiah, dan beberapa penelitian sebelumnya.

2.1 Statistika Deskriptif

Statistika deskriptif merupakan bagian statistika yang membahas tentang metode-metode untuk menyajikan data sehingga menarik dan informatif. Secara umum statistika deskriptif dapat diartikan sebagai metode-metode yang berkaitan dengan pengumpulan dan penyajian suatu gugus data sehingga memberikan informasi yang berguna. Perlu kiranya dimengerti bahwa statistika deskriptif memberikan informasi hanya mengenai data yang dipunyai dan sama sekali tidak menarik kesimpulan (Walpole, 2012).

Statistika deksriptif terbagi menjadi dua, yaitu ukuran pemusatan data dan ukuran penyebaran data. Terdapat berbagai macam penyajian statistika deskriptif melalui diagram, seperti *pie chart* dan *bar chart*. Selain menggunakan diagram, statistika deskriptif dapat disajikan dengan bentuk tabel dan grafik.

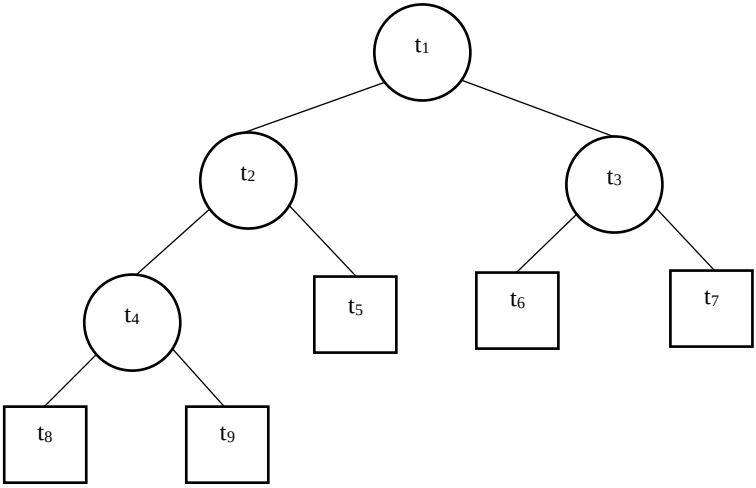
2.2 Klasifikasi

Klasifikasi dapat digunakan untuk membuktikan klasifikasi yang akurat atau untuk memprediksi tergantung dari kasus yang diteliti, klasifikasi juga dapat digunakan untuk mendapat variabel yang berpengaruh terhadap suatu kejadian (Breiman dkk., 1984). Dalam menggunakan metode *machine learning* perlu adanya pembagian data menjadi data *training* dan data *testing*, dimana data *training* untuk membangun model sedangkan data *testing* untuk validasi model.

2.3 *Classification and Regression Trees* (CART)

CART merupakan sebuah metode statistika non parametrik yang digunakan untuk melakukan analisis klasifikasi. Metode ini diperkenalkan oleh Breiman *et al*., (1984) terdiri dari dua analisis yaitu *classfication trees* dan *regression trees*. CART akan menghasilkan pohon klasifikasi (*classification trees*) apabila variabel respon yang

digunakan berskala kategorik dan akan menghasilkan pohon regresi (*regression trees*) jika variabel respon berskala kontinu. Konsep yang digunakan dalam metode ini dengan cara membuat sebuah pohon klasifikasi atau dikenal dengan istilah *Binary Recursive Partitioning*. Proses *binary* merupakan pemilahan data dimana data yang terkumpul di setiap simpul induk (*node*) akan mengalami pemecahan kedalam tepat dua simpul anak (*child node*). Sedangkan pada proses *recursive* merupakan sebuah proses dimana pemecahan tersebut diulang kembali di setiap simpul anak atau simpul dalam (*internal node*) kemudian akan membentuk simpul anak yang lain dan dapat diulang-ulang. Pemecahan ini akan dilakukan terus-menerus sampai tidak dapat dilakukan untuk pemecahan berikutnya atau sampai pada simpul terakhir (*terminal node*). *Partitioning* merupakan proses klasifikasi dapat dilakukan dengan cara memilah kumpulan data menjadi bagian atau partisi.



Gambar 2.1 Ilustrasi Struktur Pohon Klasifikasi (Sumber : Brieman dkk., 1984)

Pada gambar 2.1 ditunjukan ilustrasi pohon klasifikasi dimana pada gambar notasi t_1 mengandung seluruh data atau simpul induk. Notasi t_2, t_3, dan t_4 merupakan simpul dalam (*internal node*), sedangkan pada notasi t_5, t_6, t_7, t_8, dan t_9 dimana setelah itu tidak

terdapat pemecahan lagi. Setiap simpul berada pada kedalaman (*depth*) tertentu, dimulai dari simpul awal t_1 yang berada pada kedalam 1, t_2 dan t_3 berada pada kedalam 2, dan begitu seterusnya hingga dapat simpul t_4, t_5, t_6, t_8, dan t_9 yang berada di kedalaman 4.

Terdapat tiga tahap dalam melakukan klasifikasi dengan menggunakan metode CART. Pertama, membentuk pohon klasifikasi dengan prosedur pembentukan menggunakan pemecahan simpul secara berulang (*recursive*). Selanjutnya dilakukan pemangkasan pohon klasifikasi (*pruning*) yang menghasilkan rangkaian pohon klasifikasi yang lebih sederhana. Tahap terakhir, penentuan pohon klasifikasi optimal, dimana pohon klasifikasi tersebut dapat mempresentasikan informasi dari data namun tidak berlebihan (*overfitting*).

Pembentukan pohon klasifikasi terdapat tahap tahap yang harus dilakukan. Pertama menentukan variabel dan nilai yang layak dari variabel tersebut untuk dijadikan pemecah bagi setiap simpul. Proses yang dilakukan untuk membentuk pohon klasifikasi terdiri dari tiga tahap, yaitu pemilihan pemecah simpul, penentuan simpul terminal, dan pelabelan kelas (Breiman dkk., 1984).

a. Pemilihan pemecah simpul

Tahap pemilihan pemecah ini dilakukan pemecahan pada sampel data *training* berdasarkan kriteria dan *goodness of split* (kriteria pemecah uji terbaik). Pada data *training* yang digunakan masih bersifat heterogen. Pemilihan pemecah simpul tergantung pada jenis pohon atau pada jenis variabel respon. Hasil dari pemecahan harus menjadi lebih homogen dari simpul induknya. Nilai *impurity* atau $r(t)$ digunakan untuk mengukur tingkat keheterogenan dari simpul tersebut. Aturan pemecahan simpul induk menjadi dua simpul anak berdasarkan pada nilai yang berasal dari satu variabel prediktor. Setiap pemecahan hanya bergantung pada satu variabel prediktor saja, apabila variabel prediktornya merupakan variabel kontinu, maka pemecahan yang diperbolehkan adalah $x_n \leq y_m$ dan $x_n > y_m$ dengan $m = 1,2,3, \dots, n-1$ dengan y_m adalah nilai tengah atau median dari dua nilai amatan sampel yang berbeda dan berurutan. Sehingga jika terdapat sejumlah m sampel yang memiliki nilai

berbeda pada variabel x_m, maka terdapat pemecahan yang berbeda. Namun jika variabel prediktornya merupakan variabel kategorik, maka pemecahan berasal dari semua kemungkinan pemecahan berdasarkan terbentuknya dua simpul yang saling lepas (*disjoint*). Fungsi heterogenitas yang sering digunakan adalah indeks gini. Penggunaan indeks gini dalam pemilahan pemecah memiliki kelebihan, yaitu proses perhitungannya sederhana dan relatif cepat, serta mudah untuk diterapkan dalam berbagai kasus (Breiman dkk., 1984). Berikut ini merupakan fungsi indeks gini.

$$r(t) = \sum_{c_0}^{c_0} \sum_{c_1}^{c_1} p(c_0 \mid t)\, p(c_1 \mid t) = 1 - \sum_{i=0}^{1} (c_i)^2, c_0 \neq c_1 \qquad (2.1)$$

Keterangan:
$r(t)$: Indeks gini (fungsi heterogenitas) pada simpul *t*.
$p(c_0|t)$: Proporsi kelas 0 pada simpul *t*
$p(c_1|t)$: Proporsi kelas 1 pada simpul *t*

Langkah selanjutnya adalah menentukan pemecah terbaik dari setiap variabel prediktor. pemecah terbaik adalah pemecah yang memaksimalkan ukuran homogenitas setiap simpul anak terhadap simpul induknya dan juga memaksimalkan ukuran pemecahan antara dua simpul anak tersebut. Setiap pemecahan akan dilakukan pada setiap simpul sampai diperoleh simpul akhir.

Kemudian, pemecah terbaik pada masing-masing simpul induk berdasarkan nilai *goodness of split*. Berikut ini merupakan rumus untuk mencari nilai *goodness of split*.

$$\phi(s,t) = \Delta_1(s,t) = r(t) - p_L r(t_L) - p_R r(t_R) \qquad (2.2)$$

Keterangan:
$\phi(s,t)$: Nilai *goodness of split*.
$r(t)$: Fungsi heterogenitas pada simpul *t*.
p_L : Proporsi pengamatan simpul kiri.
p_R : Proporsi pengamatan simpul kanan.
$r(t_L)$: Fungsi heterogenitas pada simpul kiri.
$r(t_R)$: Fungsi heterogenitas pada simpul kanan

Nilai *goodness of split* yang tertinggi merupakan pemecah terbaik, karena dapat menghasilkan heterogenitas lebih tinggi. Setiap variabel akan menghasilkan skor, dimana skor tersebut menunjukan seberapa besar variabel tersebut memberikan kontribusi dalam proses pembentukan pohon klasifikasi. Untuk menentukan besarnya skor dari setiap variabel dapat digunakan persamaan sebagai berikut.

$$skor = \sum_{g=1}^{G} \phi(s,tg) \tag{2.3}$$

$\phi(s,tg)$ merupakan nilai *goodness of split* dari setiap simpul. Nilai skor didapatkan dengan menjumlahkan dari setiap nilai *goodness of split* dari masing-masing variabel.

b. Penentuan simpul terminal

Ketika pada simpul *t* tidak terdapat penurunan heterogenitas yang signifikan, atau hanya terdapat satu pengamatan di setiap simpul anak maka simpul dapat dikatakan simpul terminal, atau terdapat batasan minimum m pengamatan di setiap simpul anak yang dihasilkan (Breiman dkk., 1984).

c. Pelabelan kelas

Pelabelan kelas merupakan indentifikasi dari tiap simpul dengan suatu kelas tertentu. Pelabelan tiap simpul terminal berdasarkan aturan jumlah anggota kelas terbanyak yaitu label kelas simpul t adalah c_i.

$$p(c_i \mid t) = \max\, p(c_i \mid t) = \max \frac{M_{ci}}{M(t)} \tag{2.4}$$

Keterangan :

$p(c_i|t)$: Proporsi kelas c_i pada simpul *t*

M_{ci} : Jumlah pengamatan kelas c_i pada simpul *t*

$M(t)$: Jumlah pengamatan pada simpul t

Label kelas untuk simpul terminal t adalah ci yang memberikan nilai dugaan kesalahan pengklasifikasian pada simpul t yang paling kecil, yaitu sebesar $\mathrm{r(t)} = 1 - \max p(c_i|t)$.

Tahapan Selanjutnya adalah melakukan pemangkasan pohon klasifikasi atau yang biasa disebut dengan *prunning* perlu dilakukan karena semakin banyak pemilahan yang dilakukan mengakibatkan makin kecilnya tingkat kesalahan prediksi atau dengan kata lain nilai

prediksi melebihi nilai yang sebenarnya (*overfitting*). Pemangkasan pohon dilakukan dengan menentukan *cost complexity minimum* (Breiman dkk., 1984).

$$R_a(T) = R(T) + a\left|\tilde{T}\right| \tag{2.5}$$

Keterangan :

$R_a(T)$: Ukuran kompleksitas suatu pohon *T* pada komplesitas *a*

$R(T)$: Penduga pengganti (resubtitution estimate) pohon atau ukuran kesalahan klasifikasi pohon *T*

a : Parameter cost complexity bagi penambah satu simpul terminal pada pohon *T*

$\left|\tilde{T}\right|$: Banyak simpul terminal pada pohon *T*

Cost complexity pruning digunakan untuk menentukan pohon bagian T(a) yang dapat meminimumkan $R_a(T)$ pada pohon bagian atau setiap nilai a. Nilai parameter kompleksitas (a) akan secara perlahan meningkat selama proses pemangkasan. Selanjutnya, pencarian pohon bagian T(a)< T_{maks} yang dapat meminimumkan $R_a(T)$. Pemangkasan pohon dimulai dengan mengambil t_R dan t_L dari T_{maks} yang dihasilkan dari simpul induk t. Jika diperoleh dua simpul anak dari proses pemilahan yang dilakukan pada simpul induk yang memenuhi persyaratan $R(t) = R(t_R) + R(t_L)$, maka dua simpul anak akan dipangkass. Sehingga diperoleh pohon T_1 yang memnuhi kriteria $R(T_1) = R(T_{\text{maks}})$. Proses ini terus dilakukan secara berulang hingga tidak mungkin lagi dilakukan pemangkasan. Jika R(T) digunakan sebagai kriteria penentuan pohon klasifikasi optimal, maka nilai penduga pengganti akan cenderung memilih pohon besar T_1, karena semakin besar pohon, semakin kecil nilai penduga penggantinya. Hasil yang diperoleh dari tahap pemangkasan berupa urutan pohon yaitu $T_{maks} > T_1 > T_2 > \cdots > \{t_1\}$. Urutan pohon tersebut memiliki nilai a yang semakin menurun, yaitu $a_j < a_{j+1}$ dimana $a_1 = 0$ untuk $j \geq 1$ dan $T(a) = T(a_j) = T_j$.

