PENILAIAN KEMAMPUAN PEMBAYARAN KREDIT DENGAN MENGGUNAKAN *MACHINE LEARNING LOGISTIC REGRESSION* DAN *RANDOM FOREST CLASSIFIER* PADA HOME CREDIT

Skripsi

Oleh

AMANDA HASNA CAHYANA

NPM 2015061073

FAKULTAS TEKNIK

UNIVERSITAS LAMPUNG

BANDAR LAMPUNG

2024

PENILAIAN KEMAMPUAN PEMBAYARAN KREDIT DENGAN MENGGUNAKAN *MACHINE LEARNING LOGISTIC REGRESSION* DAN *RANDOM FOREST CLASSIFIER* PADA HOME CREDIT

Oleh
AMANDA HASNA CAHYANA

Skripsi

Sebagai Salah Satu Syarat untuk Mencapai Gelar
SARJANA TEKNIK

Pada

Jurusan Teknik Elektro
Fakultas Teknik Universitas Lampung

FAKULTAS TEKNIK

UNIVERSITAS LAMPUNG

BANDAR LAMPUNG

2024

ABSTRAK

PENILAIAN KEMAMPUAN PEMBAYARAN KREDIT DENGAN MENGGUNAKAN *MACHINE LEARNING LOGISTIC REGRESSION* DAN *RANDOM FOREST CLASSIFIER* PADA HOME CREDIT

Oleh

AMANDA HASNA CAHYANA

Perkembangan ekonomi global menyebabkan tingginya kompleksitas kebutuhan masyarakat. Lembaga keuangan hadir memberikan fasilitas kredit maupun pembiayaan untuk memenuhi kebutuhan masyarakat yang semakin kompleks. Namun, adanya kredit bermasalah dapat menjadi ancaman yang serius bagi lembaga keuangan. Teknik klasifikasi dalam *data mining* menjadi suatu solusi yang dapat digunakan untuk mengatasi kredit bermasalah. Berdasarkan permasalahan tersebut, penelitian ini mengembangkan model yang dapat memprediksi kemampuan nasabah dalam melakukan pembayaran kredit sehingga lembaga keuangan dapat terhindar dari terjadinya kredit bermasalah. Dalam penelitian ini, teknik *resampling* dengan *SMOTE* digunakan untuk melihat pengaruh *sampling* dalam menangani ketidakseimbangan kelas dan melakukan penilaian kredit. Hasil penelitian menunjukkan, model yang dibangun dengan menggunakan *SMOTE* memiliki *AUC* yang lebih baik dibandingkan dengan model tanpa *SMOTE*. Dari dua algoritma *machine learning logistic regression* dan *random forest* diperoleh hasil bahwa model *random forest* dengan *SMOTE* memiliki kinerja paling baik dengan nilai *accuracy* sebesar 90%, *precision* sebesar 92%, *recall* sebesar 88%, *F1-score* sebesar 90%, dan nilai *AUC* sebesar 0.97. Berdasarkan model terbaik tersebut didapatkan sepuluh *importance features* yang berpengaruh dalam proses penilaian kemampuan pembayaran kredit, yaitu skor yang dinormalisasi dari sumber data eksternal, rentang waktu perubahan nomor nasabah, jumlah pembayaran cicilan sebelumnya, usia nasabah, waktu registrasi, rentang waktu pengajuan kredit di biro kredit, rentang waktu perubahan dokumen identitas, waktu pembaruan informasi di biro kredit, dan lama nasabah bekerja. Selain itu, penelitian ini menghasilkan visualisasi melaui *dashboard* yang dapat digunakan untuk meningkatkan proses penilaian kemampuan pembayaran kredit.

Kata kunci : Kemampuan Pembayaran Kredit, Prediksi, *Logistic Regression*, *Random forest*, *SMOTE*.

ABSTRACT

ASSESSMENT OF CREDIT PAYMENT CAPABILITY USING MACHINE LEARNING LOGISTIC REGRESSION AND RANDOM FOREST CLASSIFIER ON HOME CREDIT

By

AMANDA HASNA CAHYANA

Global economic development has led to the high complexity of society's needs. Financial institutions are here to provide credit and financing facilities to meet the increasingly complex needs of society. However, the existence of non-performing loans can pose a serious threat to financial institutions. Classification techniques in data mining are a solution that can be used to overcome problem loans. Based on these problems, this research develops a model that can predict customers' ability to make credit payments so that financial institutions can avoid problematic credit. In this research, the SMOTE resampling technique is used to see the effect of sampling in dealing with class imbalance and conducting credit assessments. The research results show that the model built using SMOTE has better AUC compared to the model without SMOTE. From the two machine learning algorithms, logistic regression and random forest, the results show that the random forest model with SMOTE has the best performance with an accuracy value of 90%, precision of 92%, recall of 88%, F1-score of 90%, and AUC value of 0.97. Based on the best model, ten important features were obtained that influence the process of assessing credit repayment capabilities, namely the normalized score from external data sources, the period for changing customer numbers, the number of previous installment payments, the customer's age, registration time, the period for applying for credit at the credit bureau, the period for changing identity documents, the time for updating information at the credit bureau, and the length of time the customer has worked. In addition, this research produces visualizations via dashboards that can be used to improve the process of assessing credit repayment capabilities.

Keywords: Credit Payment Capability, Prediction, Logistic Regression, Random Forest, SMOTE.

