PREDIKSI TINGKAT RISIKO KREDIT DENGAN *RANDOM OVER-UNDER SAMPLING* PADA METODE *ENSEMBLE* MENGGUNAKAN ALGORITMA *DECISION TREE* ID3, *RANDOM FOREST* DAN REGRESI LOGISTIK BINER

SKRIPSI

Diajukan sebagai salah satu syarat untuk memperoleh gelar Sarjana di Jurusan Matematika pada Fakultas MIPA

Oleh:

FAHIRA ANGGRAINI

08011181823014

JURUSAN MATEMATIKA

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM

UNIVERSITAS SRIWIJAYA

2022

HALAMAN PENGESAHAN

PREDIKSI TINGKAT RISIKO KREDIT DENGAN *RANDOM OVER-UNDER SAMPLING* PADA METODE *ENSEMBLE* MENGGUNAKAN ALGORITMA *DECISION TREE* ID3, *RANDOM FOREST* DAN REGRESI LOGISTIK BINER

SKRIPSI

Diajukan sebagai salah satu syarat untuk memperoleh gelar Sarjana di Jurusan Matematika pada Fakultas MIPA

Oleh:
FAHIRA ANGGRAINI
NIM. 08011181823014

Indralaya, Juni 2022

Pembimbing Kedua

Drs. Endro Setyo Cahyono, M.Si.
NIP. 196409261990021002

Pembimbing Utama

Dr. Yulia Resti, M.Si.
NIP. 197307191997022001

Mengetahui,
Ketua Jurusan Matematika

Drs. Sugandi Yahdin, M.M
NIP. 195807271986031003

PERNYATAAN KEASLIAN KARYA ILMIAH

Yang bertanda tangan dibawah ini:

Nama mahasiswa	: Fahira Anggraini
NIM	: 08011181823014
Fakultas/Jurusan	: MIPA / Matematika

Menyatakan bahwa skripsi ini adalah hasil karya saya sendiri dan karya ilmiah ini belum pernah diajukan sebagai pemenuhan persyaratan untuk memperoleh gelar kesarjanaan strata satu (S1) dari Universitas Sriwijaya maupun perguruan tinggi lain.

Semua informasi yang dimuat dalam skripsi ini yang berasal dari penulis lain baik yang dipublikasikan atau tidak telah diberikan penghargaan dengan mengutip nama sumber penulis secara benar. Semua isi dari skripsi ini sepenuhnya menjadi tanggung jawab saya sebagai penulis.

Demikianlah surat pernyataan ini saya buat dengan sebenarnya.

Indralaya, 10 Juni 2022

Penulis,

Fahira Anggraini

NIM. 08011181823014

HALAMAN PERSEMBAHAN

MOTTO

"Menuntut ilmu adalah takwa. Menyampaikan ilmu adalah ibadah. Mengulang-ulang ilmu adalah dzikir. Mencari ilmu adalah jihad."
(Abu Hamid Al Ghazali)

"Memulai dengan Penuh Keyakinan, Menjalankan dengan Penuh Keikhlasan, Menyelesaikan dengan Penuh Kebahagiaan"

"Tidak ada kesuksesan tanpa kerja keras. Tidak ada keberhasilan tanpa sikap pantang menyerah. Tidak ada kemudahan tanpa doa."

"Kerja keras tidak akan mengkhianati hasil. Apapun kerja keras itu akan selalu membuahkan hasil yang berguna bagi kehidupan."

Skripsi ini kupersembahkan kepada:

1. **Allah SWT**
2. **Orangtuaku**
3. **Saudaraku**
4. **Keluarga Besarku**
5. **Semua Dosen dan Guruku**
6. **Almamaterku**
7. **Sahabatku**

KATA PENGANTAR

Assalamu'alaikum Warahmatullahi Wabarakatuh

Puji syukur kehadirat Tuhan Yang Maha Esa atas limpahan rahmat dan karunia-Nya sehingga penulis dapat menyelesaikan skripsi ini dengan judul "**Prediksi Tingkat Risiko Kredit dengan *Random Over-Under Sampling* pada Metode *Ensemble* menggunakan Algoritma *Decision Tree* ID3, *Random Forest* dan Regresi Logistik Biner"**. Shalawat beserta salam senantiasa tercurahkan kepada baginda kita Nabi Muhammad SAW beserta keluarga, sahabat dan para pengikutnya hingga akhir zaman.

Penulis menyadari bahwa dalam proses penyusunan skripsi ini masih ada kekurangan, serta banyaknya rintangan dan tantangan dalam mengerjakannya. Namun dengan kesabaran dan ketekunan yang dilandasi rasa tanggung jawab sehingga penulis dapat menyelesaikan skripsi ini dengan baik.

Penulis mengucapkan terima kasih yang sebesar-besarnya kepada orang tua tercinta, yaitu Ibu **Budi Harti** yang tidak pernah lupa mendoakan yang terbaik, merawat, memberikan perhatian dan kasih sayang, nasihat, serta restunya kepada penulis dan memberikan dukungan atas apa yang penulis pilih. Penulis juga mengucapkan terima kasih sebesar-besarnya kepada saudara tercinta, yaitu **Riana Widiastuty** yang selalu mendoakan, memberikan semangat, nasihat, kasih sayang serta segala bantuan baik dalam bentuk material, fisik maupun psikis.