2.4 Random Forest

Menurut Breiman (2001), metode *Random Forest* ialah gabungan dari pohon klasifikasi (CART) yang saling independen yang

berasal dari distribusi yang sama melalui proses *votting* (jumlah terbanyak) untuk memperoleh prediksi klasifikasi. *Random forest* merupakan pengembangan dari metode *ensemble* yang dikembangkan pertama kali oleh Leo Breiman (2001). Metode ini memiliki banyak pembuatan pohon sehingga dari pohon-pohon tersebut akan terbentuk suatu hutan (*forest*), kemudian analisis dilakukan pada kumpulan pohon tersebut. Pemilihan pemecah hanya melibatkan beberapa variabel prediktor, prosedur *Random Forest* yang dilakukan adalah

1. Tahap *bootstrap* dengan cara mengambil n data sampel dari dataset awal dengan pengembalian.
2. Menyusun pohon klasifikasi dari setiap dataset hasil *resampling bootstrap*, dengan penentuan pemilah terbaik didasarkan pada variabel prediktor yang diambil secara acak. Jumlah variabel yang diambil secara acak dapat ditentukan melalui perhitungan $\log_2(Z+1)$ dimana *Z* adalah banyaknya variabel prediktor (Brieman, 2001) atau $\sqrt{Z}$ (Genuer dkk, 2009)
3. Melakukkan prediksi klasifikasi data sampel berdasarkan pohon klasifikasi yang terbentuk
4. Mengulangi langkah 1 dan langkah 2 sejumlah *k* kali, sehingga terbentuk sebuah hutan yang terdiri dari *k* pohon.

2.5 Regresi Logistik Biner

Regresi logistik biner merupakan suatu metode analisis data yang berguna untuk mencari hubungan antara variabel respon (y) yang bersifat biner dengan variabel prediktor (x) yang bersifat politokomus (Hosmer & Lemeshow, 2000). Variabel respon yang terdiri dari dua kategori, yaitu sukses dan gagal yang dinotasikan dengan y =1 untuk sukses dan y =0 untuk gagal. Oleh karena itu variabel y mengikuti distribusi *Bernoulli* untuk setiap observasi tunggal. Fungsi probabilitas untuk setiap observasi dapat dituliskan pada persamaan (2.6)

$$f(y_i, \pi_i) = \pi_i^{y} (1-\pi_i)^{1-y_i}; y = 0,1 \qquad (2.6)$$

Dimana y merupakan variabel respon, Jika y = 0 maka $f(y)$ =1- π dan jika y =1 maka $f(y) = \pi$. Berikut merupakan merupakan fungsi Regresi logistik yang dapat dituliskan sebagai berikut.

$$f(z)=\frac{1}{1+e^{-z}} \text{ ekuivalen } f(y)=\frac{e^{z}}{1+e^{z}} \tag{2.7}$$

dimana $z=\beta_0+\beta_0 x_1+\cdots+\beta_p x_p$ dengan p adalah banyaknya variabel prediktor. Nilai z antara $-\infty$ dan $+\infty$ sehingga nilai *f*(z) terletak antara 0 dan 1 untuk setiap nilai *z* yang diberikan. Hal ini menunjukan bahwa model logistik menggambarkan probabilitas atau resiko dari suatu objek sehingga model regresi dapat dituliskan sebagai berikut.

$$\pi(x)=\frac{\exp(\beta_0+\beta_1 x_1+...+\beta_p x_p)}{1+\exp(\beta_0+\beta_1 x_1+...+\beta_p x_p)} \tag{2.8}$$

dimana β_p merupakan parameter ke-p yang diestimasi. Untuk mempermudah pendugaan parameter regresi maka model Regresi logistik pada persamaan (2.8) dapat ditransformasi logit $\pi(x)$ sehingga diperoleh persamaan (2.9)

$$g(x)=\ln\left(\frac{\pi(x)}{1-\pi(x)}\right)=\beta_0+\beta_1 x_1+...+\beta_p x_p \tag{2.9}$$

Model tersebut merupakan fungsi linier dari paramater-parameternya.

2.5.1 Estimasi Parameter

Estimasi parameter dilakukan dengan metode *maximum likelihood*. Metode ini memberikan landasan dalam mendekati nilai estimasi pada Regresi logistik. Setiap pengamatan pada Regresi logistik ini mengikuti distribusi *Bernoulli* sehingga fungsi *likelihood* dapat ditentukan. Jika variabel respon kategori 0 dan 1 dimana untuk variabel $y = 1$ maka fungsi *likelihood* adalah $\pi(x_i)$ dan untuk variabel $y = 0$ maka fungsi likelihood adalah $1-\pi(x_i)$. Fungsi probabilitas untuk setiap pasangan (x_i, y_i) dapat dituliskan pada persamaan berikut.

$$\pi(x_i)^{y_i}(1-\pi(x_i))^{1-y_i} \tag{2.10}$$

Setiap pasangan pengamatan diasumsikan saling bebas sehingga fungsi *likelihood* merupakan gabungan dari fungsi distribusi masing-masing pasangan dan dapat dituliskan pada persamaan

$$l(\boldsymbol{\beta}) = \prod_{i=1}^{n} \pi(x_i)^{y_i} [1-\pi(x_i)]^{1-y_i} \tag{2.11}$$

Fungsi *maximum likelihood* yang digunakan untuk mengstimasi $\boldsymbol{\beta}$ pada persamaan (2.11) akan lebih mudah di maksimumkan dalam bentuk persamaan. $\boldsymbol{\beta}$ merupakan vector yang berisi nilai β. Persamaan untuk *log likelihood* dapat dituliskan pada persamaan (2.12)

$$L(\boldsymbol{\beta}) = \ln\left[l(\boldsymbol{\beta})\right] = \sum_{i=1}^{n} \left\{ y_i \ln\left[\pi(x_i)\right] + (1-y_i) \ln\left[1-\pi(x_i)\right] \right\} \tag{2.12}$$

Nilai $\boldsymbol{\beta}$ didapatkan dari hasil diferensial $L(\boldsymbol{\beta})$ terhadap β_j dan hasilnya adalah sama dengan nol.

$$\frac{\partial L(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{n} \sum_{j=0}^{p} y_i x_{ij} - \sum_{i=1}^{n} \frac{e^{\sum_{j=0}^{p} \beta_j x_{ij}}}{1+e^{\sum_{j=0}^{p} \beta_j x_{ij}}} \left(\sum_{j=0}^{p} x_{ij} \right) \tag{2.13}$$

sehingga,

$$\sum_{i=1}^{n} \sum_{j=0}^{p} y_i x_{ij} - \sum_{i=1}^{n} \pi\left(\mathbf{x}_i\right) \left(\sum_{j=0}^{p} x_{ij} \right) = 0 \tag{2.14}$$

Estimasi varians kovarians dikembangkan melalui teori MLE dari koefisien parameternya yang didapatkan dari turunan kedua $l(\boldsymbol{\beta})$.

$$\frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_j \partial \beta_j'} = \sum_{i=1}^{n} \sum_{j=0}^{p} x_{ij} x_{ij}' \pi\left(\mathbf{x}_i\right)\left(1-\pi\left(\mathbf{x}_i\right)\right) \tag{2.15}$$

Nilai taksiran $\boldsymbol{\beta}$ dari turunan pertama fungsi $L(\boldsymbol{\beta})$ yang nonlinier didapatkan dengan menggunakan metode iterasi Newton Raphson dengan rumus sebagai berikut:

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} - \left(\mathbf{H}\left(\boldsymbol{\beta}^{(t)}\right)\right)^{-1} \mathbf{g}(\boldsymbol{\beta})^{(t)}, \; t = 1, 2, \ldots, n. \tag{2.16}$$

Iterasi dilakukan sampai konvergen ke-n. $\mathbf{H}\left(\boldsymbol{\beta}^{(t)}\right)$ merupakan matriks Hessian dengan

$$\mathbf{H}\left(\boldsymbol{\beta}^{(t)}\right)=\begin{bmatrix} h_{00} & h_{01} & \cdots & h_{0p} \\ h_{10} & h_{11} & \cdots & h_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ h_{p0} & h_{p1} & \cdots & h_{pp} \end{bmatrix},\ h_{ij}=\frac{\partial^2 L(\boldsymbol{\beta})}{\partial\beta_j\partial\beta_j'} \qquad (2.17)$$

dan $\mathbf{g}(\boldsymbol{\beta})^{(t)}$ merupakan vektor gradient dimana

$$\mathbf{g}(\boldsymbol{\beta})^{(t)}=\left[\frac{\partial L(\boldsymbol{\beta})}{\partial\beta_0},\frac{\partial L(\boldsymbol{\beta})}{\partial\beta_1},...,\frac{\partial L(\boldsymbol{\beta})}{\partial\beta_p}\right] \qquad (2.18)$$

Matrix Hessian dengan iterasi sebanyak *n* dapat ditulis menjadi

$$\mathrm{H}\left(\beta^{(t)}\right)=-\left\{\mathbf{x}^T diag\left[\pi\left(\mathbf{x}_1\right)^{(t)}\left(1-\left(\mathbf{x}_1\right)^{(t)}\right),...,\pi\left(\mathbf{x}_n\right)^{(t)}\left(1-\left(\mathbf{x}_n\right)^{(t)}\right)\right]\mathbf{x}\right\}^{-1} \qquad (2.19)$$

Sehingga didapatkan estimasi sebagai berikut :

$$\hat{\boldsymbol{\beta}}^{(t+1)}=\hat{\boldsymbol{\beta}}^{(t)}+\left\{\mathbf{x}^T diag\left[\pi\left(\mathbf{x}_1\right)^{(t)}\left(1-\left(\mathbf{x}_1\right)^{(t)}\right),...,\pi\left(\mathbf{x}_n\right)^{(t)}\left(1-\left(\mathbf{x}_n\right)^{(t)}\right)\right]\mathbf{x}\right\}^{-1}$$
$$\mathbf{x}^T\left(y-\pi\left(\mathbf{x}_n\right)^{(t)}\right) \qquad (2.20)$$

2.5.2 Uji Serentak

Uji serentak dilakukan untuk mengetahui signifikansi parameter **β** terhadap variabel respon. Uji serentak pada Regresi logistik biner memiliki hipotesis sebagai berikut.

H_0 : $\beta_1=\beta_2=...=\beta_p=0$ (tidak ada pengaruh antara variabel prediktor terhadap variabel respon)

H_1 : minimal ada satu $\beta_p\neq 0$ (terdapat pengaruh antara variabel prediktor terhadap variabel respon)

Statistik uji yang digunakan adalah G, dimana statistik uji G tersebut mengikuti distribusi *Chi-Square* (Hosmer & Lemeshow, 2000). H_0 akan ditolak jika nilai statistik uji G lebih dari sama dengan $\chi^2_{(p,\alpha)}$ nilai dengan tingkat kepercayaan α. Rumus statistik uji G dapat dituliskan pada persamaan (2.21).

$$G = -2\ln\left[\frac{\left(\frac{n_1}{n}\right)^{n_1}\left(\frac{n_0}{n}\right)^{n_0}}{\prod_{i=1}^{n}\hat{\pi}_i^{y_i}\left(1-\hat{\pi}_i\right)^{1-y_i}}\right], \tag{2.21}$$

dimana $\hat{\pi}_i$ merupakan peluang kejadian sukses dengan nilai probabilitas antara 0 sampai 1, n_1 merupakan jumlah data dari variabel respon yang berkategori 1, n_0 merupakan jumlah data dari variabel respon yang berkategori 0 serta n aalah jumlah keseluruhan data.

2.5.3 Uji Parsial

Pengujian secara parsial dilandaskan pada hipotesis adalah sebagai berikut.

H_0 : $\beta_j = 0$ (tidak terdapat pengaruh antara masing-masing variabel prediktor terhadap variabel respon)

H_1 : $\beta_j \neq 0$ (terdapat pengaruh antara masing-masing variabel prediktor terhadap variabel respon)

H_0 ditolak jika p-value yang kurang dari taraf signifikansi. Rumus statistik uji Wald dapat dituliskan pada persamaan (2.22) (Hosmer & Lemeshow, 2000).

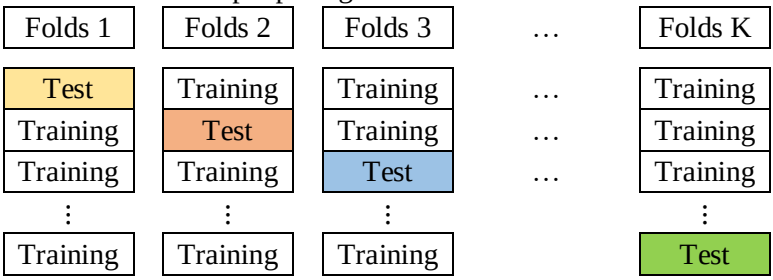
$$W = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)} \tag{2.22}$$

Dimana $\hat{\beta}_j$ merupakan koefisien parameter yang ke-j serta $SE(\hat{\beta}_j)$ adalah standar error dari koefisien parameter yang ke-*j*.

2.6 *K-Fold Cross Validation*

K-fold cross validation merupakan salah satu metode yang digunakan untuk mempartisi data menjadi data *training* dan data *testing*. Peneliti banyak menggunakan metode ini karena dapat mengurangi bias yang terjadi dalam pengambilan sampel. *K-fold cross validation* membagi data *training* dan *testing* dilakukan secara

berulang-ulang, dimana setiap data mendapat kesempatan menjadi data testing (Gokgoz & Subasi, 2015). K merupakan besar angka partisi data yang digunakan untuk pembagian *training* dan *testing*. Berikut merupakan ilustrasi pembagian data menggunakan *k-fold cross validation* terdapat pada gambar 2.2.