Judul Skripsi	:	**PENILAIAN KEMAMPUAN PEMBAYARAN KREDIT DENGAN MENGGUNAKAN *MACHINE LEARNING LOGISTIC REGRESSION* DAN *RANDOM FOREST CLASSIFIER* PADA HOME CREDIT**
Nama Mahasiswa	:	*Amanda Hasna Cahyana*
Nomor Pokok Mahasiswa	:	2015061073
Program Studi	:	Teknik Informatika
Fakultas	:	Teknik

MENYETUJUI

1. Komisi Pembimbing

Pembimbing Utama	Pembimbing Pendamping
Ir. Muhamad Komarudin, S.T., M.T.	**Ir. Titin Yulianti, S.T., M.Eng.**
NIP. 196812071997031006	NIP. 198807092019032015

2. Mengetahui

Ketua Jurusan Teknik Elektro	Ketua Program Studi Teknik Informatika
Herlinawati, S.T., M.T.	**Yessi Mulyani, S.T., M.T.**
NIP. 197103141999032001	NIP. 197312262000122001

MENGESAHKAN

1. Tim Penguji

   Ketua : **Ir. Muhamad Komarudin, S.T., M.T.**

   Sekretaris : **Ir. Titin Yulianti, S.T., M.Eng.**

   Penguji : **Yessi Mulyani, S.T., M.T.**

2. Dekan Fakultas Teknik

Dr. Eng. Ir. Helmy Fitriawan, S.T., M.Sc.

NIP. 197509282001121002

Tanggal Lulus Ujian Skripsi : **19 Januari 2024**

SURAT PERNYATAAN

Saya yang bertanda tangan di bawah ini, menyatakan bahwa skripsi saya dengan judul "Penilaian Kemampuan Pembayaran Kredit dengan Menggunakan *Machine Learning Logistic Regression* dan *Random Forest Classifier* pada Home Credit" dibuat oleh saya sendiri. Semua hasil yang tertuang dalam skripsi ini telah mengikuti kaidah penulisan karya ilmiah Universitas Lampung. Apabila di kemudian hari terbukti bahwa skripsi ini merupakan salinan atau dibuat oleh orang lain maka saya bersedia menerima sanksi sesuai dengan ketentuan hukum atau akademik yang berlaku.

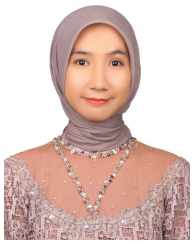
Bandar Lampung, 26 Januari 2024

Pembuat pernyataan,

Amanda Hasna Cahyana

NPM 2015061073

RIWAYAT HIDUP



Penulis dilahirkan di Kotabumi pada tanggal 27 Maret 2002. Penulis merupakan anak ketiga dari pasangan Bapak Jauhari Effendi, B. Sc dan Ibu Gurtilia, S.E.

Penulis menempuh pendidikannya di Taman Kanak – Kanak (TK) Putri Kotabumi. Pendidikan Sekolah Dasar (SD) penulis tempuh di SD Negeri 04 Tanjung Aman Kotabumi pada tahun 2014, Sekolah Menengah Pertama (SMP) di SMP Negeri 7 Kotabumi pada tahun 2017, dan Sekolah Menengah Atas (SMA) di SMA Negeri 9 Bandar Lampung pada tahun 2020. Pada tahun 2020, penulis terdaftar sebagai mahasiswa Program Studi Teknik Informatika, Jurusan Teknik Elektro, Fakultas Teknik Universitas Lampung melalui jalur Seleksi Bersama Masuk Perguruan Tinggi Negeri (SBMPTN). Selama menjadi mahasiswa, penulis aktif mengikuti beberapa kegiatan, antara lain sebagai berikut :

1. Menjadi anggota biasa Himpunan Mahasiswa Teknik Elektro Universitas Lampung, Departemen Pendidikan dan Pengembangan Diri Divisi Pendidikan periode 2020/2021 serta Departemen Pengembangan Keteknikan Divisi Penelitian dan Pengembangan periode 2021/2022.
2. Membantu dalam penelitian, pembuatan, dan penerbitan jurnal pada acara *International Conference on Converging Technology in Electrical and Information Engineering* (*ICCTEIE*) 2023.
3. Menjadi sekretaris koordinator *sponsorship* pada acara *Electrical Engineering in Action* tahun 2022.
4. Menjadi asisten Laboratorium Teknik Digital Universitas Lampung pada tahun 2022 sampai tahun 2023.
5. Mengikuti program Studi Independen Bersertifikat Kampus Merdeka *Batch* 3

pada posisi *Data Analytics* di Zenius Education pada Agustus sampai Desember tahun 2022.

6. Mengikuti program Magang Bersertifikat Kampus Merdeka *Batch* 4 pada posisi *Data Scientist* di PT Telkom Indonesia pada Februari sampai Juni tahun 2023.
7. Melaksanakan Kuliah Kerja Nyata selama 36 hari di Desa Way Napal, Kecamatan Krui Selatan, Kabupaten Pesisir Barat, Lampung, Indonesia

MOTTO

"……Allah's help is all-sufficient for us and Allah is the best protector." (QS 3 ; 173)

" Dan Dia memberikan kepadamu segala apa yang kamu mohonkan kepada-Nya. Dan jika kamu menghitung nikmat Allah, tidaklah dapat kamu menghinggakannya..... " (QS 14 ; 34)

"Nothing lasts forever. Everything is momentary. They all flow away. And that's not always a bad thing"—2521

"It's not about how much we lost. It's about how much we have left" —Tony Stark

" Dan bahwa sesungguhnya tidak ada (balasan atau hasil) bagi seseorang manusia melainkan apa yang telah diusahakannya " (QS 53 ; 39)

"..and Allah Closer to you more than your jugular vein.." (QS 50 ; 16)

PERSEMBAHAN

Bismillahirrahmannirrahiim

Puji syukur kepada Allah SWT yang telah memberikan limpahan rahmat dan hidayah-Nya. Tak lupa shalawat dan salam senantiasa tercurah kepada Nabi Muhammad SAW teladan yang menginspirasi setiap langkah perjalanan hidup.

Skripsi ini merupakan persembahan sederhana untuk mereka yang senantiasa mendoakan dan selalu sabar menunggu di rumah;
Mama, Papa, Wanda, Yunda & Dinda...