Penulis juga mengucapkan terima kasih kepada:

1. Bapak **Prof. Hermansyah, S.Si., M.Si., Ph.D.** selaku Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Sriwijaya.

2. Bapak **Drs. Sugandi Yahdin, M.M.** selaku Ketua Jurusan Matematika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Sriwijaya.
3. Ibu **Dr. Dian Cahyawati, M.Si.** selaku Sekretaris Jurusan Matematika yang telah membantu proses administrasi pendaftaran seminar, serta ilmunya.
4. Ibu **Dr. Yulia Resti, M.Si.** selaku Dosen Pembimbing Utama yang telah bersedia meluangkan waktu untuk memberikan bimbingan, nasihat, saran dan motivasi serta pengetahuan yang sangat bermanfaat bagi penulis dalam menyelesaikan skripi yang baik ini.
5. Bapak **Drs. Endro Setyo Cahyono, M.Si.** selaku Dosen Pembimbing Kedua yang telah memberikan bimbingan, saran, nasihat dan motivasi kepada penulis sehingga penulis dapat menyelesaikan skripsi dengan baik.
6. Ibu **Dra. Ning Eliyati, M.Pd.** selaku Dosen Pembimbing Akademik yang selalu memberikan bimbingan, nasihat, dan motivasi kepada penulis tentang urusan akademik selama masa pembelajaran serta dalam pengerjaan skripsi.
7. Ibu **Dr. Yuli Andriani, M.Si.** selaku Dosen Pembahas I dan Ibu **Novi Rustiana Dewi, M.Si.** selaku Dosen Pembahas II yang telah memberikan masukan, dan saran yang bermanfaat bagi penulis dalam pengerjaan skripsi.
8. Bapak **Drs. Putra B.J. Bangun, M.Si.** selaku Ketua Seminar dan Ibu **Sri Indra Maiyanti, M.Si.** selaku Sekretaris Seminar yang telah membantu pelaksanaan seminar bagi penulis, sehingga dapat berjalan dengan baik.
9. **Seluruh Dosen di Jurusan Matematika FMIPA UNSRI** atas ilmu dan didikan yang telah diberikan selama penulis menempuh pendidikan di Jurusan Matematika FMIPA UNSRI.

10. Bapak **Irwansyah** dan Ibu **Hamida** yang telah banyak membantu penulis dalam hal administrasi di Jurusan Matematika FMIPA UNSRI.
11. Keluarga besarku serta saudara saya yang saya cintai **Riana Widiastuty, Tineke Andari, dan Oktavia Tantri** atas do'a, semangat, nasihat, motivasi serta dukungan yang telah diberikan kepada penulis.
12. Sahabat-sahabatku **Siti Hasmawati, Desi Herlina Saraswati, Nurafni Rahayu Khotimah, Nafasa Istiqoza, Santi Puji Lestari, Miftahul Jannah, Sukmalina, Imelda Putri Rizky, Rani Lestari, dan Mahdiyah Afifah Sari** serta seluruh teman-teman Angkatan **2018** yang telah saling menguatkan, membantu, mengajarkan, dan kebersamaannya.
13. Kakak tingkat Angkatan **2016** dan **2017** yang telah memberikan ilmu dan pengalamannya serta adik tingkat Angkatan **2019** dan **2020** yang telah memberikan dukungan, semangat serta doa.
14. **Semua Pihak** yang telah membantu penulis dalam menyelesaikan skripsi ini yang tidak dapat disebutkan satu persatu. Semoga semua kebaikan yang diberikan mendapat balasan dari Allah SWT.

Penulis berharap agar skripsi ini dapat bermanfaat bagi semua pihak yang membutuhkan terutama mahasiswa/mahasiswi Jurusan Matematika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Sriwijaya.

Wassalamu'alaikum Warahmatullahi Wabarakatuh

Indralaya, Juni 2022

Penulis

PREDICTION OF CREDIT RISK
WITH RANDOM OVER-UNDER SAMPLING
ON ENSEMBLE METHOD USING
DECISION TREE ID3, RANDOM FOREST, AND
BINARY LOGISTIC REGRESSION ALGORITHM

By:
FAHIRA ANGGRAINI
08011181823014

ABSTRACT

Credit-granting activities are included in business activities that have a high risk and affect the sustainability of the company as well as other financial institutions. In credit activities, non-performing loans often occur due to the failure to repay a number of loans in accordance with the agreed time. The problem of providing credit can be overcome, one of which is by identifying and predicting prospective customers before giving credit. Datasets used to predict sometimes have class imbalance problems. This problem is usually solved by resampling method. Therefore, this research was conducted with the aim of predicting the level of credit risk by implementing Random Over-Under Sampling in the Ensemble method using Decision Tree ID3, Random Forest, and Binary Logistics Regression. The data used is a dataset of credit card approval UCI Repository. The results showed that the Ensemble method has a better overall classification effectiveness level than others, as seen from the higher accuracy, precision, and fscore values, while the better classification effectiveness level in the form of recall is Binary Logistics Regression. Prediction classification using decision tree resulted in accuracy, precision and recall of 77.79%, 49.82, 45.95%, 47.76%, respectively. Prediction classification using random forest resulted in accuracy, precision and recall of 78.10%, 50.55%, 45.31%, 47.76%, respectively. Prediction classification using binary logistic regression resulted in accuracy, precision and recall of 74.16%, 42.66%, 48.90%, 45.55%, respectively. Prediction classification using ensemble majority vote resulted in accuracy, precision and recall of 78.22%, 50.86%, 45.54%, 48.03%, respectively.