Gambar 2.2 Ilustrasi Pembagian Data

2.7 *Synthetic Minority Oversampling Technique* (SMOTE)

Synthetic Minority Over-sampling Technique (SMOTE) merupakan salah satu cara mengatasi class imbalance yang diusulkan oleh Chawla dkk., (2002). Konsep dari SMOTE adalah melakukan *oversampling* pada *minority class* dengan membuat contoh atau data *synthetic* dibanding melakukannya dengan perulangan (Chawla, dkk., 2002). SMOTE menambah data buatan dengan *k-nearest neighbor*, sehingga jumlah kelas minor setara dengan kelas mayor. SMOTE-N merupakan pengembangan dari SMOTE yang awalnya hanya dapat digunakan pada data numerik. SMOTE-N digunakan untuk melakukan oversampling pada data dengan kategori nominal.

Jika pada SMOTE, untuk menentukan *j*-data terdekat digunakan jarak *euclidean*, sedangkan pada SMOTE-N jarak terdekat dihitung menggunakan versi modifikasi dari *Value Difference Metric* yang disebut MVDM (Cost & Salzberg, 1993).

$$\delta(V_1, V_2) = \sum_{i=1}^{h} \left| \frac{C_{1i}}{C_1} - \frac{C_{2i}}{C_2} \right| \tag{2.23}$$

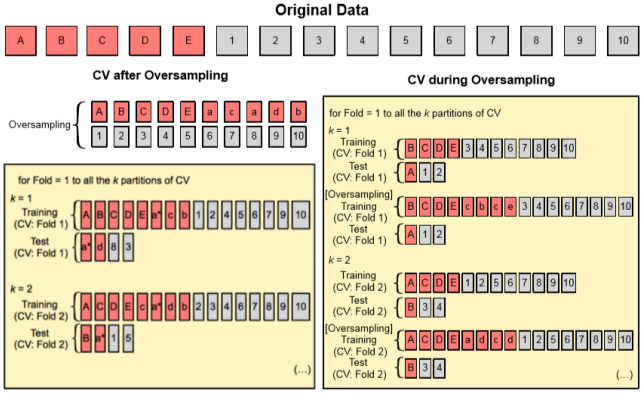
keterangan:
$\delta(V_1, V_1)$ = Jarak antar kategori V_1 dan V_2

C_1 = Banyaknya V_1 terjadi
C_2 = Banyaknya V_2 terjadi
C_{1i} = Banyaknya V_2 terjadi
C_{2i} = Banyaknya V_2 terjadi
h = Jumlah kelas pada variabel respon
jarak antar data dihitung menggunakan persamaan berikut.

$$\Delta(X,Y) = w_x, w_y \sum_{b=1}^{p} \delta(x_b, y_b)^r \qquad (2.24)$$

$\Delta(X,Y)$ =jarak antar observasi X dan Y
w_x, w_y = bobot (dapat diabaikan)
p = banyaknya variabel independen
$\delta(x_b, y_b)$ = jarak antara kategori x dan y pada variabel independen ke-b
r = 1 (manhatan) atau 2 (euclidean)

SMOTE akan diterapkan di prosedur *K-Fold Cross Validation* pada masing-masing data training di setiap fold. Jika *K-Fold CrossValidation* dilakukan bersamaan dengan SMOTE, hanya pola data training yang dipertimbangkan untuk menghasilkan pola baru dan model dari data training, sehingga dapat menghindari *overoptimistic* (Santos, dkk., 2018). Ilustrasi penerapan SMOTE pada *K-Fold Cross Validation* ditunjukkan pada Gambar 2.3



Gambar 2.3 Ilustrasi SMOTE dalam KCV

2.8 Ukuran Ketepatan Klasifikasi

Kemampuan prediksi dari algoritma klasifikasi biasanya diukur dengan akurasi prediksinya. Ketepatan klasifikasi digunakan untuk mengukur kebaikan dari model dalam memprediksi kelas berdasarkan kelas data aktual. *Area Under ROC Curve* (AUC) adalah ukuran ketepatan klasifikasi yang konsisten secara statistik dimana ukuran yang lebih baik daripada akurasi (Huang & Ling, 2005). Nilai AUC dihitung dengan menggunakan rata-rata perkiraan bidang berbentuk kurva yang dibentuk oleh TP_{rate} dan FP_{rate} (Dubey dkk., 2014). Nilai TP_{rate} dan FP_{rate} didapatkan dari *confusion matrix* yang dapat dilihat pada Tabel 2.1.

Tabel 2.1 Confussion Matrix Klasifikasi

Aktual	Prediksi	
	Positif= kelas 0	Negatif = kelas 1
Positif= kelas 0	*True Positif* (TP)	*False Negative* (FN)
Negatif = kelas 1	*False Positif* (FP)	*True Negative* (TN)

dengan

TP : Jumlah anggota kelas 0 yang diprediksi dengan benar
FP : Jumlah anggota kelas 0 yang diprediksi dengan salah
FN : Jumlah anggota kelas 1 yang diprediksi dengan salah
TN : Jumlah anggota kelas 1 yang diprediksi dengan benar

$$Akurasi = \frac{TP + TN}{TP + FN + FP + TN}. \tag{2.25}$$

$$TP_{rate} = Sensitivity = \frac{TP}{TP + FN}. \tag{2.26}$$

$$Specificity = \frac{TN}{TN + FP}. \tag{2.27}$$

$$FP_{rate} = 1 - Specificity. \tag{2.28}$$

$$AUC = \frac{1 + TP_{rate} - FP_{rate}}{2}. \tag{2.29}$$

Nilai maksimum AUC adalah sebesar 1 dimana dapat dikatakan model klasifikasi dapat memprediksi data dengan sempurna. Sementara itu apabila nilai AUC yang diperoleh adalah 0,5 menunjukkan bahwa model klasifikasi merupakan model acak tanpa kekuatan diskriminatif untuk memisahkan data.

2.9 *Early Warning System* (EWS)

Early warning system (EWS) adalah suatu mekanisme atau sistem deteksi atau pengenalan terhadap tanda-tanda atau gejala awal yang diperkirakan dapat mempengaruhi kondisi debitur. *Early warning system* dilakukan untuk meminimalisir dan mencegah terjadinya kerugian akibat kredit macet, maka bank harus menerapkan suatu sistem yang efektif dan berkesinambungan untuk memonitoring fasilitas kredit yang telah diberikan.

Manfaat bank melaksanakan EWS adalah

1. Manfaat bagi bank
    a. Penurunan *non performing loan*
    b. Mengambil tindakan korektif lebih awal
    c. Dapat menurunkan kewajiban pembentukan biaya Penyisihan penghapusan aktiva produktif (PPAP)
    d. Memelihara dan meningkatkan kredit lancar
    e. Meminimalisir kredit macet
2. Manfaat bagi debitur
    a. Debitur memiliki waktu yang lebih banyak untuk membayarkan kewajiban setiap bulan
    b. Meminimalisir terjadinya gagal bayar

EWS merupakan bentuk aktualisasi dari pasal 2 undang-undang perbankan, yaitu mengenai penerapan *Self Regulatory Banking* (SRB). *Self regulatory banking* merupakan kebijakan masing-masing bank untuk diterapkan kepada nasabahnya. EWS yang dilakukan oleh Bank "x" sebagai salah satu rangkaian untuk melakukan pencegahan atau meminimalisir terjadinya kredit macet. Bentuk tindakan EWS yang dilakukan oleh Bank "x" adalah yang pertama melakukan pemblokiran saldo rekening nasabah sebesar 1x angsuran kredit dari setiap nasabah. Kebijakan ini diterapkan kepada seluruh nasabah kredit dari bank tersebut. Tindakan pemblokiran saldo dengan jumlah 1x angsuran akan memberikan akibat memperpanjang jangka waktu untuk melakukan pembayaran kembali atas kredit yang diterima oleh nasabah. Tetapi, apabila nasabah tetap terlambat membayar angsuran untuk kedua kalinya setelah dikenakan tindakan EWS, maka nasabah dikategorikan nasabah "telat bayar".

2.10 Kerangka Konsep Variabel

Salah satu dasaran dalam penggunaan variabel penelitian ini yaitu didasarkan pada penelitian terdahulu. Penelitian terkait resiko kredit disajikan pada Tabel 2.2 berikut.

Tabel 2.2 Kerangka Konsep Variabel

Penulis	Judul Penelitian	Variabel Prediktor yang digunakan
Solvi M. Makandolu & Johanes G. Sogen (2015)	Analisis Faktor-Faktor yang Mempengaruhi Tingkat Pengembalian Kredit Mikro Utama (KMU)	- Umur - Pendidikan - Tanggungan Keluarga - Pengalaman Usaha - Aset - Omset - Pendapatan Usaha - Setoran Pokok - Setoran Bunga - Setoran Tunggakan - Status diri - Jangka Waktu
Nazeeh Ghataseheh(2014)	Business Analytics using *Random Forest* Trees for Credit Risk Prediction: A Comparison Study	- Jumlah tabungan di rekening - Umur - Pekerjaan - Jenis Kelamin - Jumlah kredit - Penggunaan kredit - Riwayat kredit - Jaminan - Tipe rumah - Jangka waktu

Tabel 2.2 Kerangka Konsep Variabel (lanjutan)

Raden Yogi Arrieffiandi dkk(2016)	Faktor-Faktor yang Mempengaruhi Tingkat Kolektabilitas Pembiayaan sektor Umkm (studi kasus : Bank syariah XYZ kantor cabang Jakarta Barat)	- Jenis pembiayaan - Tingkat bagi hasil - Jangka waktu - Jenis kelamin - Jenis usaha - Kepemilikan perusahaan

Berdasarkan penelitian terdahulu didapatkan variabel-variabel prediktor yang akan digunakan pada penelitian ini yaitu, jenis kelamin, jaminan, jangka waktu, jumlah pembiayaan, angsuran per bulan, sisa angsuran dan EWS.

(Halaman ini sengaja dikosongkan)

BAB III
METODOLOGI PENELITIAN

3.1 Sumber Data

Data yang digunakan pada penelitian ini adalah data sekunder, dimana data diperoleh dari Bank "x" mengenai pembiayaan nasabah. Data nasabah pembiayaan yang digunakan pada penelitian ini adalah data pembukuan pada tahun 2015-2018. Dataset dimana terdiri dari 1137 data nasabah Bank "x".

3.2 Variabel Penelitian

Variabel penelitian yang digunakan dalam penelitian ini ditunjukan pada tabel 3.2

Tabel 3. 1 Variabel Penelitian

Variabel	Keterangan	Deskripsi	Skala Data
y	Status Nasabah	Status keadaan pembayaran angsuran	Nominal 0: Tidak telat bayar 1 : Telat Bayar
x_1	Jenis Kelamin	Jenis Kelamin Nasabah	Nominal 0 : Perempuan 1: Laki-laki
x_2	Angunan/Jaminan	Jaminan yang digunakan untuk melakukan kredit	Nominal 1 : Emas 2 : Sertifikat Tanah 3 : BPKB 4 : BPJS 5 : Deposito
x_3	Jangka Waktu	Lama waktu nasabah mulai meminjam hingga waktu jatuh tempo	Nominal 1 : < 60 bulan 2 : < 120 bulan 3 : ≥ 120 bulan

Tabel 3.1 Variabel Penelitian (Lanjutan)

x_4	Jumlah Pembiayaan	Jumlah uang yang dipinjam oleh nasabah	Nominal 1 : < 100 juta 2 : < 200 juta 3 : < 300 juta 4 : < 400 juta 5 : ≥ 400 juta
x_5	Angsuran per bulan	Angsuran yang harus dibayarkan perbulan oleh nasabah	Nominal 1 : < 564 ribu 2 : < 3,88 juta 3 : ≥ 3,88 juta
x_6	EWS	Pemblokiran dana pinjam dari bank x	Nominal 0 : No 1 : Yes
x_7	Sisa Angsuran	Sisa angsuran yang harus dibayarkan	Nominal 1 : < 100 juta 2 : < 200 juta 3 : < 300 juta 4 : < 400 juta 5 : ≥ 400 juta

3.4 Struktur Data

Struktur Data secara keseluruhan yang digunakan dalam penelitian ini disajikan pada tabel 3.3 berikut.

Tabel 3.2 Struktur Data Penelitian

Nasabah ke-	$\boldsymbol{x_1}$	$\boldsymbol{x_2}$	$\boldsymbol{x_3}$	$\boldsymbol{x_4}$	$\boldsymbol{x_5}$	$\boldsymbol{x_6}$	$\boldsymbol{x_7}$	$\boldsymbol{y}$
1	$x_{1,1}$	$x_{2,1}$	$x_{3,1}$	$x_{3,1}$	$x_{5,1}$	$x_{6,1}$	$x_{7,1}$	y_1
2	$x_{1,2}$	$x_{2,2}$	$x_{3,2}$	$x_{3,2}$	$x_{5,2}$	$x_{6,2}$	$x_{7,2}$	y_2
3	$x_{1,3}$	$x_{2,3}$	$x_{3,3}$	$x_{3,3}$	$x_{5,3}$	$x_{6,3}$	$x_{7,3}$	y_3
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1137	$x_{1,1137}$	$x_{2,1137}$	$x_{3,1137}$	$x_{4,1137}$	$x_{5,1137}$	$x_{6,1137}$	$x_{7,1137}$	y_{1137}

3.5 Langkah Analisis

Langkah Analisis yang digunakan dalam penelitian ini antara lain adalah sebagai berikut.

a. Langkah analisis karakteristik data
   1. Melakukan pengumpulan data sekunder
   2. Melakukan *preprocessing* data
   3. Menganalisis karakteristik data pembiayaan nasabah di Bank "x" (sub bab 2.1).
      - Membuat *bar chart* atau *pie chart* di variabel prediktor
      - Membuat *pie chart* pada variabel respon untuk mengetahui data *imbalance* atau *balance*.