Juga

Untuk seluruh keluarga, teman-teman, dan pihak lain yang telah berkontribusi dalam hidupku...

Dengan tulus persembahan ini diberikan sebagai bentuk ucapan terima kasih kepada semua pihak yang memberikan dukungan, bimbingan, dan doa yang senantiasa mengiringi perjalanan ini.

SANWACANA

Puji syukur Penulis panjatkan kepada Allah SWT, yang telah memberikan rahmat dan hidayah–Nya sehingga skripsi ini dapat diselesaikan. Shalawat serta salam semoga selalu tercurahkan kepada Nabi Muhammad S.A.W.

Skripsi yang berjudul "**Penilaian Kemampuan Pembayaran Kredit dengan Menggunakan *Machine Learning Logistic Regression* dan *Random Forest Classifier* pada Home Credit**" merupakan salah satu syarat yang digunakan untuk memperoleh gelar sarjana Teknik pada Jurusan Teknik Elektro, Fakultas Teknik, Universitas Lampung. Dalam menyelesaikan skripsi ini penulis banyak mendapatkan masukan, bantuan, dorongan, saran, bimbingan dan kritik dari berbagai pihak. Oleh karena itu, pada kesempatan ini dengan segala kerendahan hati penulis ingin menyampaikan rasa terima kasih kepada :

1. Bapak Dr. Eng. Helmy Fitriawan, S.T., M.Sc. selaku Dekan Fakultas Teknik Universitas Lampung;
2. Ibu Herlinawati, S.T., M.T. selaku Ketua Jurusan Teknik Elektro Universitas Lampung;
3. Ibu Yessi Mulyani, S.T., M.T. selaku Ketua Program Studi Teknik Informatika Universitas Lampung dan dosen penguji atas kesediaannya meluangkan waktu, memberikan masukan, kritik, dan saran yang bermanfaat dalam skripsi ini;
4. Bapak Ir. M. Komarudin, S.T., M.T. selaku dosen pembimbing utama dan pembimbing akademik atas kesediaannya meluangkan waktu, membimbing, memotivasi, dan memberikan ilmu selama pengerjaan skripsi dan perkuliahan;
5. Ibu Ir. Titin Yulianti, S.T., M.Eng. selaku dosen pembimbing kedua atas kesediaannya meluangkan waktu, memberikan bimbingan dan bantuan, serta motivasi kepada penulis selama proses pengerjaan skripsi;
6. Seluruh jajaran dosen dan civitas Jurusan Teknik Elektro Universitas Lampung

atas segala bantuan, pengetahuan, dan pengalaman yang diberikan selama penulis menjalani proses perkuliahan;

7. Mbak Rika selaku admin program studi teknik informatika yang telah banyak membantu penulis dalam urusan administrasi selama perkuliahan
8. Kedua orang tuaku Mama, Gurtilia, S.E. dan Papa, Jauhari Effendi, B. Sc. sebagai kedua pelita dalam hidupku. Terima kasih atas segala keikhlasan doa, kasih sayang, dukungan, dan pengorbanan yang tak terhingga dan tak ada habisnya;
9. Kakak dan Adikku tersayang, dr. Atika Marcherya (Wanda), Chintia Dwi (Yunda), Adinda Husna (Dinda) yang selalu memberikan doa, keceriaan, canda tawa, semangat, dan menjadi tempatku berkeluh kesah selama ini;
10. Teman-teman "UTS Gabut" (Feny, Michel, Sherly, Fajar, Hamzah, Handrio, Niki) yang selalu memotivasi, menemani, membantu dan berjuang bersama penulis dalam suka maupun duka selama masa perkuliahan ini;
11. Teman–teman "Anbuc" (Divia, Farah, Mazi, Niken, Triana) dan "Ciwi-ciwi" (Fatia, Salsabila) yang selalu setia menyemangati, mendukung, mengapresiasi, dan menghibur penulis sejak masa sekolah hingga saat ini;
12. Teman–teman konsentrasi kecerdasan buatan dan asisten laboratorium teknik digital yang telah bertukar pikiran dan berbagi informasi berharga bagi penulis.
13. Teman–teman Jurusan Teknik Elektro 2020 atas perjuangannya dalam masa sulit dan bahagia, semoga kita semua dapat bermanfaat bagi setiap insan kehidupan.

Akhir kata, penulis menyadari bahwa skripsi ini masih jauh dari kesempurnaan. Akan tetapi, penulis berharap semoga skripsi ini dapat bermanfaat dan berguna bagi kita semua. *Aamiin Ya Rabbal Alamin*.

Bandar Lampung, 26 Januari 2024
Penulis,

Amanda Hasna Cahyana

DAFTAR ISI

Halaman

DAFTAR TABEL

Tabel Halaman

DAFTAR GAMBAR

Gambar Halaman

BAB I
PENDAHULUAN

1.1 Latar Belakang

Perkembangan ekonomi global akan mengakibatkan tingginya kebutuhan masyarakat terhadap keuangan dalam menghadapi perubahan yang terjadi. Dalam menghadapi perubahan tersebut, keuangan sangat diperlukan untuk memenuhi kebutuhan masyarakat mulai dari kebutuhan pokok, seperti sandang, pangan dan papan hingga kebutuhan sekunder dan tersier. Kompleksnya kebutuhan tersebut menyebabkan semua kebutuhan tidak dapat terpenuhi secara tunai karena adanya keterbatasan keuangan. Oleh karena itu, lembaga keuangan hadir memberikan fasilitas kredit maupun pembiayaan untuk memenuhi kebutuhan masyarakat yang kompleks [1].