Keywords: Credit Risk, Ensemble, Decision Tree, Random Forest, Binary Logistics Regression

PREDIKSI TINGKAT RISIKO KREDIT
DENGAN *RANDOM OVER-UNDER SAMPLING*
PADA METODE *ENSEMBLE* MENGGUNAKAN
ALGORITMA *DECISION TREE* ID3, *RANDOM FOREST* DAN
REGRESI LOGISTIK BINER

Oleh:
FAHIRA ANGGRAINI
08011181823014

ABSTRAK

Kegiatan pemberian kredit termasuk dalam kegiatan usaha yang memiliki risiko tinggi serta berpengaruh pada keberlangsungan perusahaan juga lembaga keuangan lainnya. Di dalam kegiatan perkreditan sering terjadi kredit bermasalah yang disebabkan gagalnya pengembalian sejumlah pinjaman sesuai dengan waktu yang sudah disepakati. Masalah pemberian kredit ini dapat diatasi, salah satunya dengan melakukan identifikasi dan prediksi pada calon nasabah sebelum memberikan kredit. *Dataset* risiko kredit yang digunakan untuk memprediksi terkadang memiliki permasalahan ketidakseimbangan kelas. Permasalahan ini biasanya diselesaikan dengan metode *resampling*. Oleh karena itu, penelitian ini dilakukan dengan tujuan prediksi tingkat risiko kredit yang mengimplementasikan *Random Over-Under Sampling* pada metode *Ensemble* menggunakan algoritma *Decision Tree* ID3, *Random Forest*, dan Regresi Logistik Biner. Data yang digunakan yaitu *dataset approval credit card* UCI Repository. Hasil penelitian menunjukkan bahwa metode *Ensemble* memiliki tingkat ketepatan klasifikasi keseluruhan yang lebih baik dibandingkan lainnya, terlihat dari nilai *accuracy*, *precision*, dan *fscore* yang lebih tinggi, sedangkan tingkat ketepatan klasifikasi yang lebih baik berupa *recall* adalah Regresi Logistik Biner. Klasifikasi prediksi menggunakan *decision tree* menghasilkan *accuracy*, *precision*, *recall*, dan *fscore* berturut-turut sebesar 77,79%; 49,82%; 45,95%; 47,76%. Klasifikasi prediksi menggunakan *random forest* menghasilkan *accuracy*, *precision*, *recall*, dan *fscore* berturut-turut sebesar 78,10%; 50,55%; 45,31%; 47,76%. Klasifikasi prediksi menggunakan regresi logistik biner menghasilkan *accuracy*, *precision recall*, dan *fscore* berturut-turut sebesar 74,16%; 42,66%; 48,90%; 45,55%. Klasifikasi prediksi menggunakan metode *ensemble* menghasilkan *accuracy*, *precision*, *recall*, dan *fscore* berturut-turut sebesar 78,22%; 50,86%; 45,54%; 48,03%.

Kata Kunci: Risiko Kredit, *Ensemble*, *Decision Tree*, *Random Forest*, Regresi Logistik Biner

DAFTAR ISI

HALAMAN JUDUL ..... **i**

HALAMAN PENGESAHAN ..... **ii**

HALAMAN PERSEMBAHAN ..... **iii**

KATA PENGANTAR ..... **iv**

ABSTRACT ..... **vii**

ABSTRAK ..... **viii**

DAFTAR TABEL ..... **xii**

DAFTAR GAMBAR ..... **xiii**

DAFTAR LAMPIRAN ..... **xiv**

BAB I PENDAHULUAN **1**

1.1 Latar Belakang 1

1.2 Rumusan Masalah 5

1.3 Batasan Masalah 6

1.4 Tujuan Penelitian 6

1.5 Manfaat Penelitian 7

BAB II TINJAUAN PUSTAKA **8**

2.1 Kredit 8

2.1.1 Unsur – Unsur Kredit 8

2.1.2 Prinsip – Prinsip Pemberian Kredit 9

2.1.3 Risiko Kredit 12

2.2 *Statistical Learning* 13

2.3 *Data Mining* 13

2.4 *Machine Learning* 14

2.5 *Repeated K-Fold Cross Validation* 14
2.6 Klasifikasi 15
2.7 *Imbalanced Class* 16
2.8 *Random Over-Under Sampling* 18
2.9 Metode *Ensemble* 19
2.10 Algoritma *Decision Tree* 20
2.10.1 Algoritma ID3 21
2.11 Algoritma *Random Forest* 23
2.12 Analisis Regresi Logistik Biner 26
2.12.1 Estimasi Parameter 29
2.12.2 Uji Model Regresi Logistik (Uji Serentak) 30
2.12.3 Uji Hipotesis Parsial 31
2.12.4 Uji Kesesuaian Model 32
2.12.5 *Odds Ratio* 33
2.13 *Confusion Matrix* 33
BAB III METODOLOGI PENELITIAN 36
3.1 Tempat 36
3.2 Waktu 36
3.3 Metode Penelitian 36
BAB IV HASIL DAN PEMBAHASAN 41
4.1 Deskripsi Data 41
4.2 Diskritisasi Data 42
4.3 Ringkasan Data 45
4.4 Partisi Data 46
4.5 *Resampling* dengan *Random Over-Under Sampling* (ROUS) 47