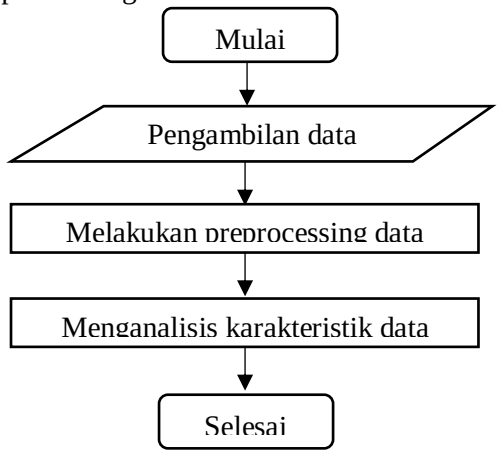
b. Langkah analisis hasil klasifikasi tanpa SMOTE
   1. Membagi data menjadi data training dan testing menggunakan K-fold cross validation dengan K=2-10 (sub bab 2.6).
   2. Melakukan analisis klasifikasi nasabah telat bayar menggunakan metode *Random Forest* dengan k sebanyak 1000 (sub bab 2.4).
   3. Melakukan analisis klasifikasi nasabah telat bayar menggunakan metode Regresi logistik biner (sub bab 2.5).

c. Langkah analisis hasil klasifikasi menggunakan SMOTE
   1. Membagi data menjadi data training dan testing menggunakan K-fold cross validation dengan K=2-10 (sub bab 2.6).
   2. Melakukan teknik SMOTE pada data *training* (sub bab 2.7).
   3. Melakukan analisis klasifikasi nasabah telat bayar menggunakan metode *Random Forest* dengan k sebanyak 1000 (sub bab 2.4).
   4. Melakukan analisis klasifikasi nasabah telat bayar menggunakan metode Regresi logistik biner (sub bab 2.5).

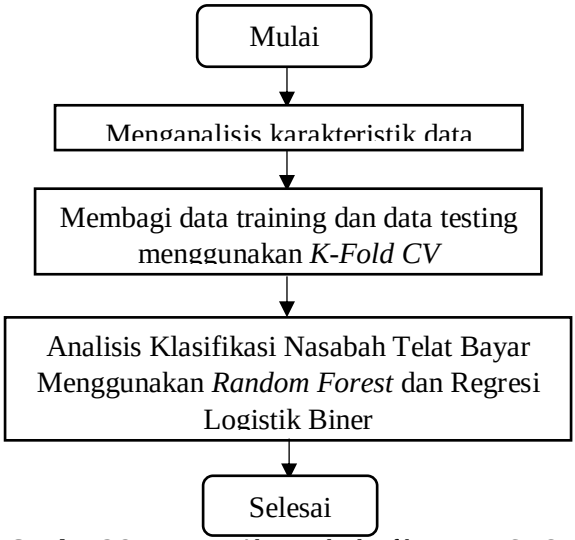
d. Langkah analisis menentukan metode terbaik
   1. Menghitung nilai performasi masing-masing metode (sub bab 2.8).

2. Membandingkan kedua metode berdasarkan nilai AUC untuk mengetahui metode terbaik
3. Menarik kesimpulan dan saran.

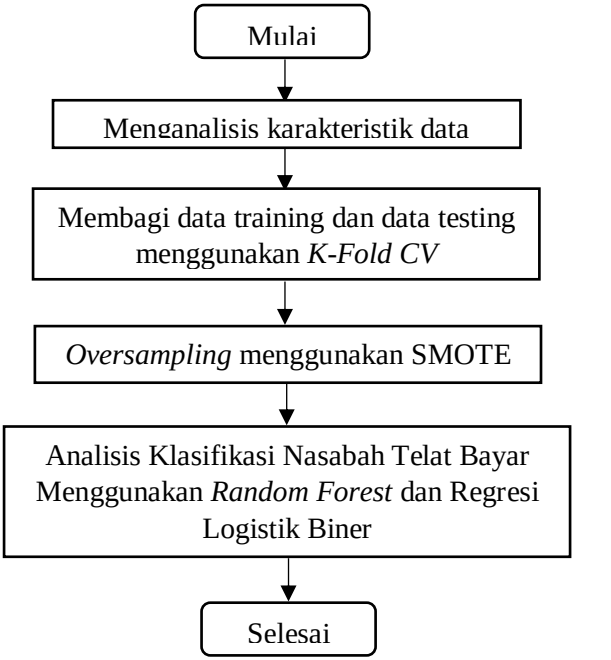
Berdasarkan langkah analisis diatas yang telah dijelaskan, dapat digambarkan pada 4 diagram alir dibawah.



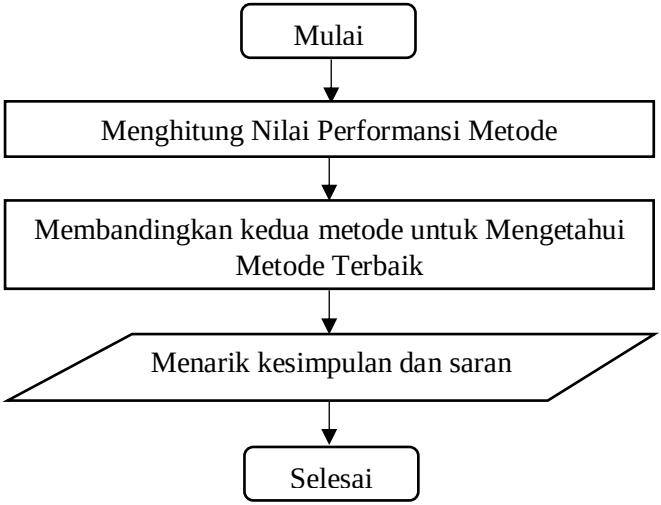
Gambar 3.1 Diagram Alir Karakteristik Data



Gambar 3.2 Diagaram Alir Hasil Klasifikasi tanpa SMOTE



Gambar 3.3 Diagram Alir Hasil Klasifikasi menggunakan SMOTE



Gambar 3.4 Diagram Alir Metode Terbaik

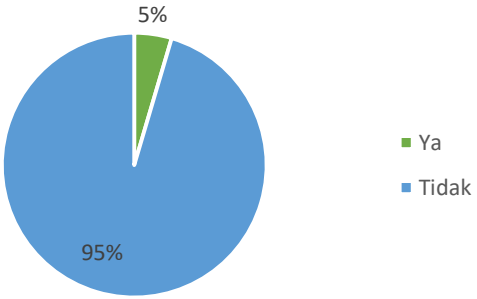
(Halaman ini sengaja dikosongkan)

BAB IV
ANALISIS DAN PEMBAHASAN

Analisis dan pembahasan menyajikan hasil dari output dari proses yang telah dilakukan, dimana output tersebut menjawab tujuan penelitian. Pembahasan pada penelitian adalah membandingkan dua metode *machine learning* yakni *Random Forest* dan Regresi logistik biner. Kebaikan hasil klasifikasi tersebut didapatkan dari hasil nilai akurasi dan AUC.

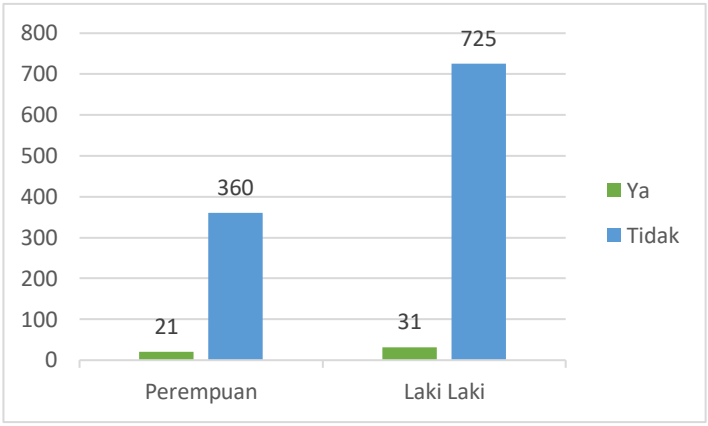
4.1 Karakteristik Data

Eksplorasi data bertujuan untuk melihat karakteristik dari data yang digunakan atau untuk mendapatkan gambaran umum sebagai informasi awal dari sebuah data sebelum menentukan atau menerapkan metode analisis yang akan digunakan. Eksplorasi data awal dilakukan pada beberapa variabel awal seperti variabel respon dan jenis kelamin nasabah.



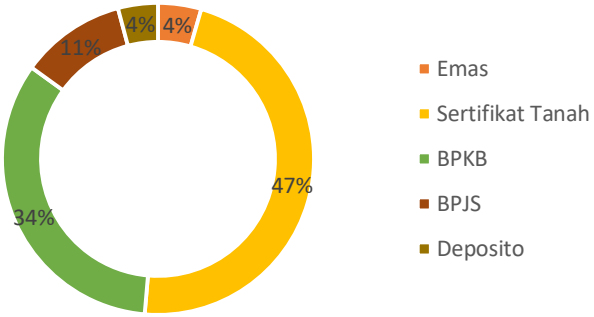
Gambar 4.1 Persentase Nasabah Telat Bayar

Pada data pembiayaan nasabah tahun 2015-2018, nasabah yang mengalami telat bayar atau tidak digambarkan pada gambar 4.1. Dari gambar tersebut dapat dikatakan bahwa secara umum nasabah yang mengalami telat bayar lebih kecil daripada yang tidak telat bayar, yakni 5% dibanding 95%. Maka dapat dikatakan bahwa data pembiayaan nasabah tahun 2015-2018 merupakan data yang *imbalance*. Hal ini akan menjadi masalah apabila dilakukan analisis klasifikasi, karena akan menghasilkan ouput yang tidak sesuai, sehingga perlu adanya penanganan *imbalance*.



Gambar 4.2 Histogram Jenis Kelamin Nasabah

Gambar 4.2 menunjukan bahwa nasabah yang melakukan peminjaman dana terhadap Bank "x" paling banyak adalah berjenis kelamin laki-laki dengan jumlah sebanyak 756 nasabah dan jumlah nasabah perempuan sebanyak 381 nasabah. Dimana masing-masing jenis kelamin laki-laki dan perempuan terdapat yang mengalami telat bayar, dimana nasabah dengan jenis kelamin perempuan sebanyak 21 nasabah dan dengan jenis kelamin laki-laki sebanyak 31 nasabah.



Gambar 4.3 Persentase Jaminan yang digunakan Nasabah

Berdasarkan gambar 4.3 dapat diketahui bahwa sebagian besar nasabah yang melakukan pinjaman di Bank "x" menggunakan jaminan sertifikat tanah, hal itu ditunjukan dengan besarnya persentase sebesar 47%. Sedangkan jaminan yang paling sedikit

digunakan oleh nasabah adalah jaminan emas dan jaminan deposito dengan persentase sebesar 4%.

4.2 Klasifikasi Data Nasabah Telat Bayar

Setelah mengetahui karakteristik data pembiayaan nasabah, dimana berdasarkan karakteristik data bahwa data pembiayaan nasabah merupakan data yang *imbalance*. Pertama, data yang digunakan merupakan data yang *imbalance*. Sebelum melakukan analisis perlu dilakukan pembagian data *training* dan *testing* menggunakan metode KCV dengan nilai K sebesar 2 hingga 10. Kemudian dilakukan analisis klasifikasi nasabah telat bayar menggunakan metode *Random Forest* dan Regresi logistik biner. Berikut pembahasan lebih lanjut untuk masing-masing metode yang digunakan dalam penelitian ini.

4.2.1 Analisis Klasifikasi Nasabah Telat Bayar menggunakan Metode *Random Forest*

Metode klasifikasi yang pertama menggunakan metode klasifikasi *Random Forest*. Sebelum dilakukan analisis menggunakan metode *Random Forest*, perlu dilakukan analisis CART, karena metode CART sendiri merupakan metode yang mendasari dari analisis *Random Forest*. Pada analisis CART digunakan data sebanyak 1137 data nasabah, dengan pembagian data *training* dan *testing* menggunakan KCV dengan nilai *K* dari 2 hingga 10

Terdapat tiga tahapan dalam metode klasifikasi CART, yaitu pembentukan pohon, pemangkasan pohon, hingga penentuan pohon klasifikasi yang optimal. Berikut merupakan ilustrasi penggunaan metode CART dengan nilai K=2.

Pada pembentukan pohon klasifikasi, diperlukan variabel-variabel yang berperan sebagai pemilah. Jika variabel berskala nominal bertaraf L, maka akan diperoleh pemilah sebanyak 2^{L-1}-1, sedangkan variabel berskala ordinal akan diperoleh pemilah sebanyak L-1. Jumlah kemungkinan pemilah untuk membentuk pohon yang berbeda pada pohon klasifikasi nasabah telat bayar ditampilkan pada tabel 4.1.

Tabel 4.1 Jumlah Kemungkinan Pemilah Variabel Independen

Variabel	Jumlah Kategori	Kemungkinan Pemilah
Jenis Kelamin	2	1
Jaminan	5	15
Jangka Waktu	3	3
Jumlah Pembiayaan	5	15
Angsuran perbulan	3	3
EWS	2	1
Sisa Angsuran	5	15

Setelah dilakukan perhitungan jumlah kemungkinan pemilah dalam pembentukan pohon klasifikasi, selanjutnya adalah pemilihan pemilah menggunakan indeks gini. Indeks gini merupakan karakteristik dari CART. Berikut merupakan contoh atau ilustrasi perhitungan indeks gini pada variabel jenis kelamin.