Kredit dapat diartikan sebagai kegiatan memperoleh barang atau pinjaman yang pembayarannya dapat dilakukan di kemudian hari dengan cara dicicil atau diangsur sesuai dengan perjanjian [2]. Kredit membantu masyarakat dengan menyediakan layanan pembiayaan keuangan bagi mereka yang ingin memperoleh suatu barang. Layanan pembiayaan tersebut disediakan oleh lembaga keuangan bukan bank yang sering disebut dengan lembaga pembiayaan. Berbeda dengan bank, penyaluran kredit di lembaga pembiayaan diberikan pada pihak ketiga dan bukan berupa uang tunai. Salah satu contoh lembaga pembiayaan adalah Home Credit.

Pemberian kredit pada nasabah dapat menjadi salah satu sumber pendapatan terbesar bagi suatu lembaga keuangan atau lembaga pembiayaan. Dengan penyaluran kredit suatu lembaga keuangan dapat memperoleh pendapatannya dari angsuran pokok dan bunga yang diberikan oleh setiap nasabah saat melakukan

pembayaran angsuran. Namun, kredit juga dapat menimbulkan kerugian bagi lembaga keuangan jika terdapat kredit bermasalah. Kredit bermasalah atau yang sering disebut juga dengan kredit macet merupakan suatu kondisi dimana pihak peminjam tidak mampu melunasi pinjaman kepada pihak yang memberi pinjaman dalam sesuai waktu yang telah ditentukan [3].

Berdasarkan data pada Statistik Perbankan Indonesia yang diterbitkan oleh Departemen Perizinan dan Informasi Perbankan, rasio jumlah kerugian yang terjadi akibat kredit bermasalah pada Februari 2023 secara keseluruhan sebesar 2,58%. Rasio kredit tersebut mengalami kenaikan sebesar 0,15% sejak Desember 2022. Menurut Peraturan Bank Indonesia Nomor 23/2/PBI/2021 rasio kredit bermasalah untuk total kredit atau pembiayaan secara bruto harus kurang dari 5% sehingga kenaikan rasio yang terjadi pada Februari 2023 tidak membuat perubahan yang cukup besar karena masih berada di ambang rasio kredit bermasalah yang baik. Namun, apabila rasio kredit bermasalah pada Februari 2022 hanya dihitung berdasarkan jenis orientasi penggunaan kredit konsumsi, nilai rasio kredit bermasalah mencapai 5,45% sehingga masuk kedalam kategori kredit bermasalah yang tidak baik. Nilai rasio tersebut naik sebesar 3.91% dibandingkan dengan rasio pada Desember 2022 yang hanya sebesar 1.54%.

Kredit bermasalah menjadi ancaman dengan risiko tinggi bagi lembaga keuangan yang tidak hanya berpengaruh terhadap pendapatan, tetapi juga berpengaruh terhadap keberlangsungan lembaga keuangan tersebut. Dengan demikian, dalam proses pemberian kredit dibutuhkan analisis untuk menilai apakah calon nasabah memiliki kemampuan untuk membayar kembali kredit yang akan diberikan karena tidak semua masyarakat yang mengajukan kredit dapat menerima pinjaman kredit. Pemberian kredit harus memperhatikan beberapa hal untuk menghindari kredit bermasalah yang dapat merugikan penyalur kredit. Salah satu tindakan yang dapat dilakukan oleh penyalur kredit untuk menyelesaikan kredit bermasalah adalah dengan memperhatikan data historis dari calon nasabah yang mengajukan kredit [4]. Namun, besarnya data historis calon nasabah menyebabkan keterbatasan analisis manual. Oleh karena itu, pemanfaatan teknik *data mining* dapat digunakan untuk mengatasi keterbatasan dalam menganalisis jumlah data yang besar.

Data mining merupakan proses penguraian kompleks dari sekumpulan data yang sebelumnya belum diketahui menjadi suatu informasi berpotensi secara implisit [5]. Terdapat beberapa teknik dalam *data mining* yang dapat digunakan untuk menemukan informasi dalam kumpulan data, seperti klasterisasi, regresi, dan klasifikasi. Klasifikasi menggunakan *supervised machine learning* dapat digunakan untuk memisahkan kemampuan calon nasabah dalam membayar kredit. Dalam melakukan klasifikasi terdapat beberapa algoritma *machine learning* yang dapat digunakan seperti *Naïve Bayes*, *Logistic Regression*, *Random Forest* dan lain-lain. Setiap algoritma *machine learning* memiliki kelemahan dan kelebihan. *Naïve Bayes* merupakan algoritma yang sederhana dengan biaya perhitungan lebih kecil [6], tetapi banyaknya fitur yang digunakan pada algoritma ini akan mempengaruhi nilai akurasi menjadi lebih kecil [7]. Kemudian, algoritma *Logistic Regression* dapat digunakan untuk klasifikasi yang memiliki variabel dikotomi atau hanya terdiri dari dua nilai [8] sehingga sesuai diimplementasikan untuk menilai kemampuan pembayaran kredit. Namun, pada data yang tidak seimbang, nilai akurasi *Logistic Regression* lebih rendah karena rentan terhadap *underfitting* [9]. Sementara algoritma *Random Forest* baik dalam mengatasi *noise*, *missing value*, dan *overfitting* dengan hasil akurasi yang maksimal meskipun data yang diolah berukuran besar [6], tetapi waktu latih data dengan algoritma ini lebih lama jika dibandingkan dengan algoritma lain [10].

Pada penelitian ini akan dilakukan proses klasifikasi menggunakan *machine learning*. Data yang digunakan diperoleh dari data Home Credit *Default Risk* pada website *Kaggle* [11]. Untuk mengatasi kelas tidak seimbang pada data tersebut, penelitian ini akan menggunakan proses *resampling,* yaitu dengan *SMOTE.* Berdasarkan penjelasan pendahuluan di atas maka akan dilakukan penelitian mengenai penilaian kemampuan nasabah dalam melakukan pembayaran pinjaman kredit dengan menggunakan *machine learning* pada Home Credit. Adapun algoritma *machine learning* yang digunakan pada penelitian ini adalah *logistic regression* dan *random forest*.