4.6 Algoritma *Decision Tree* ID3 48
4.7 Algoritma *Random Forest* 55
4.8 Algoritma Regresi Logistik Biner.....64
4.8.1 Uji Serentak.....64
4.8.2 Uji Parsial.....65
4.8.3 Uji Kesesuaian Model.....72
4.8.4 *Odds Ratio*.....73
4.8.5 Model Terbaik.....77
4.8.6 Probabilitas.....78
4.9 Metode *Ensemble*.....81
4.10 Analisis Hasil.....85
BAB V KESIMPULAN DAN SARAN 86
5.1 Kesimpulan 86
5.2 Saran 87
DAFTAR PUSTAKA 88
LAMPIRAN..... 92

DAFTAR TABEL

Tabel 2.1 *Confusion Matrix* 34
Tabel 4.1 Keterangan Variabel 41
Tabel 4.1 Keterangan Variabel (Lanjutan)..... 42
Tabel 4.2 Ringkasan Data Variabel *Dependent* 45
Tabel 4.3 Data *Train* 46
Tabel 4.4 Data *Test* 46
Tabel 4.5 Data *Train* ROUS 47
Tabel 4.6 Nilai *Entropy* Kategorik Variabel *Independent* 49
Tabel 4.7 Nilai *Gain* Variabel *Independent* Terhadap Variabel *Dependent*..... 50
Tabel 4.8 Prediksi dengan *Decision Tree* 53
Tabel 4.9 *Confusion Matrix* Algoritma *Decision Tree* 53
Tabel 4.10 *Accuracy*, *precision*, *recall*, *fscore* seluruh model *decision tree* 54
Tabel 4.11 Sampel Pohon ke-1 55
Tabel 4.12 Nilai *Entropy* Kategorik Variabel *Independent* 57
Tabel 4.13 Nilai *Gain* Variabel *Independent* Terhadap Variabel *Dependent*..... 58
Tabel 4.14 Prediksi dengan *Random Forest* 62
Tabel 4.15 *Confusion Matrix* Algoritma *Random Forest* 62
Tabel 4.16 *Accuracy*, *precision*, *recall*, *fscore* seluruh model *random forest* 63
Tabel 4.17 Estimasi Model Regresi Logistik Biner 64
Tabel 4.18 Hasil Uji Serentak 65
Tabel 4.19 Hasil Uji Parsial 66
Tabel 4.20 Uji Hosmer dan Lemeshow..... 73
Tabel 4.21 Prediksi Algoritma Regresi Logistik Biner..... 79
Tabel 4.22 *Confusion Matrix* Algoritma Regresi Logistik Biner 80
Tabel 4.23 *Accuracy*, *precision*, *recall*, *fscore* seluruh model regresi logistik..... 81
Tabel 4.24 Metode *Ensemble*..... 82
Tabel 4.25 *Confusion Matrix* Metode *Ensemble*..... 83
Tabel 4.26 *Accuracy*, *precision*, *recall*, *fscore* seluruh model *Ensemble* 84
Tabel 4.27 Perbandingan Tingkat Ketepatan Prediksi Klasifikasi 85

DAFTAR GAMBAR

Gambar 2.1 Algoritma *Random Forest* 24
Gambar 4.1 Diagram Batang Dari Diskritisasi Data 44
Gambar 4.2 Diagram Ringkasan Data Variabel *Independent* 45
Gambar 4.3 Pohon Keputusan *Node* Cabang 1.1.1.1 51
Gambar 4.4 Pohon Keputusan (1) *Node* Cabang 1.1.1.1 59

DAFTAR LAMPIRAN

Lampiran 1. Hasil Pohon Keputusan *Decision Tree* 92
Lampiran 2. Hasil Pohon Keputusan *Random Forest*..... 122
Lampiran 3. Tabel *Chi Square* 126
Lampiran 4. Hasil Prediksi..... 127

BAB I

PENDAHULUAN

1.1 Latar Belakang

Kegiatan pemberian kredit termasuk dalam kegiatan usaha yang memiliki risiko tinggi serta berpengaruh pada keberlangsungan perusahaan juga lembaga keuangan lainnya. Di dalam kegiatan perkreditan sering terjadi kredit bermasalah atau disebut juga kredit macet yang disebabkan karena gagalnya pengembalian sejumlah pinjaman yang diberikan kepada nasabah sesuai dengan waktu tempo yang sudah disepakati. Masalah pemberian kredit ini sebenarnya dapat diatasi, salah satunya yaitu dengan melakukan identifikasi dan prediksi pada calon nasabah sebelum memberikan kredit (Wulan, Bettiza & Hayaty, 2017).

Dataset risiko kredit yang digunakan untuk mengidentifikasi dan memprediksi terkadang memiliki suatu permasalahan yaitu ketidakseimbangan kelas (*imbalance class data*). Data pada kasus nyata penilaian kredit umumnya memiliki kelas respon yang tidak seimbang karena kejadian tidak gagal bayar jauh lebih banyak dibandingkan gagal bayar. Kelas respon tidak seimbang menyebabkan hasil prediksi akan akurat hanya pada satu kelas tertentu yaitu kelas dengan respon terbanyak (Syukron, Santoso & Widiharih, 2020).