Tabel 4.2 Ilustrasi Pemilihan pada Simpul Jenis Kelamin

Jenis Kelamin	Telat Bayar		Total
	Ya	Tidak	
Perempuan	21	360	381
Laki-laki	31	725	756

Selanjutnya melakukan perhitungan untuk nilai indeks gini pada masing-masing simpul kanan dan kiri sebagai berikut.

$$r(t_L) = 2 \times \left(\frac{21}{381} \times \frac{360}{381} \right) = 0,10416$$

$$r(t_R) = 2 \times \left(\frac{31}{756} \times \frac{725}{756} \right) = 0,07865$$

Kemudian menentukan kriteria *goodness of split* untuk evaluasi pemilahan yang telah dilakukan oleh pemilah *s* pada simpul *t*. Karena hanya ada satu kemungkinan pemilahan, maka untuk simpul jenis kelamin hanya ada satu kriteria *goodness of split*.

$$\phi(s,t) = 0,087286 - \left(\frac{381}{1137}\right) \times 0,10416 - \left(\frac{756}{1137}\right) \times 0,07865 = 8,78 \times 10^{-5}$$

Goodness of split pada variabel jenis kelamin adalah $8,78 \times 10^{-5}$. Selanjutnya indeks gini dan *goodness of split* variabel lainnya dihitung dengan perhitungan serupa variabel jenis kelamin. Variabel yang menjadi simpul akar adalah variabel dengan *goodness of split* paling tinggi. Hal tersebut berulang terus menerus hingga didapatkan simpul terminal. Simpul *t* dikatakan sebagai simpul terminal jika tidak terdapat penurunan heterogenitas atau dengan kata lain hanya terdapat satu kelas pada simpul anak.

Besarnya kontribusi variabel sebagai pemilah, baik pemilah uatam maupun pengganti pada pohon klasifikasi maksimal yang terbentuk ditunjukan melalui suatu angka skor yang ditampilkan pada tabel 4.3

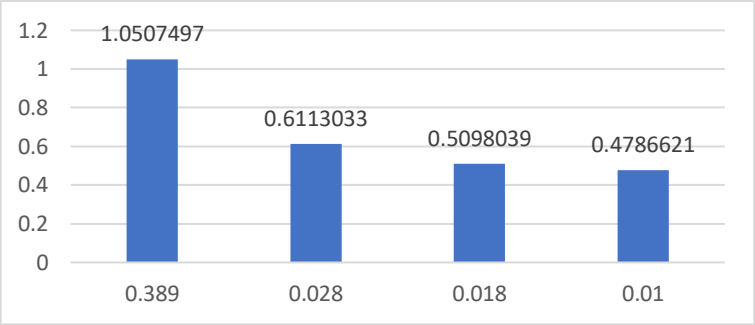
Tabel 4.3 Nilai Important Variabel

Variabel	Skor Variabel
EWS	0.4118
Angunan/Jaminan	0.1433
Jumlah Pembiayaan	0.1076
Jenis Kelamin	0.0948
Sisa Angsuran	0.0912
Jangka Waktu	0.0877
Angsuran perbulan	0.0635

Tabel 4.3 menunjukan bahwa skor yang dihasilkan diketahui variabel EWS merupakan variabel yang terepenting dan menjadi pemilah utama dalam mengklasifikasi nasabah telat bayar.

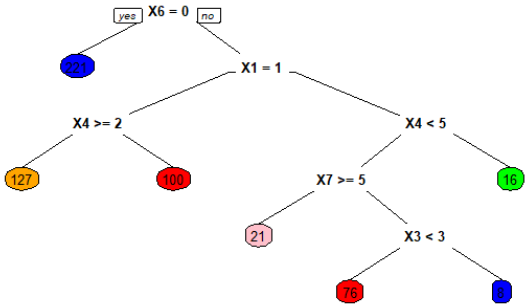
Langkah selanjutnya adalah melakukan pemangkasan pada pohon. Dalam hal ini nilai yang digunakan adalah *complexity paramter*. Pemangkasan pohon tersebut bertujuan untuk menghindari adanya overfitting yang diakibatkan oleh jumlah pemilahan yang terlalu banyak. Gambar 4.4 menunjukan nilai *complexity paramter* dan besar error, dimana dari gambar 4.4 menunjukan bahwa

complexity paramter terbaik adalah sebesar 0,01 karena memiliki error terkecil.



Gambar 4.4 Nilai Complexity Parameter beserta error Data Imbalance

Selanjutnya dibentuk pohon keputusan dengan menggunakan cp optimum sebagai berikut.



Gambar 4.5 CART Data *Imbalance*

Setelah melakukan analisis dengan menggunakan CART, dapat dilanjutkan dengan melakukan analisis dengan menggunakan *Random Forest.* Dalam penelitian ini jumlah pohon yang akan dibentuk adalah sebanyak 1000 pohon. *K-Fold Cross Validation* (KCV) diaplikasikan untuk membagi data training dan testing pada metode *Random Forest* dengan nilai *K* dari 2 hingga 10. Berikut merupakan hasil analisis menggunakan *Random Forest* yang ditunjukan pada tabel 4.3 dimana ditampilkan nilai *accuracy*, *sensitivity* dan AUC.

Tabel 4.4 Nilai Kebaikan Model *Random Forest Imbalance setiap K*

K	*Accuracy*		AUC		*Sensitivity*	
	Training	*Testing*	*Training*	*Testing*	*Training*	*Testing*
2	0.9561	0.9577	0.537	0.52	0.0741	0.04
3	0.9578	0.9578	0.57	0.5294	0.1429	0.0588
4	0.9555	0.9648	0.5366	0.5455	0.0732	0.0909
5	0.956	0.9648	0.5349	0.5556	0.0698	0.1111
6	0.9578	0.963	0.5779	0.5625	0.1591	0.125
7	0.9579	0.963	0.5853	0.5	0.1739	0
8	0.9578	0.9648	0.5633	0.5	0.1277	0
9	0.9585	0.9603	0.5724	0.4959	0.1458	0
10	0.959	0.9558	0.5922	0.4954	0.1875	0

Tabel 4.4 menunjukan kebaikan dari model *Random Forest* dengan nilai K dari 2 hingga 10, menunjukan bahwa K sebesar 6 memiliki nilai AUC testing paling besar diantara yang lain, dimana nilai tersebut adalah sebesar 0,5625 dengan nilai akurasi pada testing adalah sebesar 0,963. Tingginya nilai akurasi namun nilai AUC rendah mengindikasikan masing-masing kelas tidak diklasifikasikan secara tepat, atau kelas minoritas diklasifikasikan pada kelas mayoritas. Sebagian besar data akan diklasifikasikan pada kelas mayoritas karena kelas yang tidak seimbang. Oleh karena itu, data imbalance tidak dapat dibiarkan, karena klasifikasi yang dihasilkan akan menjadi salah.

4.2.2 Analisis Klasifikasi Nasabah Telat Bayar menggunakan Metode Regresi Logistik Biner

Metode yang berikutnya yang akan digunakan dalam melakukan analisis klasifikasi nasabah telat bayar adalah metode Regresi logistik biner. Metode Regresi logistik biner diterapkan untuk memprediksi nasabah telat bayar di Bank “x” dengan data yang *imbalance*. Berikut merupakan hasil kebaikan model menggunakan metode Regresi logistik biner dengan *K* dari 2 hingga 10.

Tabel 4.5 Nilai Kebaikan Model Regresi Logistik Biner

K	*Accuracy*		AUC		*Sensitivity*	
	Training	*Testing*	*Training*	*Testing*	*Training*	*Testing*
2	0.9525	0.956	0.5	0.5	0	0
3	0.9538	0.9551	0.5	0.5	0	0
4	0.9519	0.9613	0.5	0.5	0	0
5	0.9527	0.9604	0.5	0.5	0	0
6	0.9536	0.9577	0.5	0.5	0	0
7	0.9528	0.963	0.5	0.5	0	0
8	0.9528	0.9648	0.5	0.5	0	0
9	0.9525	0.9683	0.5	0.5	0	0
10	0.9531	0.9646	0.5	0.5	0	0

Tabel 4.5 menunjukan kebaikan dari model Regresi logistik biner dengan data yang masih imbalance, berdasarkan hasil tersebut didapatkan bahwa nilai AUC dari K 2 hingga10 menunjukan nilai yang sama yaitu sebesar 0,5, begitu pun juga pada nilai sensitivity menunjukan nilai sebesar 0, hal ini menunjukan bahwa kelas minoritas tidak diklasifikasikan secara tepat, melainkan kelas minoritas diklasifikasikan menjadi kelas mayoritas seluruhnya. Pada nilai akurasi menunjukan nilai yang tinggi, akurasi testing tertinggi terdapat *K*=9 dengan nilai sebesar 0,9683 sedangkan akurasi training tertinggi terdapat pada *K*=3 dengan nilai sebesar 0,9538. Tingginya nilai akurasi dan rendahnya AUC menunjukan tidak tepatnya klasifikasi antara kelas minoritas dan kelas mayoritas. Maka dari itu data imbalance tidak boleh untuk dibiarkan harus diatasi dengan penanganan imbalance agar klasifikasi yang didapatkan tidak salah.

4.3 Penanganan Imbalance Menggunakan SMOTE

Pada karakteristik data telah dijelaskan bahwa persentase nasabah yang mengalami telat bayar dan yang tidak telat bayar berbeda jauh, dimana nilainya yang tidak telat bayar adalah sebesar 95% dan yang telat bayar adalah sebesar 5% atau dapat dikatakan bahwa data pembiayaan nasabah tidak seimbang (*imbalance*). Selain

itu pada sub bab 4.2 juga terlihat bahwa penggunaan data imbalance pada klasifikasi *Random Forest* dan Regresi logistik biner keduanya tidak memberikan hasil yang baik. Oleh karena itu, perlu dilakukan penanganan *imbalance* sebelum melakukan analisis klasifikasi. Pada penelitian ini akan dilakukan *oversampling* (pembuatan data sintetis) untuk mengatasi imbalanced data. Penanganan hanya dilakukan pada data training, sedangkan data testing merupaka dataset asli. Metode penangan *imbalance* yang digunakan adalah SMOTE. Berikut merupakan penjelasan mengenai penangan *imbalance* menggunakan SMOTE.

SMOTE (*Synthetic Minority Oversampling Technique*) merupakan sebuah teknik membuat data sintetis (*oversampling*) sehingga jumlah nasabah yang mengalami telat bayar dab tidak telat bayar menjadi seimbang. Data baru dibangkitkan pada sekitar *j*-tetangga terdekat yang memiliki kemiripan berdasarkan jarak VDM. Berikut merupakan ilustrasi proses SMOTE untuk menangani kasus imbalance pada klasifikasi data nasabah telat bayar.

1. Mengambil data kelas minoritas secara random

Langkah pertama dalam SMOTE adalah menghitung jarak VDM antar kelas minoritas. Sebagai ilustrasi, data training dengan *K*=2 terdapat 27 data yang merupakan kelas minor (nasabah telat bayar), kemudian diambil salah satu observasi secara random. Pada kali ini misalkan terpilih data ke-65.

2. Menghitung Jarak VDM

Selanjutnya adalah menghitung jarak VDM data ke-65 dengan 26 data minoritas lainnya. Misal jarak data ke-65 dengan data ke-290. Berikut merupakan gambaran masing-masing variabel data ke-65 dan data ke-290.

Tabel 4.6 Data Ilustrasi SMOTE

Variabel	**Data ke-**	
	65	**290**
Jenis Kelamin	0	0
Angunan/Jaminan	**4**	**2**

Tabel 4.6 Data Ilustrasi SMOTE (Lanjutan)

Jangka Waktu	**1**	**3**
Jumlah Pembiayaan	**1**	**2**
Angsuran per Bulan	2	2
EWS	1	1
Sisa Angsuran	**1**	**3**

Tabel 4.6 menunjukan nilai atau kategori pada masing-masing variabel prediktor pada data ke-65 dan data ke-290. Terlihat bahwa terdapat beberapa variabel memiliki nilai/kategori yang sama, seperti variabel jenis kelamin, angsuran per bulan, serta EWS. Jarak anatara kategori yang sama adalah 0.

Selanjutnya, pada variabel dengan kelas berbeda dihitung nilai jarak VDM. Misalkan pada variabel Angunan/Jaminan, terlebih dahulu dibuat tabel *crosstabulation* seperti berikut.

Tabel 4.7 Ilustrasi SMOTE Jaminan

Jaminan	**Telat Bayar**	
	Tidak	**Ya**
BPJS	60	1
Sertifikat	250	16

Kemudian dihitung jarak antara dua kategori jaminan sebagai berikut.

$$\delta(\text{BPJS,Sertifikat}) = \left|\frac{60}{61} - \frac{250}{266}\right| + \left|\frac{1}{61} - \frac{16}{266}\right|$$

$$= 0{,}087514$$

Setelah didapatkan jarak antar kategori pada variabel jaminan, selanjutnya dilakukan perhitungan jarak pada variabel lainnya (Jangka Waktu, Jumlah Pembiayaan, Sisa Angsuran) dengan cara yang sama seperti variabel jaminan. Berikut merupakan jarak antar kategori pada masing-masing variabel.

Tabel 4.8 Ilustrasi SMOTE VDM

Variabel	$\delta(\mathrm{x}_b, y_b)$
Jenis Kelamin	0
Angunan/Jaminan	0,087514

Tabel 4.8 Ilustrasi SMOTE VDM (Lanjutan)

Jangka Waktu	0,006455
Jumlah Pembiayaan	0,012987
Angsuran per Bulan	0
EWS	0
Sisa Angsuran	0,004006

Untuk menghitung jarak VDM antara data ke-65 dan data ke-290 dapat dilakukan sebagai berikut.

$$\Delta(x_{65}, x_{290}) = \sum_{b=1}^{7} \delta(x_{65,b}, x_{290,b})^2 = 0,00788508$$

Jarak VDM anatar data ke-65 dan data ke-290 adalah sebesar 0,00788508. Perhitungan serupa juga dilakukan antara data ke-65 dengan 25 data minor lainnya.