1.2 Perumusan Masalah

Berdasarkan latar belakang yang ada maka diperoleh rumusan masalah sebagai berikut:

1. Apakah proses penanganan ketidakseimbangan kelas pada data dengan metode *SMOTE* berpengaruh dalam melakukan penilaian kemampuan pembayaran kredit?
2. Bagaimana perfoma model yang dibangun dengan *machine learning logistic regression* dan *random forest* dalam melakukan penilaian kemampuan pembayaran kredit?
3. Apa saja atribut yang dapat mempengaruhi hasil prediksi penilaian kemampuan pembayaran kredit nasabah?
4. Bagaimana penerapan Google Looker Studio dalam melakukan visualisasi model *machine learning* terbaik untuk klasifikasi penilaian kemampuan pembayaran kredit?

1.3 Tujuan Penelitian

1.3.1 Tujuan Umum

Mengetahui hasil penilaian kemampuan calon nasabah Home Credit dalam melakukan pembayaran kredit.

1.3.2 Tujuan Khusus

1. Mengetahui pengaruh penanganan ketidakseimbangan kelas pada data dengan metode *SMOTE* dalam penilaian kemampuan pembayaran kredit.
2. Mengetahui perfoma model yang dibangun dengan *machine learning logistic regression* dan *random forest* sehingga dapat menentukan model terbaik dari dua metode yang digunakan untuk melakukan penilaian kemampuan pembayaran kredit.
3. Mengetahui atribut yang memengaruhi hasil prediksi penilaian kemampuan pembayaran kredit nasabah.
4. Melakukan pembuatan visualisasi atribut-atribut yang didapatkan dari model *machine learning* terbaik dengan Google Looker Studio.

1.4 Manfaat Penelitian

1.4.1 Bagi Peneliti

Hasil penelitian ini diharapkan dapat memberikan keilmuan dan pengalaman bagi peneliti mengenai klasifikasi untuk penilaian kemampuan pembayaran kredit dengan menggunakan *machine learning logistic regression* dan *random forest classifier* pada Home Credit.

1.4.2 Bagi Institusi

Hasil penelitian ini diharapkan dapat bermanfaat untuk meningkatkan jumlah penelitian terkait implementasi *machine learning logistic regression* dan *random forest classifier* dalam klasifikasi.

1.4.3 Bagi Masyarakat

Hasil penelitian ini diharapkan dapat meningkatkan pengetahuan dan memberikan informasi pada masyarakat mengenai faktor apa saja yang berpengaruh dalam penilaian kemampuan pembayaran kredit dan membantu pengambilan keputusan dalam menentukan kelayakan pemberian kredit.

1.5 Batasan Masalah

Adapun batasan masalah dalam penelitian ini adalah sebagai berikut :

1. Klasifikasi dilakukan dengan menggunakan algoritma *machine learning Logistic Regression* dan *Random Forest*.
2. Pembuatan visualisasi yang dibuat berupa *dashboard* dan tidak dilakukan pengujian fungsionalitas.
3. *Dashboard* hanya menampilkan atribut-atribut yang terdapat pada model *machine learning* terbaik.

1.6 Sistematika Penulisan

Adapun sistematika penulisan laporan yang digunakan pada penelitian ini adalah sebagai berikut :

BAB I : PENDAHULUAN

Pada bagian ini membahas latar belakang permasalahan terkait kondisi kredit yang ada di Indonesia. Kemudian, rumusan masalah yang akan dipecahkan dan tujuan penelitian yang akan dihasilkan dengan memanfaatkan *machine learning* dan *dashboard* untuk menyelesaikan permasalahan kredit. Selanjutnya membahas manfaat yang akan diperoleh setelah penelitian terlaksana, batasan masalah yang memperjelas fokus dan ruang lingkup penelitian, serta sistematika penulisan yang menjelaskan proses penelitian secara terstruktur.

BAB II : TINJAUAN PUSTAKA

Pada bagian ini membahas teori-teori penunjang yang digunakan sebagai sumber informasi untuk memahami beberapa hal yang terkait dengan penelitian, seperti penjelasan mengenai kredit, *data mining*, *machine learning*, tahapan *CRISP-DM*, *tools* yang digunakan, dan penelitian lain yang terkait dengan proses penilaian kemampuan nasabah yang akan dilakukan.

BAB III : METODOLOGI PENELITIAN

Pada bab ini membahas mengenai waktu dan tempat penelitian yang dilaksanakan, alat dan bahan yang berisi *software*, *hardware* dan data yang digunakan, serta langkah-langkah yang dilakukan untuk klasifikasi penilaian kemampuan pembayaran dengan menggunakan metode pengembangan *CRISP-DM*.

BAB IV : HASIL DAN PEMBAHASAN

Pada bab ini membahas mengenai hasil yang diperoleh setelah menerapkan metode pengembangan *CRISP-DM* dalam melakukan klasifikasi untuk menilai kemampuan pembayaran nasabah, mulai dari tahapan *business understanding*, *data understanding*, *data preparation*, *modeling*, *evaluation*, dan *deployment*.

BAB V : KESIMPULAN

Pada bab ini berisi tentang kesimpulan yang dapat diperoleh berdasarkan keseluruhan hasil dari penerapan proses *CRISP-DM* yang telah dilakukan serta saran-saran sebagai masukan untuk penelitian lebih lanjut di masa mendatang.

DAFTAR PUSTAKA

LAMPIRAN

BAB II
TINJAUAN PUSTAKA

2.1 Home Credit

Home Credit merupakan lembaga keuangan bukan bank yang didirikan pada tahun 1997 di Republik Ceko. Perusahaan ini telah melebarkan bisnisnya di sembilan negara, termasuk di Indonesia. Home Credit merupakan perusahaan yang menyediakan pembiayaan berbasis teknologi dengan keterbukaan akses terhadap layanan keuangan terutama kepada nasabah yang memiliki sedikit atau bahkan belum memiliki riwayat kredit. Home Credit memberikan pengalaman meminjam yang positif dan aman bahkan bagi nasabah yang belum memiliki rekening keuangan. Dengan adanya Home Credit, masyarakat dapat terhindar dari pemberi pinjaman yang tidak dapat dipercaya. Layanan yang diberikan oleh Home Credit juga sederhana, mudah, dan cepat sehingga tak heran perusahaan ini telah dipercaya lebih dari 135,4 juta nasabah dari berbagai negara [12].