Permasalahan *imbalance class data* dapat diselesaikan dengan metode umum yang biasanya digunakan yaitu metode *sampling* atau *resampling*. Menerapkan *resampling* dapat membuat data yang *imbalance* semakin kecil dan klasifikasi dapat dilakukan dengan lebih tepat. Beberapa metode *resampling* dengan tujuan untuk menangani ketidakseimbangan kelas yaitu *Random Over*

Sampling, *Random Under Sampling*, ADASYN, SMOTE, dan *Random Over-Under Sampling* (*Combine Sampling*). Metode *resampling* yang digunakan dalam mengolah data risiko kredit ini adalah *Random Over-Under Sampling*. Menggabungkan kedua metode pengambilan sampel acak dapat menghasilkan peningkatan kinerja secara keseluruhan. Konsep *Random Over-Under Sampling* dilakukan dengan menyeimbangkan jumlah distribusi data dengan meningkatkan jumlah data kelas minor (*oversampling*) dan mengurangi data mayor (*undersampling*) (Sastrawan, Baizal & Bijaksana, 2010).

Adanya distribusi kelas yang tidak seimbang dapat mempengaruhi kinerja algoritma klasifikasi, karena algoritma klasifikasi mengasumsikan bahwa distribusi kelas dalam *dataset* relatif seimbang. Hal ini tentu saja dapat menyebabkan risiko terjadinya kesalahan klasifikasi (*misclassification*) terhadap *dataset* yang akan berakibat pada kinerja algoritma klasifikasi menjadi tidak optimal (Pristyanto, 2019).

Wahyuningsih dan Utari (2018) mengemukakan bahwa klasifikasi adalah metode pembelajaran data untuk memprediksi nilai dari sekelompok variabel. Algoritma klasifikasi banyak yang berfokus pada pengklasifikasi tunggal (*single classifier*), tetapi sebenarnya dengan menggunakan *Ensemble* yaitu metode yang menggabungkan beberapa klasifikasi dapat meningkatkan ketepatan (Da Silva, Hruschka & Hruschka, 2014). Karena setiap algoritma *classifier* memiliki keunggulan masing-masing, dan apabila beberapa algoritma *classifier* tersebut dikombinasikan dan dibandingkan dengan metode *Ensemble* tentu hasil yang dicapai akan lebih baik lagi serta akan memiliki kemungkinan mendapatkan hasil

klasifikasi yang lebih akurat. Beberapa algoritma *classifier* yang digunakan secara umum adalah algoritma *Decision Tree*, *Random Forest*, dan *Logistic Regression*.

Decision Tree atau pohon keputusan merupakan salah satu algoritma dalam melakukan klasifikasi data untuk memprediksi risiko kredit. Ada beberapa jenis *Decision Tree* salah satunya adalah *Iterative Dichotomizer* 3 (ID3) yang merupakan algoritma paling dasar. Algoritma ID3 melakukan pencarian secara menyeluruh (*greedy)* pada semua kemungkinan pohon keputusan (Fatmandini, Saputra & Yulistria, 2020). Algoritma ID3 dapat memberikan data yang lengkap sehingga akan lebih mempermudah pihak bank/lembaga keuangan untuk menentukan penilaian kelayakan kredit pada nasabah.

Algoritma lain yang bisa digunakan dalam mengolah data untuk memprediksi risiko kredit adalah *random forest*. *Random forest* adalah kumpulan dari pohon keputusan yang berkerja menjadi suatu gabungan yang fungsional. Algoritma *Random forest* menggunakan pohon keputusan yang tidak berkorelasi, sehingga algoritma *Random forest* memberikan hasil yang lebih baik daripada model individu lainnya. Kesalahan prediksi dalam satu pohon keputusan dapat ditutupi oleh kebenaran yang diperoleh dari pohon keputusan lainnya, selama arah pembuatan pohon keputusan itu benar (Sanjaya dkk, 2020).

Pengolahan data dalam memprediksi risiko kredit yang dilakukan dengan algoritma *decision tree* atau *random forest* bisa dilakukan juga dengan regresi logistik. Regresi Logistik adalah analisis regresi yang digunakan untuk menganalisis hubungan pengaruh antara satu atau lebih variabel *independent* terhadap satu variabel *dependent* yang bersifat kategorik (nominal atau ordinal)

dengan variabel *independent* yang bersifat kontinu ataupun kategorik. Menganalisis data menggunakan regresi logistik bertujuan untuk mendapatkan model yang terbaik dan sederhana, namun model tersebut dapat menjelaskan hubungan antara hasil variabel *dependent* dengan variabel *independent*. Dinamakan sebagai regresi logistik biner karena variabel *dependent* berupa biner atau bersifat dikotomi yaitu skala pengukuran nominal dengan dua kategori (Hosmer & Lemeshow, 2000).

Sudah ada beberapa penelitian sebelumnya dengan menggunakan data yang sama dan berbagai macam metode yang dilakukan. Penelitian yang dilakukan oleh Ajay, Venkatesh & Gracia pada tahun 2016 yaitu pengujian model menggunakan fitur algoritma Info Gain dan *Correlation-Based Feature Selection Best First* (CFS-*Best First*) dalam algoritma *BayesNet*, *Stacking*, *Naive Bayes*, *Random Forest*, *Random Tree*, ZeroR, *Instance Based* KNN (IBK), dan Social Media Optimization (SMO). Algoritma klasifikasi yang memiliki kinerja terbaik adalah *Random Forest*, *Random Tree* dan IBK dengan hasil akurasi 81,50%.