3. Menentukan *j*-tetangga terdekat yang memiliki kemiripan

Setelah dilakukan perhitungan jarak VDM antara data ke-65 dengan seluruh data minoritas, selanjutnya jarak tersebut diurutkan sehingga didapatkan jarak terkecil hingga terbesar. Berikut merupakan ilustrasi hasil perhitungan jarak VDM dari data ke-65 dengan 10 data minor lainnya.

Tabel 4.9 Ilustrasi SMOTE Jarak VDM data ke-65

Data ke-	**Jarak VDM**
290	0.007885077
291	0.008217083
294	0.008244221
316	0.009188277
300	0.00965607
302	0.009840781
320	0.012619519
292	0.01530231
301	0.016146513
293	0.016878039

Tabel 4.9 menunjukan jarak VDM antara data ke-65 dengan 10 data minor lain secara urut dari nilai yang terkecil. Semakin kecil jarak, maka suatu data dikatakan semakin dekat atau semakin mirip. Dengan mengambil 7 data terdekat, maka data yang dekat dengan data ke-65 adalah data ke-290, ke-291, ke-294, ke-316, ke-300, ke-302, dan ke-320.

4. Membentuk Data Sintetis

Langkah selanjutnya adalah memberuk satu data sintetis dengan mempertimbangkan kategori mayoritas masing-masing variabel data terdekat. Diambil contoh 3 dari 7 variabel.

Tabel 4.10 Ilustrasi SMOTE Data Tetangga Terdekat

Data ke-	Variabel		
	Jenis Kelamin	**Jaminan**	**EWS**
290	0	2	1
291	0	2	1
294	0	2	1
316	0	2	1
300	0	2	1
302	0	2	1
320	1	2	1

Variabel data sintetis dibentuk berdasarkan mayoritas masing-masing variabel seperti pada tabel 4.10 misal pada variabel jenis kelamin terlihat bahwa 6 data terdekat merupakan kategori “0” yang berarti jenis kelamin perempuan dan 1 data terdekat merupakan kategori “1” yang berarti jenis kelamin laki-laki. Kemudian pada variabel jaminan, 7 data terdekat semua berkategori “2” yang berarti menggunakan jaminan sertifikat tanah. Selanjutnya variabel EWS, 7 data terdekat merupakan berkategori “1” yang berarti bahwa sudah mengalami sistem EWS.

Berdasarkan informasi tersebut, maka mayoritas pada variabel jenis kelamin adalah kategori “0”. Pada variabel jaminan adalah

kategori “2” dan pada variabel EWS adalah kategori “1”. Data sintetis baru yang terebntuk dari data ke-65 memiliki nilai sebagai berikut.

Tabel 4.11 Ilustrasi SMOTE Data Sintetis

Variabel	Kategori Data Sintetis
Jenis Kelamin	0
Jaminan	2
EWS	1

Ilustrasi pembentukan data sintetis di atas hanya memasukan 3 variabel, pada 4 variabel lainnya dilakukan dengan cara yang serupa sehingga didapatkan 1 data sintetis baru lengkap dengan 7 variabel yang terisi dengan memiliki variabel respon “1” (telat membayar). Proses 1 sampai 4 dilakukan berulang hingga didapatkan jumlah data *training* kelas minoritas seimbang dengan kelas mayoritas. Selanjutnya dilakukan analisis klasifikasi kembali dengan metode *Random Forest* dan Regresi logistik biner. Berikut merupakan hasil analisis klasifikasi metode *Random Forest* dan Regresi logistik biner.

a. Hasil Klasifikasi *Random Forest*

Analisis klasifikasi menggunakan metode *Random Forest* dilakukan kembali setelah penanganan imbalance, berikut merupakan kebaikan model yang dihasilkan dengan K dari 2 hingga 10.

Tabel 4.12 Nilai Kebaikan Model *Random Forest* setiap K

K	*Accuracy*		AUC		*Sensitivity*	
	Training	*Testing*	*Training*	*Testing*	*Training*	*Testing*
2	0.9013	0.7447	0.9013	0.5803	0.9797	0.4
3	0.8568	0.723	0.8568	0.5747	0.9142	0.4118
4	0.8208	0.6655	0.8208	0.6079	0.9212	0.5455
5	0.8201	0.7093	0.8201	0.6888	0.9296	0.6667
6	0.8142	0.709	0.8142	0.6689	0.9148	0.625
7	0.8089	0.6975	0.8089	0.6026	0.9128	0.5
8	0.8049	0.6831	0.8049	0.6431	0.9283	0.6
9	**0.8136**	**0.7063**	**0.8136**	**0.7275**	**0.919**	**0.75**
10	0.8053	0.6549	0.8053	0.5803	0.9324	0.5

Tabel 4.12 Menunjukan nilai kebaikan model metode *Random Forest* setiap K yang sudah dilakukan oversampling. Dimana berdasarkan tabel 4.11 dapat dikatehui bahwa hasil terbaik terdapat pada *K*=9. Hal itu ditunjukan dengan nilai AUC *testing* tertinggi dengan nilai sebesar 0,7275 dengan nilai AUC *training* 0,8136. Kemudian nilai akurasi yang dihasilkan tidak jauh beda dengan AUC, akurasi pada testing sebesar 0,7063 sedangkan pada akurasi training sebesar 0,8136. Nilai *sensitivity test* dan nilai *sensitivity training* masing-masing adalah sebesar 0,75 dan 0,919, nilai sensitivity adalah untuk menunjukan seberapa baik model untuk mengklasifikasikan kelas minoritas. Nilai AUC dan akurasi yang tidak jauh beda menunjukan bahwa hasil klasifikasi setiap kelas dapat diklasifikasikan hampir tepat, kelas minoritas dapat diklasifikasikan pada kelas minoritas dan juga kelas mayoritas diklasifikasin pada kelas mayoritas. Berdasarkan hal ini dapat dikatakan bahwa nilai AUC yang dihasilkan setelah dilakukan *oversampling* lebih baik dibandingkan nilai AUC yang belum dilakukan *oversampling*.

Tabel 4. 13 Hasil Prediksi *Random Forest*

Data Aktual	Prediksi	
	0	1
0	86	36
1	1	3

Berdasarkan tabel 4.13 dapat diketahui bahwa hasil prediksi pada data testing menunjukan dimana nilai data aktual yang berkategori “0” (tidak telat bayar) didapatkan prediksi yang sesuai adalah sebanyak 86 nasabah dan prediksi yang tidak sesuai adalah sebanyak 36 nasabah, sedangkan pada nilai data aktual yang berkategori “1” (telat bayar) didapatkan prediksi yang sesuai adalah sebanyak 3 nasabah dan yang tidak sesuai adalah sebanyak 1 nasabah.

b. Hasil Klasifikasi Regresi Logistik Biner

Setelah dilakukan analisis klasifikasi menggunakan metode *Random Forest*, selanjutnya adalah menggunakan metode Regresi

logistik biner. Berikut merupakan hasil kebaikan model dengan *K* dari 2 hingga 10.

Tabel 4.14 Nilai Kebaikan Model Regresi Logistik Biner setiap K

K	*Accuracy*		AUC		*Sensitivity*	
	Training	*Testing*	*Training*	*Testing*	*Training*	*Testing*
2	0.857	0.6849	0.857	0.549	0.9576	0.4
3	0.8257	0.6148	0.8257	0.5461	0.9668	0.4706
4	0.7617	0.5493	0.7617	0.5911	0.9323	0.6364
5	0.7636	0.5771	0.7636	0.5668	0.8997	0.5556
6	**0.7412**	**0.5767**	**0.7412**	**0.5998**	**0.9259**	**0.625**
7	0.7153	0.642	0.7153	0.4936	0.789	0.3333
8	0.7068	0.6197	0.7068	0.4175	0.7859	0.2
9	0.756	0.5952	0.756	0.5492	0.9065	0.5
10	0.7357	0.646	0.7357	0.4553	0.8289	0.25

Pada tabel 4.14 merupakan nilai kebaikan model metode Regresi logistik biner dari setiap *K* yang sudah dilakukan *oversampling*. Berdasarkan tabel 4. Dapat diketahui bahwa model yang paling baik diantara *K* dari 2 hingga 10 adalah terdapat pada *K*=6, hal itu ditunjukan dengan nilai AUC *testing* terbesar, nilai AUC *testing* adalah sebesar 0,5998 dan nilai AUC *training* adalah sebesar 0,7412. Kemudian nilai akurasi training dan testing masing-masing adalah sebesar 0,7412 dan 0,5767. Nilai AUC dan akurasi tidak jauh berbeda jauh, hal itu menunjukan bahwa klasifikasi Regresi logistik biner setelah dilakukan *oversampling* lebih baik dibanding Regresi logistik biner pada saat *imbalance*. Kemudian pada sensitivity nilai nya pada data training dan testing masing-masing adalah sebesar 0,9259 dan 0,625. Berikut merupakan hasil estimasi parameter yang didapatkan. Nilai yang didapatkan diatas menunjukan bahwa metode Regresi logistik biner belum mampu menghasilkan prediksi klasifikasi data nasabah yang telat bayar dengan baik. Setelah diketahui nilai kebaikan model terbaik, selanjutnya menghitung estimasi parameter, dimana estimasi parameter didapatkan dari data pada fold ke-6.

Tabel 4.15 Estimasi Parameter Regresi Logistik Biner

	coef	std error	z	p-value
Const	-42.8371	127000	0.000	1
Gender	-0.9766	0.123	-7.921	0.000*
EWS	44.4215	127000	0.000	1
Jaminan_2	-0.0827	0.351	-0.235	0.814
Jaminan_3	-0.5138	0.307	-1.671	0.095
Jaminan_4	-2.71	0.367	-7.377	0.000*
Jaminan_5	17.0602	3236.844	0.005	0.996
Jangka_Waktu_2	0.1842	0.242	0.762	0.446
Jangka_Waktu_3	-0.0427	0.312	-0.137	0.891
Jumlah_Pembiayaan_2	-0.4613	0.369	-1.250	0.211
Jumlah_Pembiayaan_3	-1.7966	0.564	-3.183	0.001*
Jumlah_Pembiayaan_4	0.5725	0.81	0.707	0.48
Jumlah_Pembiayaan_5	1.2513	0.852	1.469	0.142
Angsuran_perbulan_2	0.4653	0.235	1.980	0.048*
Angsuran_perbulan_3	-1.5033	0.63	-2.388	0.017*
Sisa_angsuran_2	-0.7168	0.414	-1.732	0.083*
Sisa_angsuran_3	-0.0898	0.556	-0.162	0.872
Sisa_angsuran_4	-0.0628	0.607	-0.103	0.918
Sisa_angsuran_5	-0.2066	0.723	-0.286	0.775

Hasil pengujian serentak didapatkan *p-value* pada *likelihood ratio test* adalah sebesar 0, dapat disimpulkan bahwa terdapat minimal satu variabel yang signifikan terhadap model. Pada uji parsial dimana nilai *p-value* ditampilkan pada tabel 4.15 Menunjukan bahwa terdapat 5 variabel yang signifikan terhadap model yaitu pada variabel jenis kelamin, jaminan, jumlah pembiayaan, angsuran per bulan dan sisa angsuran dimana hal itu dapat dilihat dari ketiga variabel tersebut memiliki *p-value* < 0,1. Berdasarkan tabel 4.15 juga dapat diketahui model yang dihasilkan adalah sebagai berikut.

$$g(x) = -42.83 - 0.9766Gender + 44.4215EWS - 0.0827\text{Jaminan\_2}$$
$$-0.5138\text{Jaminan\_3} - 2.71\text{Jaminan\_4} + 17.0602\text{Jaminan\_5}$$
$$+0.1842Jangka\_Waktu\_2 - 0.0427Jangka\_Waktu\_3$$
$$-0.4613Jumlah\_Pembiayaan\_2 - 1.7966Jumlah\_Pembiayaan\_3$$
$$+0.5725Jumlah\_Pembiayaan\_4 + 1.2513Jumlah\_Pembiayaan\_5$$
$$+0.4653Angsuran\_perbulan\_2 - 1.5033Angsuran\_perbulan\_3$$
$$-0.7168Sisa\_Angsuran\_2 - 0.0898Sisa\_Angsuran\_3$$
$$-0.0628Sisa\_Angsuran\_4 - 0.2066Sisa\_Angsuran\_5$$

Nilai *odd ratio* pada model Regresi logistik biner dapat dihitung berdasarkan nilai eksponen dari koefisien parameternya. Nasabah perempuan akan cenderung mengalami telat bayar 2,655 kali untuk mengalami telat bayar dibanding nasabah laki-laki. Kemudian, nasabah yang menggunakan jaminan emas akan cenderung mengalami telat bayar 15,029 kali dibandingkan nasabah yang menggunakan jaminan BPJS. Sementara itu, nasabah yang mendapatkan jumlah pembiayaan pada golongan 1 (< 100 juta) akan lebih mengalami telat bayar 6,029 kali dibanding nasabah yang mendapatkan jumlah pembiayaan pada golongan 3 (< 300 Juta). Apabila nasabah yang angsurannya per bulan pada golongan 2 (< 3,88 juta) akan mengalami telat bayar 1,59 kali dibanding nasabah yang angsuran per bulannya pada golongan 1 (< 564 ribu), dan juga nasabah yang angsuran per bulannya pada golongan 1 akan mengalami telat bayar 4,49 kali dibandingkan nasabah yang angsuran per bulannya pada golongan 3 (> 3,88 juta). Nasabah yang memiliki sisa angsuran golongan 1 (< 100 juta) akan cenderung mengalami telat bayar 2,04 kali dibanding nasabah yang memiliki sisa angsuran di golongan 2 (< 200 juta).