Home Credit memberikan layanan pembiayaan kepada nasabah yang ingin melakukan pembelian berbagai produk untuk memenuhi kebutuhannya, seperti peralatan elektronik, rumah tangga, ataupun furnitur. Selain itu, perusahaan ini juga memberikan layanan pembiayaan multiguna yang dapat digunakan untuk pembayaran biaya pendidikan, pembangunan bahkan perjalanan. Home Credit terus berkembang untuk memenuhi kebutuhan secara terencana termasuk mengelola keuangan dan cicilan nasabah dengan baik. Untuk memastikan nasabah mendapat pengalaman pinjaman yang positif, Home Credit menggunakan berbagai data alternatif, seperti infromasi transaksional untuk menilai kemampuan nasabah dalam melakukan pembayaran pinjaman yang diberikan. Data-data tersebut

digunakan dalam menentukan nasabah yang dapat bertanggung jawab dalam melakukan pengembalian pinjaman kredit.

2.2 Kredit

2.2.1 Pengertian Kredit

Kredit berasal dari bahasa Latin, yaitu *Credere* yang artinya percaya. Oleh karena itu, kredit didasari oleh rasa kepercayaan dengan syarat-syarat yang telah disetujui bersama oleh pihak-pihak yang bersangkutan di dalamnya. Menurut Undang-Undang Nomor 10 Tahun 1998 Pasal 1 butir 11, kredit dapat didefinisikan sebagai penyediaan uang atau tagihan yang dapat dipersamakan dengan itu, berdasarkan perjanjian atau kesepakatan pinjam-meminjam antara bank dengan pihak lain yang mewajibkan pihak peminjam untuk melunasi utangnya setelah jangka waktu tertentu dengan pemberian bunga [13].

Beberapa ahli juga mengemukakan pendapatnya mengenai definisi kredit. H.M.A Savelberg menyatakan bahwa kredit merupakan dasar dari segala persekutuan dimana seseorang mempunyai hak untuk menuntut sesuatu sebagai jaminan telah memberikan pinjaman kepada orang lain, dengan tujuan mengembalikan apa yang telah diserahkannya. Sedangkan menurut J.A Levy, kredit adalah sejumlah uang yang secara sukarela dapat dipergunakan dengan bebas oleh penerima pinjaman untuk keuntungannya dengan kewajiban mengembalikan jumlah pinjaman tersebut di waktu yang akan datang [14]. Sementara itu, menurut Thomas secara umum kredit dapat didefinisikan sebagai suatu kepercayaan atas kemampuan pihak penerima pinjaman untuk membayar sejumlah uang pada masa yang akan datang [15].

Dari pengertian di atas dapat disimpulkan bahwa, kredit merupakan suatu kesepakatan untuk membayar sejumlah tagihan atas pinjaman yang digunakan secara bebas dengan jangka waktu tertentu beserta bunga yang telah disepakati.

2.2.2 Prinsip Pemberian Kredit

Dalam pemberian kredit, lembaga pemberi kredit harus yakin bahwa kredit yang diberikan kepada nasabah pasti akan dilunasi. Oleh karena itu, lembaga tersebut juga harus mempunyai prinsip pemberian kredit untuk menjamin pemberian kredit pada nasabah. Prinsip tersebut disebut juga sebagai konsep 5*C's* yang meliputi [16]:

1. *Character*

   Character menjelaskan moral, watak, ataupun sifat pribadi pada diri peminjam yang positif, kooperatif, dan juga penuh rasa tanggung jawab.
2. *Capacity*

   Capacity menjelaskan kemampuan calon peminjam berdasarkan kegiatan usaha yang dilakukan untuk melunasi pinjaman terhadap pemberi kredit.
3. *Capital*

   Capital menjelaskan jumlah dana atau modal yang dimiliki oleh calon peminjam.
4. *Collateral*

   Collacteral menjelaskan kumpulan jaminan yang diserahkan oleh peminjam sebagai tanggungan atas kredit yang diterima apabila peminjam gagal dalam memenuhi kewajibannya.
5. *Condition of economy*

   Condition of economy menjelaskan situasi politik, sosial, ekonomi, budaya dan lain-lain yang dapat mempengaruhi mempengaruhi kelancaran usaha dari peminjam.

2.2.3 Kredit Bermasalah

Kredit bermasalah atau yang sering disebut dengan *non performing loan* adalah suatu keadaan dimana penerima pinjaman atau nasabah tidak sanggup membayar sebagian atau seluruh kewajibannya kepada pemberi, seperti yang telah tertera dalam perjanjian kredit. Kredit bermasalah dapat terjadi karena hambatan dua pihak, yakni pihak pemberi kredit dalam melakukan analisis calon nasabah dan pihak nasabah yang dengan sengaja atau tidak sengaja mengabaikan kewajibannya dalam melakukan pembayaran [2]. Kredit bermasalah memberikan risiko yang

besar karena menyebabkan kerugian pihak pemberi kredit. Kerugian tersebut terjadi karena dana yang telah disalurkan beserta bunga yang akan diperoleh tidak diterima kembali sehingga akan menurunkan pendapatan lembaga penyalur kredit.