Penelitian lainnya dilakukan oleh Maruf Pasha dan tim pada tahun 2017 dengan menggunakan algoritma FLDA, J48, *Logistic Regression*, *Naive Bayes*, MLP dan IBK. Hasil penelitian menunjukkan bahwa algoritma tertinggi dengan MLP yaitu akurasi 81,70% (Ipin & Gata, 2019). Penelitian selanjutnya dilakukan oleh Ipin Sugiyarto dan Windu Gata pada tahun 2018-2019 yaitu melakukan pengujian menggunakan *neural network* dengan *feature selection* PCA dan dioptimasi dengan *Particle Swarm Optimize* (PSO) untuk memprediksi persetujuan kartu kredit. Hasil penelitian ini membuktikan dengan method *Neural Network* + PCA + PSO terbukti memiliki akurasi tertinggi yaitu 82,67%.

Penelitian mengenai *Random Over-Under Sampling* yang dilakukan oleh Akhmad Syukron dan Agus Subekti pada tahun 2018 berjudul "Penerapan Metode *Random Over-Under Sampling* dan *Random Forest* Untuk Klasifikasi Penilaian Kredit". Hasil pengujian menunjukkan bahwa metode Random Forest memiliki nilai akurasi yang lebih baik yaitu sebesar 76%, sedangkan klasifikasi dengan penerapan metode *Random Over-Under Sampling Random Forest* dapat meningkatkan kinerja akurasi sebesar 14,1% dengan nilai akurasi sebesar 90,1 %.

Penelitian menggunakan metode *Ensemble* yaitu oleh Novianti (2019) dengan judul "Prediksi Status Berlangganan Klien Bank Menggunakan Algoritma *Naive Bayes*, C4.5, dan KNN Berbasis *Ensemble Classifier*". Hasil evaluasi dengan ROC *curve* menunjukan penggabungan algoritma menggunakan *ensemble vote* menghasilkan 94,20%.

Berdasarkan penelitian-penelitian sebelumnya dan kelebihan dari beberapa metode yang sudah dijabarkan, maka penulis bertujuan untuk melakukan penelitian mengenai prediksi tingkat risiko kredit dengan mengimplementasikan *Random Over-Under Sampling* pada metode *Ensemble* menggunakan algoritma *Decision Tree* ID3, *Random Forest*, dan Regresi Logistik Biner.

1.2 Rumusan Masalah

Rumusan masalah pada penelitian ini adalah sebagai berikut:

1. Berapa tingkat ketepatan klasifikasi dalam memprediksi risiko kredit dengan mengimplementasikan *Random Over-Under Sampling* pada klasifikasi tunggal menggunakan algoritma *Decision Tree*, *Random Forest* dan Regresi Logistik Biner?

2. Berapa tingkat ketepatan klasifikasi dalam memprediksi risiko kredit dengan mengimplementasikan *Random Over-Under Sampling* pada metode *Ensemble* menggunakan algoritma *Decision Tree*, *Random Forest* dan Regresi Logistik Biner?
3. Bagaimana perbandingan tingkat ketepatan klasifikasi dalam memprediksi risiko kredit pada semua metode pengklasifikasi yaitu pada *Decision Tree*, *Random Forest*, Regresi Logistik Biner, dan *Ensemble?*

1.3 Batasan Masalah

Batasan masalah dalam penelitian ini adalah sebagai berikut:

1. Data yang digunakan dari *dataset approval credit card* UCI *Repository*. Data ini terdiri dari 23 variabel *independent* yaitu *limit ball*, *sex*, *education*, *marriage*, *age*, *payment* 1 – 6, *bill amount* 1 – 6 dan *payment amount* 1 – 6 dan variabel *dependent* yaitu *default payment*.
2. Data ini berjumlah 30000 data, dengan validasi data menggunakan *repeated k-fold cross validation* yaitu dengan 10 *folds* dan 5 *repeated*.
3. Tingkat ketepatan klasifikasi pada penelitian ini dibatasi hanya pada nilai *Accuracy*, *Precision*, *Recall*, dan *Fscore*.

1.4 Tujuan Penelitian

Tujuan dari penelitian ini adalah sebagai berikut.

1. Menghitung tingkat ketepatan klasifikasi dalam memprediksi risiko kredit dengan mengimplementasikan *Random Over-Under Sampling*

pada klasifikasi tunggal menggunakan algoritma *Decision Tree*, *Random Forest* dan Regresi Logistik Biner?

2. Menghitung tingkat ketepatan dalam memprediksi risiko kredit dengan mengimplementasikan *Random Over-Under Sampling* pada metode *Ensemble* menggunakan algoritma *Decision Tree*, *Random Forest* dan Regresi Logistik Biner.
3. Membandingkan tingkat ketepatan klasifikasi dalam memprediksi risiko kredit pada semua metode pengklasifikasi yaitu pada *Decision Tree*, *Random Forest*, Regresi Logistik Biner, dan *Ensemble*.

1.5 Manfaat Penelitian

Manfaat dari penelitian ini adalah sebagai berikut:

1. Sebagai media pembelajaran dan menambah pengetahuan bagi penulis serta pembaca dalam memprediksi risiko kredit dengan mengimplementasikan *Random Over-Under Sampling* pada metode *Ensemble* menggunakan algoritma *Decision Tree*, *Random Forest* dan Regresi Logistik Biner.
2. Sebagai bahan referensi dan menambah pengetahuan mengenai permasalahan data yang memiliki *imbalance class data* khususnya dengan melakukan *Random Over-Under Sampling*.
3. Membantu memberikan informasi kepada perusahaan ataupun lembaga keuangan dalam memprediksi risiko kredit dari seorang nasabah.