Tabel 4.16 Hasil Prediksi Regresi Logistik Biner

Data Aktual	Prediksi	
	0	1
0	104	77
1	3	5

Berdasarkan tabel 4.16 dapat diketahui bahwa hasil prediksi pada data testing menunjukan dimana nilai data aktual yang berkategori "0" (tidak telat bayar) didapatkan prediksi yang sesuai adalah sebanyak 104 nasabah dan prediksi yang tidak sesuai adalah sebanyak 77 nasabah, sedangkan pada nilai data aktual yang berkategori "1" (telat bayar) didapatkan prediksi yang sesuai adalah sebanyak 5 nasabah dan yang tidak sesuai adalah sebanyak 3 nasabah.

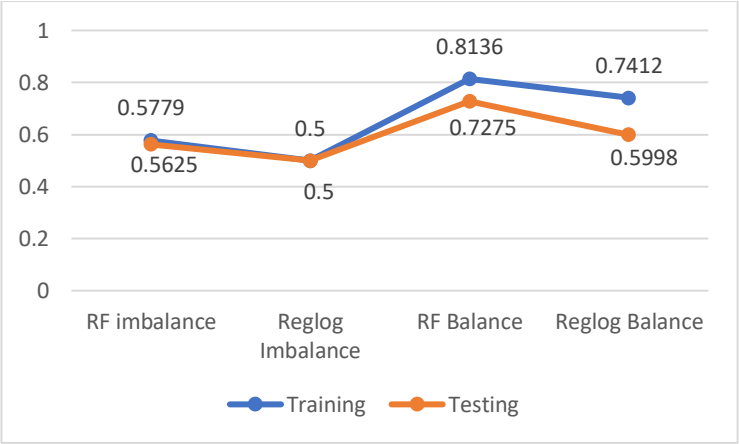
4.4 Perbandingan Performa Klasifikasi Antar Metode

Setelah dilakukan klasifikasi dengan menggunakan dua metode klasifikasi yaitu *Random Forest* dan Regresi logistik biner dan juga sudah dilakukan penanganan *imbalance*, didapatkan nilai *accuracy*, *sensitivity* dan AUC, maka akan dilakukan perbandingan kebaikan metode berdasarkan nilai-nilai tersebut untuk memilih metode mana yang terbaik untuk data pembiayaan nasabah di Bank "x". Berikut rangkuman hasil performa untuk setiap metode yang digunakan dapat dilihat pada tabel 4.17

Tabel 4. 17 Perbandingan Performa Metode RF dan Reglog

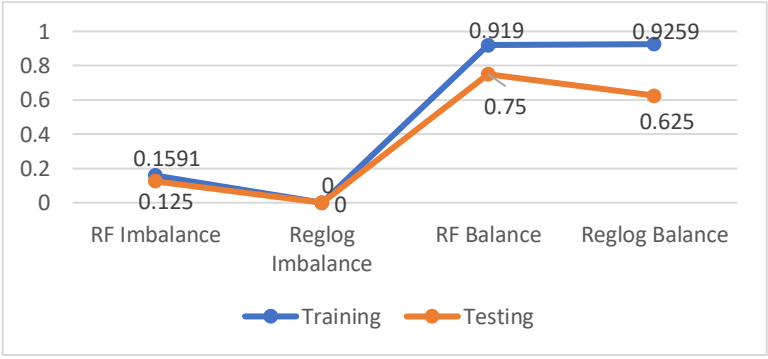
Performansi	Data	Metode	
		RF	Reglog
Accuracy	*Training*	**0,8136**	0,7412
	Testing	**0,7063**	0,5767
AUC	*Training*	**0,8136**	0,7412
	Testing	**0,7275**	0,5998
Sensitvity	*Training*	**0,919**	0,9259
	Testing	**0,75**	0,625

Tabel 4.17 menunjukan bahwa apabila dilihat dari nilai performansi kebaikan klasifikasi (*accuracy*, *sensitivity*, dan AUC) menunjukan bahwa nilai yang didapatkan menggunakan metode *Random Forest* lebih besar daripada nilai yang didapat menggunakan metode Regresi logistik biner, maka dapat disimpulkan bahwa metode *Random Forest* dapat menghasilkan klasifikasi data pembiayaan nasabah lebih baik daripada dengan menggunakan metode Regresi logistik biner.



Gambar 4.6 Nilai AUC *Imbalance* hingga *Balance*

Gambar 4.6 menunjukan bahwa nilai AUC secara keseluruhan pada data *training* lebih tinggi daripada data *testing*. Kebaikan model didapatkan lebih baik saat kondisi sudah dilakukan oversampling atau penanganan *imbalance*, hal itu terlihat di gambar bahwa nilai AUC RF *balance* dan Reglog *balance* lebih besar dibandingkan nilai AUC RF *imbalance* dan Reglog *imbalance*. Dimana kenaikan nilai kebaikan model pada metode *Random Forest* adalah sebesar 23,57% pada data *training* dan 16,5% pada data *testing*. Sedangkan pada metode Regresi logistik biner terjadi kenaikan nilai kebaikan model sebesar 24,12% pada data *training* dan 9,98% pada data *testing*.



Gambar 4.7 Nilai Sensitivity Imbalance hingga Balance

Pada gambar 4.7 dapat diketahui bahwa nilai sensitivity secara keseleruhan pada data *training* lebih tinggi daripada data *testing*. Nilai *sensitivity* terlihat lebih baik saat setelah dilakukan penanganan *imbalance*, hal itu terlihat bahwa nilai *sensitivity* RF *balance* dan Reglog *balance* lebih besar dibandingkan nilai RF *imbalance* dan Reglog *imbalance*. Dimana kenaikan nilai *sensitivity* pada metode *Random Forest* adalah sebesar 75,9% pada data *training* dan 62,5% pada data *testing*. Sedangkan pada metode Regresi logistik biner terjadi kenaikan nilai *sensitivity* sebesar 92,5% pada data *training* dan 62,5% pada data *testing*.

BAB V
KESIMPULAN DAN SARAN

5.1 Kesimpulan

Berdasarkan hasil analisis dan pembahasan yang telah dilakukan maka diperoleh kesimpulan bahwa, secara garis besar karakteristik nasabah yang melakukan peminjaman lebih banyak yang tidak telat bayar, apabila nasabah yang mengalami telat bayar lebih banyak maka kondisi Bank "x" akan tidak stabil dan bisa saja mengalami kebangkrutan. Jumlah nasabah yang melakukan peminjaman juga paling banyak adalah berjenis kelamin laki-laki. Nilai kebaikan model yang didapatkan pada saat data *imbalance* dan pada saat data sudah dilakukan *oversampling* dengan SMOTE, dari kedua metode menunjukan nilai kebaikan modelnya lebih baik pada saat sudah *balance* (sudah dilakukan *oversampling* dengan SMOTE).

Berdasarkan hasil setelah penanganan *imbalance*, metode yang lebih baik untuk melakukan klasifikasi data pembiayaan nasabah adalah metode *Random Forest* dibandingkan dengan metode Regresi logistik biner. Dimana jumlah pohon sebanyak 1000 pohon, dengan menggunakan nilai *K* sebesar 9 pada *K-Fold Cross Validation* didapatkan nilai performansi klasifikasi menggunakan akurasi, *sensitivity* dan AUC berturut turut pada data *training* adalah sebesar 81,36%, 91,9%, 81,36% dan pada data *testing* secara berturut turut sebesar 70,63%, 75%, 72,75%, juga didapatkan hasil prediksi yang berkategori "0" adalah sebanyak 86 nasabah dan hasil prediksi yang berkategori "1" adalah sebanyak 3 nasabah.

5.2 Saran

Berdasarkan kesimpulan yang telah diperoleh, maka saran yang dapat disampaikan untuk penelitian selanjutnya adalah dapat menggunakan metode machine learning yang lain seperti SVM, KNN, ANN, BPNN dan masih banyak lagi, yang mungkin dari metode-metode tersebut dapat memberikan hasil yang lebih baik. Penggunaan nilai parameter perlu diperhatikan agar mendapatkan hasil yang optimum.

Saran yang juga dapat diberikan adalah dapat melakukan penelitian mengenai telat bayar pada nasabah di lembaga keuangan yang lainnya, karena resiko kredit (nasabah telat bayar) tidak hanya terdapat di lembaga perbankan saja, namun juga terdapat di koperasi, lembaga perkreditan atau leasing.

DAFTAR PUSTAKA

Abdelmoula, A. K. (2015). Bank credit risk analysis with k-nearestneighbor classifier: Case of Tunisian banks. *Accounting and Management Information System Vol. 14, No. 1*, 79-106.

Adrianto. (2020). *Manajemen Kredit: Teori dan Konsep bagi Bank Umum.* Pasuruan: Qiara Media.

Belaid, F., Boussaada, R., & Belguith, H. (2017). Bank-firm relationship and credit risk: An analysis on Tunisian firms. *Research in International Business and Finance (42)*, 532-543.

Breiman, L. (2001). Random Forests. *Machine Learning, 45*, 5-32.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and Regression Trees.* Boca Raton: Chapman & Hall/CRC.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, P. W. (2002). SMOTE : Synthetic Minority Over-Sampling Technique. *Journal of Artificial Intelegence Research, Volume 16*, pp. 321-357.

Cost, S., & Salzberg, S. (1993). A Weighted Nearest Neighbor Algorithm for Learning with Symbolic Features. *Machine Learning Volume 10*, 57-78.

Dubey, R., Zhou, J., Wang, Y., Thompson, P. M., & Ye, J. (2014). Analysis of Sampling Techniques for Imbalanced Data: An n=648 ADNI Study. *Neuro Image, 87*, 220-241.

Geneur, R., Poggi, J., & Malot, C. T. (2009). *Variable Selection using Random Forest.* France: Laboratorie de Mathematiques, Universite Paris-Sud.

Ghatasheh, N. (2014). Business Analytics using Random Forest Trees for Credit Risk Prediction: A Comparison Study. *International Journal of Advanced Science and Technology Vol.72*, 19-30.

Gokgoz, E., & Subasi, A. (2015). Comparison of Decission Tree Algorithms for EMG Signal Classification Using DWT. *Biomedical Signal Processing and Control, 18*, 138-144.

Hilbe, J. M. (2009). *Logistic Regression Models.* USA: CRC Press.

Hosmer, D. W., & Lemeshow, S. (2000). *Applied Logistic Regression.* USA: John Wiley & Sons.

Huang, J., & Ling, C. X. (2005). Using AUC and Accuracy in Evaluating Learning Algotithms. *IEEE Transactions on Knowledge and Data Engineering*, 299-310.

Kruppa, J., Schwarz, A., Arminger, G., & Ziegler, A. (2013). Consumer credit risk : Individual Probability Estimates Using Machine Learning. *Expert Systems with Applications 40*, 5125-5131.

Laitinen, E. K. (2006). Partial Least Squares Regression in Payment Default Prediction. *Investment Management and Financial Innovations, 3(1),* 66-77.

Liu, M., Wang, M., Wang, J., & Duo, L. (2013). Comparison of Random Forest, Support Vector Machine an Back Propagation Neural Network for Electronic Tongue Data Classification : Application to the Recognition of Orange Beverage and Chinese Vinegar. *Sensors and Actuators B 177*, 970-980.

Muchtar, Rahmidani, B. R., & Kurnia, M. (2016). *Bank dan Lembaga Keuangan Lain.* Jakarta: Kencana.

Santos, M. S., Soares, J. P., Abreu, P. H., Araujo, H., & Santos, J. (2018). Cross-Validation for Imbalanced Datasets: Avoiding Overoptimistic and Overfittig Approaches. *IEEE Computational Intelligence Magazine.*

Sinungan, M. (2000). *Manajemen Dana Bank. Edisi Kedua.* Jakarat: PT. Bumi Aksara.

Speiser, J. L., Miller, M. E., Tooze, J., & Ip, E. (2019). A comparison of random forest variable selection methods for. *Expert Systems With Applications 134*, 93-101.

Surat Keputusan Menteri Keuangan Republik Indonesia No. 792 Tahun 1990 tentang Lembaga Keuangan. (1990). Jakarta: Sekertariat Negara.

Todaro, M. P. (1994). *Economic Development.* London: Longman.

Undang-Undang Republik Indonesia No. 10 Tahun 1998 tentang Perbankan. (1998). Jakarta: Sekertariat Negara.

Walpole, R. E. (2012). *Pengantar Metode Statistika Edisi ke-3. Diterjemahkan oleh : Bambang Sumantri.* Jakarta: Gramedia Pustaka Utama.

Yu, H., Huang, X., Hu, X., & Cai, H. (2010). A Comparative Study on Data Mining Algorithms for Indovidual Credit Risk Evaluation. *International Conference on Management of e-Commerce and e-Government*, 35-38.