2.3 *Data Mining*

2.3.1 Pengertian *Data Mining*

Data mining merupakan proses yang menggunakan satu atau lebih teknik pembelajaran komputer (*machine learning*) untuk menganalisis dan memperoleh pengetahuan (*knowledge*) yang ada pada *database* sehingga dapat dihasilkan informasi yang sangat berharga secara otomatis [17]. Sementara itu, Mukhlifah mendefinisikan *data mining* sebagai dekomposisi kompleks dari kumpulan data menjadi suatu informasi dengan potensi tersembunyi yang tidak diketahui. Dengan *data mining*, analisis pola-pola yang dari suatu data dapat ditemukan untuk membantu pengambilan keputusan. *Data mining* sering disebut juga dengan *knowledge discovery* di dalam *database* (KDD) [5]. Berdasarkan penjelasan yang telah terpapar di atas dapat disimpulkan bahwa *data mining* adalah sebuah proses penggalian informasi tidak terlihat pada suatu data besar dengan bantuan pembelajaran komputer untuk diterapkan dalam pengambilan suatu keputusan.

2.3.2 Teknik *Data Mining*

Berikut adalah beberapa pengelompokan teknik *data mining* berdasarkan tugas yang dilakukan :

1. Klasterisasi

   Klasterisisasi adalah proses memisahkan kumpulan data menjadi beberapa bagian atau kelompok sehingga pada kumpulan data yang memiliki tingkat kemiripan tinggi akan berada pada sekumpulan atribut yang sama. Teknik ini sering juga disebut dengan *unsupervised learning*.
2. Regresi

   Regresi adalah memperkirakan nilai dari suatu variabel kontinu berdasarkan identifikasi relasi atau hubungan yang ada pada variabel lain yang ada dengan asumsi sebuah model ketergantungan linier atau nonlinier.

3. Klasifikasi

   Klasifikasi adalah proses penambahan sebuah *record* pada *field* data dengan beberapa kategori yang telah ditentukan sebelumnya atau sering disebut juga dengan *supervised learning*.

4. Asosiasi

   Asosiasi adalah proses menemukan sekumpulan atribut yang memiliki hubungan ketergantungan dan membentuk beberapa aturan dari kumpulan tersebut [17].

2.4 *Machine Learning*

Machine learning merupakan pemanfaatan bidang komputer dan algoritma matematika untuk mengambil hasil pembelajaran yang berasal dari data sehingga menghasilkan informasi yang dapat digunakan pada masa depan [18]. Untuk mendapatkan informasi dari data yang ada, terdapat dua proses pembelajaran yang perlu dilakukan pada komputer, yaitu proses latihan (*training*) dan proses pengujian (*testing*). Dengan *machine learning* komputer dapat meningkatkan kemampuannya dalam mencari pengetahuan dari data. Proses pembelajaran yang diterima akan membuat komputer dapat mengenali pola-pola tidak terlihat dari suatu kumpulan data dan membuat suatu keputusan cerdas berdasarkan hasil analisis pola yang ditemukan. *Machine learning* dapat dibagi menjadi tiga jenis, yaitu *supervised*, *unsupervised*, dan *reinforcement* [19].

2.4.1 *Supervised Machine Learning*

Supervised learning adalah sebuah metode pembelajaran mesin untuk data yang memiliki informasi label. Label tersebut adalah atribut yang menjadi pembeda tiap data dalam kumpulannya. Tujuan dari metode pembelajar ini adalah memetakan masukan x ke keluaran y, dengan diberi label pasangan masukan-keluaran, sebagai berikut :

$$D = \{(x_i, y_i)\}_{i=1}^{N} \quad (2.1)$$

Keterangan :

D : *Training set*

N : Jumlah *training*

x_i : Variabel *input*

y_i : Variabel *output*

2.4.2 *Unsupervised Machine Learning*

Unsupervised learning adalah proses pembelajaran mesin yang dilakukan tanpa pengawasan sebab kumpulan data yang diberikan tidak memiliki label. Tujuan dari metode pembelajaran ini adalah menemukan pola menarik yang ada pada data. Pada metode *unsupervised learning* hanya diberikan *input* sehingga output yang dihasilkan ini tidak dapat di tebak.

$$D = \{(x_i)\}_{i=1}^{N} \quad (2.2)$$

Keterangan :

D : *Training set*

N : Jumlah *training*

x_i : Variabel *input*

2.4.3 *Reinforcement*

Reinforcement learning adalah proses pembelajaran mesin yang berguna untuk mempelajari pengalaman baru bagaimana pengambilan keputusan dilakukan ketika mesin sesekali diberikan sinyal berbeda. Informasi yang diperoleh pada pembelajaran ini didapat melalui percobaan yang dilakukan terus menerus.

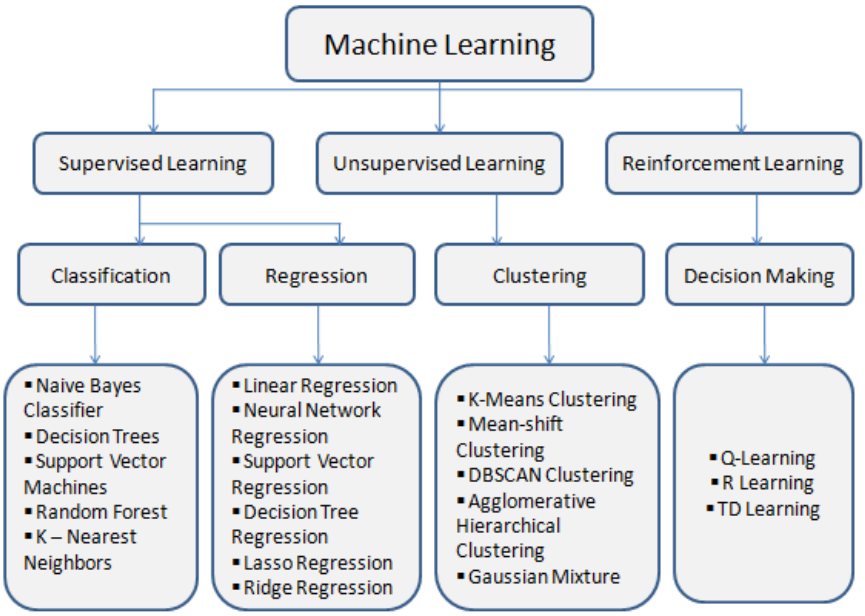
2.5 Klasifikasi

Klasifikasi merupakan sebuah proses untuk menemukan model yang dapat membedakan kelas atribut target berdasarkan dari atribut bebas yang ada pada data. Model dibuat berdasarkan analisis kumpulan data pelatihan yang label kelasnya diketahui. Kemudian model hasil *training set* digunakan untuk memprediksi label kelas suatu objek belum diketahui. Oleh karena itu, terdapat dua langkah yang perlu dilakukan pada klasifikasi. Langkah pertama adalah pembelajaran untuk membangun model klasifikasi. Kemudian, langkah kedua adalah klasifikasi