DAFTAR PUSTAKA

Ahmed *et al*. (2019). Stroke prediction using distributed machine learning based on apache spark. *International Journal of Advanced Science and Technology*, *28*(15), 89–97.

Ajay, Venkatesh, A., & Gracia, S. (2016). Prediction of credit-card defaulters: A comparative study on performance of classifiers. *International Journal of Computer Applications*, *145*(7), 36–41. https://doi.org/10.5120/ijca2016910702

Aradea dkk. (2011). Penerapan decision tree untuk penentuan pola data penerimaan mahasiswa baru. *Jurnal Penelitian Sitrotika*, *7*(1).

Aziz, F. (2017). *Penilaian kredit menggunakan ensemble logistic regression dengan metode boosting.* Skripsi Program Studi Teknik Elektro Universitas Hasanuddin Makassar.

Azwar. (2017). Klasifikasi pengambilan keputusan permohonan kredit dengan metode ID3 pada bank BPR Kepri Bintan Tanjungpinang. *Jurnal Umrah.*

Baiya, & Fernos, J. (2019). Analisis faktor-faktor penyebab kredit macet pada Bank Nagari cabang Siteba. 1–18. https://doi.org/10.31227/osf.io/4xuks

Christian, T. M. (2014). Exploration of classification using NBTree for predicting students ' performance. *IEEE*, 0–5.

Da Silva, N. F. F., Hruschka, E. R., & Hruschka, E. R. (2014). Tweet sentiment analysis with classifier ensembles. *Decision Support Systems*, 1–31. https://doi.org/10.1016/j.dss.2014.07.003

Fatmandini, N. A., Saputra, R. A., & Yulistria, R. (2020). Komparasi criteria splitting pada algoritma iterative dichotomizer 3 (ID3) untuk klasifikasi kelayakan kredit. *Jurnal Informatika Dan Komputer*, *22*(1), 79–84. https://doi.org/https://doi.org/10.31294/p.v21i2

Han, J., Kamber, M., & Pei, J. (2012). *Data mining concepts and technique. Third edition* (3rd ed.).

Hastie *et al*. (2008). *The Elements of Statistical Learning Data Data Mining, Inference, and Prediction.* (2nd ed.). Springer-Verlag.

Himasta. (2021). *Regresi logistik biner dan aplikasinya* (pp. 1–16). http://himasta.unimus.ac.id/wp-content/uploads/2021/01/Regresi_Logistik_Biner.pdf

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

Ipin, S., & Gata, W. (2019). Perbandingan kinerja algoritma data mining prediksi persetujuan kartu kredit. *Faktor Exacta*, *12*(3), 180–192. https://doi.org/10.30998/faktorexacta.v12i3.4310

James *et al*. (2013). *An Introduction To Statistical Learning*.

Kemala, I., & Wijayanto, W. A. (2021). Perbandingan kinerja metode bagging dan non-ensemble machine learning pada klasifikasi wilayah di Indonesia menurut indeks pembangunan manusia development. *Jurnal Sistem dan Teknologi Informasi*, *9*(2), 269–275. https://doi.org/10.26418/justin.v9i2.44166

Khadijah, & Kusumaningrum, R. (2019). Ensemble classifier untuk klasifikasi kanker payudara. *IT Journal Research and Development (ITJRD)*, *4*(1), 61–71. https://doi.org/10.25299/itjrd.2019.vol4(1).3540 61 Ensemble

Kotimah, K. M., & Wulandari, P. S. (2014). Model regresi logistik biner stratifikasi pada partisipasi ekonomi perempuan di provinsi Jawa Timur. *Jurnal Sains Dan Seni Pomits*, *3*(1), 2337–3520.

Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling* (1st ed.). Berlin: Springer.

Kusyanti, A. (2019). Metode ensemble classifier untuk mendeteksi jenis attention deficit hyperactivity disorder (ADHD) pada anak usia dini. *Jurnal Teknologi Informasi Dan Ilmu Komputer (JTIIK)*, *6*(3), 301–308. https://doi.org/10.25126/jtiik.201961313

Mutmainah, S. (2021). Penanganan imbalance data pada klasifikasi kemungkinan penyakit stroke. *Jurnal SNATi*, *1*(1), 10–16.

Novianti, D. (2019). *Prediksi status berlangganan klien bank menggunakan algoritma naïve bayes, C4.5, dan KNN berbasis ensemble classifier*. In *Sustainability (Switzerland)*. Skripsi Program Studi Magister Ilmu Komputer STMIK Nusa Mandiri Jakarta.

Pato, S. (2013). Analisis pemberian kredit mikro pada Bank Syariah Mandiri cabang Manado. *EMBA*, *1*(4), 875–885.

Pristyanto, Y. (2019). Penerapan metode ensemble untuk meningkatkan kinerja algoritme klasifikasi pada imbalanced dataset. *Jurnal Teknoinfo*, *13*(1), 11–16. https://doi.org/10.33365/jti.v13i1.184

Priyambodo, Y. D. (2010). *Risiko kredit ditinjau dari jenis kredit dan jaminan kredit studi kasus di KBPR bank pasar PATMA Klaten*. Skripsi Jurusan Akuntansi Universitas Sanata Dharma Yogyakarta.