LAMPIRAN

Lampiran 1 Data Penelitian

No	X1	X2	…	X6	X7	Y
1	1	3	…	0	1	0
2	1	3	…	1	1	0
3	0	3	…	1	1	0
4	1	3	…	0	1	0
5	0	1	…	1	1	0
6	0	1	…	0	5	0
7	1	4	…	0	1	0
8	0	4	…	0	1	0
9	0	4	…	1	1	0
10	0	4	…	0	1	0
11	0	4	…	0	1	0
…	…	…	…	…	…	…
…	…	…	…	…	…	…
1124	1	2	…	1	3	0
1125	1	2	…	1	5	0
1126	1	2	…	1	2	0
1127	1	2	…	1	3	0
1128	1	2	…	1	3	0
1129	0	2	…	1	2	0
1130	1	2	…	1	5	0
1131	1	2	…	1	4	0
1132	1	2	…	1	5	0
1133	0	2		1	4	0
1134	1	2		1	1	0
1135	1	2		1	3	0
1136	0	2		1	5	1
1137	1	2		1	5	1

Lampiran 2 Data Training Replikasi SMOTE K=9

No	X1	X2	...	X6	X7	Y
1	1	3	...	0	1	0
2	1	3	...	1	1	0
3	0	3	...	1	1	0
4	1	3	...	0	1	0
5	0	1	...	1	1	0
6	0	1	...	0	5	0
7	1	4	...	0	1	0
8	0	4	...	0	1	0
9	0	4	...	1	1	0
10	0	4	...	0	1	0
11	0	4	...	0	1	0
...	...	...	...	...	...	...
...	...	...	...	...	...	...
1721	0	3	...	1	3	0
1722	1	2	...	1	5	0
1723	1	3	...	1	2	0
1724	1	3	...	1	3	0
1725	0	3	...	1	3	0
1726	1	3	...	1	2	0
1727	1	3	...	1	5	0
1728	0	3	...	1	4	0
1729	1	2	...	1	5	0
1730	0	1	...	1	4	0
1731	1	2	...	1	1	0
1732	0	3	...	1	3	0
1733	1	3	...	1	5	1
1734	0	2	...	1	5	1

Lampiran 3 Syntax Kfold & SMOTE

```
import pandas as pd
from collections import Counter
from imblearn.over_sampling import SMOTE
from numpy import where
from sklearn.model_selection import KFold
data = pd.read_csv("D:\Documents\Kuliah\SEMESTER 8\TA\DATA
TA\Data Siap Olah FIX 2 Skenario 7.csv")
data.head()
X = data.drop(columns='Y',axis=1)
y = data['Y']
kf = KFold(n_splits=5,random_state=0, shuffle=True)
kf.get_n_splits(X)
print(kf)
for train_index, test_index in kf.split(X):
    print ("TRAIN:", train_index, "TEST:", test_index)
    X_train, X_test = X.iloc[train_index], X.iloc[test_index]
    y_train, y_test = y.iloc[train_index], y.iloc[test_index]
sm = SMOTE()
X_train_oversampled, y_train_oversampled = sm.fit_resample(X_train,
y_train)
counter = Counter(y_train_oversampled)
print(counter)
X_train_oversampled.to_csv('XtrainSMOTE.csv',sep=',')
y_train_oversampled.to_csv('YtrainSMOTE.csv',sep=',')
X_test.to_csv('XtestKfold.csv',sep=',')
y_test.to_csv('YtestKfold.csv',sep=',')
```

Lampiran 4 Syntax RF

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.ensemble import RandomForestClassifier

data = pd.read_csv("D:\Documents\Kuliah\SEMESTER 8\TA\DATA
TA\HASIL RUN\Data Train REPLIKASI.csv")
data1 = pd.read_csv("D:\Documents\Kuliah\SEMESTER 8\TA\DATA
TA\HASIL RUN\Data Test REPLIKASI.csv")
X_train = data.drop(columns = 'Y')
y_train = data['Y']
X_test = data1.drop(columns = 'Y')
y_test = data1['Y']

RF = RandomForestClassifier(random_state=0,n_estimators=1000)
RF.fit(X_train, y_train)
RF.score(X_test,y_test)

from sklearn.metrics import confusion_matrix, accuracy_score,
precision_score, recall_score, roc_auc_score, auc, roc_curve
import numpy as np
y_predRF=RF.predict(X_test)
y_predtRF=RF.predict(X_train)
conf_matrixRF=confusion_matrix(y_test, y_predRF)
ClassRF=data['Y'].unique()
conf_matrix_test=pd.DataFrame(data=conf_matrixRF,columns=ClassR
F,index=ClassRF)
conf_matrix_test

print('Akurasi Test RF:',accuracy_score(y_test, y_predRF))
print('Akurasi Train RF:',accuracy_score(y_train, y_predtRF))
print('AUC test RF :',roc_auc_score(y_test, y_predRF))
print('AUC train RF :',roc_auc_score(y_train, y_predtRF))
print('Sensitivity Test:',recall_score(y_test,y_predRF))
print('Sensitivity Train:',recall_score(y_train,y_predtRF))
```

Lampiran 5 Syntax Reglog Biner

```
import pandas as pd
import numpy as np
from sklearn import linear_model
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score, roc_auc_score,
confusion_matrix

data = pd.read_csv("D:\Documents\Kuliah\SEMESTER 8\TA\DATA
TA\HASIL RUN\Data Train REPLIKASI.csv")
data1 = pd.read_csv("D:\Documents\Kuliah\SEMESTER 8\TA\DATA
TA\HASIL RUN\Data Test REPLIKASI.csv")
X_train = data.drop(columns = 'Y')
y_train = data['Y']
X_test = data1.drop(columns = 'Y')
y_test = data1['Y']

##Membuat Dummy Variabel
#training
dummy_X2_train=pd.get_dummies(X_train['X2'], prefix='X2')
dummy_X2_train= dummy_X2_train.iloc[:,1:5]
dummy_X3_train=pd.get_dummies(X_train['X3'], prefix='X3')
dummy_X3_train= dummy_X3_train.iloc[:,1:3]
dummy_X4_train=pd.get_dummies(X_train['X4'], prefix='X4')
dummy_X4_train= dummy_X4_train.iloc[:,1:5]
dummy_X5_train=pd.get_dummies(X_train['X5'], prefix='X5')
dummy_X5_train= dummy_X5_train.iloc[:,1:3]
dummy_X7_train=pd.get_dummies(X_train['X7'], prefix='X7')
dummy_X7_train= dummy_X7_train.iloc[:,1:5]
```

Lampiran 5 Syntax Reglog Biner (Lanjutan)

```
#test
dummy_X2_test=pd.get_dummies(X_test['X2'], prefix='X2')
dummy_X2_test= dummy_X2_test.iloc[:,1:5]
dummy_X3_test=pd.get_dummies(X_test['X3'], prefix='X3')
dummy_X3_test= dummy_X3_test.iloc[:,1:3]
dummy_X4_test=pd.get_dummies(X_test['X4'], prefix='X4')
dummy_X4_test= dummy_X4_test.iloc[:,1:5]
dummy_X5_test=pd.get_dummies(X_test['X5'], prefix='X5')
dummy_X5_test= dummy_X5_test.iloc[:,1:3]
dummy_X7_test=pd.get_dummies(X_test['X7'], prefix='X7')
dummy_X7_test= dummy_X7_test.iloc[:,1:5]

x_train_lg=X_train.drop(['X2','X3','X4','X5','X7'],axis=1)
x_test_lg=X_test.drop(['X2','X3','X4','X5','X7'],axis=1)
x_train_lg=pd.concat([x_train_lg,dummy_X2_train,dummy_X3_train,du
mmy_X4_train,dummy_X5_train,dummy_X7_train],axis=1)
x_test_lg=pd.concat([x_test_lg,dummy_X2_test,dummy_X3_test,dumm
y_X4_test,dummy_X5_test,dummy_X7_test],axis=1)

#Estimasi Parameter
import statsmodels.api as sm
xtrain_all=sm.add_constant(x_train_lg)
data_train=pd.concat([y_train,xtrain_all],axis=1)
data_train=data_train.astype('float64')
x_train_est=data_train.columns[1:]
est_reglog = sm.Logit(data_train['Y'],data_train[x_train_est])
result=est_reglog.fit()
print(result.summary())

from sklearn.metrics import confusion_matrix, accuracy_score,
precision_score, auc, roc_curve, recall_score
#Kebaikan Model Training
lg = LogisticRegression(random_state=0)
model=lg.fit(x_train_lg,y_train)
y_pred_train=model.predict(x_train_lg)
print('AUC Training',roc_auc_score(y_train, y_pred_train))
print('Akurasi Training:',accuracy_score(y_train,y_pred_train))
print('Sensitivity Training:',recall_score(y_train,y_pred_train))
```

Lampiran 5 Syntax Reglog Biner (Lanjutan)

```
conf_matrixlg=confusion_matrix(y_train,y_pred_train)
Classlg=data['Y'].unique()
conf_matrix_train=pd.DataFrame(data=conf_matrixlg,columns=Classlg,
index=Classlg)
conf_matrix_train

#Kebaikan Model Testing
y_pred_test=model.predict(x_test_lg)
print('AUC Testing',roc_auc_score(y_test, y_pred_test))
print('Akurasi Testing:',accuracy_score(y_test, y_pred_test))
print('Sensitivity Testing:',recall_score(y_test, y_pred_test))

conf_matrixlg_test=confusion_matrix(y_test,y_pred_test)
Classlg=data['Y'].unique()
conf_matrix_test=pd.DataFrame(data=conf_matrixlg_test,columns=Clas
slg,index=Classlg)
conf_matrix_test
```

Lampiran 6 Output Estimasi Parameter

Logit Regression Results

Dep. Variable:	Y	No. Observations:	1808
Model:	Logit	Df Residuals:	1789
Method:	MLE	Df Model:	18
Date:	Tue, 09 Jun 2020	Pseudo R-squ.:	0.3117
Time:	19:14:38	Log-Likelihood:	-862.55
converged:	False	LL-Null:	-1253.2
Covariance Type:	nonrobust	LLR p-value:	2.992e-154

	coef	std err	z	P>|z|	[0.025	0.975]
const	-42.8371	1.27e+05	-0.000	1.000	-2.49e+05	2.49e+05
X1	-0.9766	0.123	-7.921	0.000	-1.218	-0.735
X6	44.4215	1.27e+05	0.000	1.000	-2.49e+05	2.49e+05
X2_2	-0.0827	0.351	-0.235	0.814	-0.771	0.606
X2_3	-0.5138	0.307	-1.671	0.095	-1.116	0.089
X2_4	-2.7100	0.367	-7.377	0.000	-3.430	-1.990
X2_5	17.0602	3236.844	0.005	0.996	-6327.037	6361.158
X3_2	0.1842	0.242	0.762	0.446	-0.289	0.658
X3_3	-0.0427	0.312	-0.137	0.891	-0.655	0.570
X4_2	-0.4613	0.369	-1.250	0.211	-1.185	0.262
X4_3	-1.7966	0.564	-3.183	0.001	-2.903	-0.690
X4_4	0.5725	0.810	0.707	0.480	-1.015	2.160
X4_5	1.2513	0.852	1.469	0.142	-0.418	2.920
X5_2	0.4653	0.235	1.980	0.048	0.005	0.926
X5_3	-1.5033	0.630	-2.388	0.017	-2.737	-0.269
X7_2	-0.7168	0.414	-1.732	0.083	-1.528	0.094
X7_3	-0.0898	0.556	-0.162	0.872	-1.179	0.999
X7_4	-0.0628	0.607	-0.103	0.918	-1.252	1.127
X7_5	-0.2066	0.723	-0.286	0.775	-1.624	1.211

Lampiran 7 Surat Pernyataan Data

SURAT KETERANGAN

Assalamu'alaikum Warahmatullahhi Wabarakatuh,

Yang bertanda tangan dibawah ini menerangkan bahwa :

1. Mahasiswa Departemen Statistika Fakultas Sains dan Analitika Data Institut Sepuluh Nopember Surabaya dengan identitas sebagai berikut:
   Nama : M. Kholilul Mutaal
   NIM : 06211640000013
   Telah mengambil data di instansi perusahaan kami. Sejak tanggal 2 Maret 2020 sampai dengan 20 Maret 2020. Untuk keperluan Tugas Akhir/ Skripsi.
2. ~~Tidak Keberatan~~/Keberatan* nama perusahaan dicantumkan dalam Tugas Akhir/Skripsi mahasiswa statistika yang akan di simpan di Perpustakaan ITS dan dibaca di lingkungan ITS
3. Tidak Keberatan~~/Keberatan~~* bahwa hasil analisis data dari perusahaan dipublikasikan dalam Tugas Akhir/Skripsi mahasiswa Statistika.

Demikian surat keterangan ini dibuat untuk digunakan sebagaimana mestinya.

Wassalamu'alaikum Warahmatullahhi Wabarakatuh.

Surabaya, 20 Maret 2020

BIODATA PENULIS

M. Kholilul Muta'al biasa dipanggil dengan nama Kholil yang merupakan anak ketiga dari tiga bersaudara dan dilahirkan di Kabupaten Banyuwangi pada tanggal 10 Januari 1998. Pendidikan yang telah ditempuh oleh penulis adalah SDN 2 Genteng (2004-2010), SMP Bustanul Makmur Genteng (2010-2013), dan SMAN 1 Genteng (2013-2016). Kemudian dilanjutkan dengan menempuh pendidikan di Institut Teknologi Sepuluh Nopember Departemen Statistika. Selain dalam bidang akademik, penulis juga aktif organisasi di Himpunan Mahasiswa Statistika ITS (HIMASTA-ITS) sebagai staff kesenian dan olahraga (KESRA) masa periode 2017/2018. Selain itu, penulis juga aktif dalam mengikuti kepanitiaan yang diadakan oleh tingkat jurusan maupun tingkat ITS, seperti fasilitator keamanan dan perizinan dalam kegiatan GERIGI ITS 2017, staff perlengkapan PAMMITS 2017, staff lapangan pada pagelaran IFC 2018, serta menjadi kelompok pelaksanaan pemungutan suara di pemilu ITS 2018. Penulis juga pernah menjadi ketua penanggung jawab pada STATION 2018. Selama menjalani perkuliahan penulis juga berkesempatan dalam menjalani program magang di UPT Badan Pendapatan Daerah Tuban. Penulis juga pernah mengikuti kegiatan survei sebagai pengaplikasian ilmu statistika. Jika ingin memberikan saran, kritik, dan diskusi lebih lanjut, dapat menghubungi melalui email: mkholil28@gmail.com