Rajagukguk, N., Ispriyanti, D., & Wilandari, Y. (2015). Perbandingan metode klasifikasi regresi logistik biner dan naive bayes pada status pengguna KB di

kota Tegal tahun 2014. *Jurnal Gaussian*, *4*(2), 365–374.

Ratnawati, E., & Sunarko. (2008). Evaluasi kinerja fasilitas iradiasi sistem rabbit menggunakan bahan acuan standard dengan metode AAN. *Buletin Pengelolaan Reaktor Nuklir*, *5*(2), 49–55.

Rifqo, M. H., & Arzi, T. (2017). Implementasi algoritma C4.5 untuk menentukan calon debitur dengan mengukur tingkat risiko kredit pada Bank BRI cabang Curup. *Jurnal Pseudocode*, *3*(2), 83–90. https://doi.org/10.33369/pseudocode .3.2.83 -90

Rohmi, A. L. (2017). *Analisis regresi logistik multinomial pada jenis pelanggaran lalu lintas di kota Surabaya*. Skripsi Fakultas Vokasi Institut Teknologi Sepuluh Nopember Surabaya.

Roihan, A., Sunarya, P. A., & Rafika, A. S. (2020). Pemanfaatan machine learning dalam berbagai bidang : Review paper. *Indonesian Journal on Computer and Information Technology (IJCIT)*, *5*(1), 75–82.

Saifudin, A. (2018). Metode data mining untuk seleksi calon mahasiswa pada penerimaan mahasiswa baru di Universitas Pamulang. *Jurnal Teknologi*, *10*(1), 25–36.

Sanjaya dkk. (2020). Prediksi kelalaian pinjaman bank menggunakan random forest dan adaptive boosting. *Jurnal Teknik Informatika Dan Sistem Informasi*, *6*(1), 50–60. https://doi.org/10.28932/jutisi.v6i1.2313

Sastrawan, A. S., Baizal, Z. A., & Bijaksana, M. A. (2010). Analisis pengaruh metode combine sampling dalam churn prediction untuk perusahaan telekomunikasi. *Seminar Nasional Informatika 2010 (SemnasIF 2010)*, A14–A22.

Sidik, Z. (2019). Klasifikasi kelancaran kredit furniture menggunakan algoritma k-nearest neighbor berbasis forward selection. https://doi.org/10.31227/osf. io/huvsa

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing and Management*, *45*(4), 427–437. https://doi.org/10.1016/j.ipm.2009.03.002

Susanto dkk. Penerapan perhitungan metode decision tree menggunakan algoritma iterative dichotomiser 3 (ID3) berbasis website. *Jurnal Sains Indonesia*, *1*(2), 59–68.

Syukron, A., & Subekti, A. (2018). Penerapan metode random over-under sampling dan random forest untuk klasifikasi penilaian kredit. *Jurnal Informatika*, *5*(2), 175–185. https://doi.org/10.31311/ji.v5i2.4158

Syukron, M., Santoso, R., & Widiharih, T. (2020). Perbandingan metode smote

random forest dan smote xgboost untuk klasifikasi tingkat penyakit hepatitis C pada imbalance class data. *Jurnal Gaussian*, *9*(3), 227–236. https://ejournal3.undip.ac.id/index.php/gaussian/

Tampil, Y., Komaliq, H., & Langi, Y. (2017). Analisis regresi logistik untuk menentukan faktor-faktor yang mempengaruhi indeks prestasi kumulatif (IPK) mahasiswa FMIPA Universitas Sam Ratulangi Manado. *D'CARTESIAN*, *6*(2), 56–62. https://doi.org/10.35799/dc.6.2.2017.17023

Triscowati, D. W., & Jayanti, L. D. (2021). Penilaian kredit pada data tak seimbang menggunakan random forest. *Jurnal Ilmiah Komputasi dan Statistika*, *1*(1), 25–31.

W., Y. Y. (2007). Perbandingan performansi algoritma decision tree C5.0, CART, dan CHAID: kasus prediksi status resiko kredit di bank X. *Seminar Nasional Aplikasi Teknologi Informasi 2007 (SNATI 2007)*, B59–B62.

Wahyuningsih, S., & Utari, D. R. (2018). Perbandingan metode k-nearest neighbor , naive bayes dan decision tree untuk prediksi kelayakan pemberian kredit. *Konferensi Nasional Sistem Informasi 2018*, 619–623.

Widiastuti, J. (2018). *Klasifikasi pembiayaan warung mikro menggunakan metode random forest dengan teknik sampling kelas imbalanced (Studi Kasus: Data Nasabah Pembiayaan Warung Mikro Bank Syariah Mandiri KC Jambi)*. Skripsi Program Studi Statistika FMIPA Universitas Islam Indonesia Yogyakarta.

Wulan, S. T., Bettiza, M., & Hayaty, N. (2017). Optimasi seleksi fitur klasifikasi naïve bayes risiko kredit konsumen (Studi Kasus : PT. Finansia Multi Finance (KreditPlus) Tanjungpinang). *Jurnal Umrah*, 1–17.

Yeh, I. C., & Lien, C. hui. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, *36*(2 part 1), 2473–2480. https://doi.org/10.1016/j.eswa.2007.12.020

Zhang, Y., & Yang, Y. (2015). Cross-validation for selecting a model selection procedure. *Journal of Econometrics*, *187*(1), 95–112. https://doi.org/10.1016/j.jeconom.2015.02.006