menggunakan model untuk memprediksi kelas yang labelnya belum diketahui dan memperkirakan akurasi dari prediksi yang dilakukan [3].

2.5.1 Algoritma *Supervised Machine Learning*

Klasifikasi merupakan salah satu teknik data mining yang menggunakan jenis *supervised learning*. Ada beragam algoritma *supervised learning* yang dapat digunakan untuk melakukan pengklasifikasian data, seperti yang ditunjukkan dalam gambar 1. Namun, pada penilaian kemampuan pembayaran kredit ini hanya akan membahas dua algoritma saja sesuai dengan latar belakang dan batasan masalah yang telah dituliskan pada bab sebelumnya. Kedua lgoritma tersebut adalah *logistic regression* dan *random forest*.



Gambar 1. Jenis algoritma *machine learning* [20].

2.5.1.1 *Logistic Regression*

Logistic regression merupakan suatu metode untuk menganalisis pola hubungan antara sekumpulan variabel bebas (independen) dengan variabel respon (dependen). Pada *logistic regression* variabel bebas (independen) dapat berupa data kategorik atau numerik sementara variabel respon (dependen) berupa data kategorik. Oleh karena itu, *logistic regression* sesuai untuk diterapkan pada variabel dependen yang bertipe kategorikal ataupun biner. Algoritma ini mampu untuk mengeksplorasi

hubungan variabel dependen dengan variabel independen yang dapat berupa atribut nominal, ordinal, atau rasio [8]. Namun, algoritma *logistic regression* memiliki keterbatasan untuk menangani klasifikasi data kompleks yang memiliki banyak variabel input. Keterbatasan lain dalam algoritma ini adalah tidak dapat menangani masalah multikolinieritas antara atribut-atribut yang terlibat. Selain itu, algoritma *logistic regression* juga rentan terhadap *overfitting* terutama pada dataset yang tidak seimbang [7].

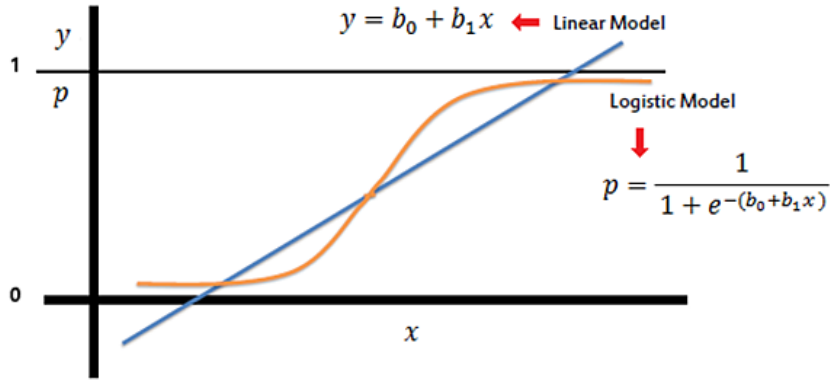
Berdasarkan variabel respon yang digunakan *logistic regression* dapat dibagi menjadi 3 macam, yaitu :

1. *Binary Logistic Regression*

   Binary logistic regression adalah metode analisis untuk mencari hubungan antara variabel respon (y) yang bersifat biner atau dikotom dengan variabel respon (x) yang bersifat kategorik atau numerik. Oleh karena itu, *output* yang dihasilkan pada *binary logistic regression* hanya terdiri dari dua nilai saja. Contoh : penentuan pasien yang positif atau negatif Covid, penentuan pelajar yang lulus dan tidak lulus, dan penilaian kemampuan kredit calon nasabah yang layak atau tidak layak untuk diberikan pinjaman.
2. *Multinomial logistic regression*

   Multinomial logistic regression adalah metode analisis untuk mencari hubungan antara variabel respon (y) yang memiliki dua atau lebih *output* dengan variabel respon (x) yang bersifat kategorik atau numerik. Contoh : penentuan dalam penggolongan kartu keanggotaan, seperti *silver, gold,* atau *platinum*.
3. *Ordinal Logistic Regression*

   *Ordinal logistic regression a*dalah *logistic* metode analisis untuk mencari hubungan antara variabel respon (y) yang memiliki dua atau lebih *output* yang memperhatikan urutan dengan variabel respon (x) yang bersifat kategorik atau numerik. Contoh : penentuan status akreditas sekolah.



Gambar 2. Model *linear* dan *logistic regression* [21].

Logistic regression adalah model linier yang digunakan dalam proses klasifikasi. *Logistic regression* berbeda dengan *linear regression* yang digunakan dalam proses regresi pada *machine learning*. Dalam *linear regression*, proses analisis dilakukan untuk mencari hubungan pada variabel respon (dependen) yang berupa informasi numerik berbeda dengan *logistic regression* yang variabel responnya berupa kategori. Pada *logistic regression* terdapat suatu fungsi yang digunakan untuk melakukan proses klasifikasi, yaitu *logistic function*. *Logistic function* merupakan suatu fungsi untuk mengatasi permasalahan klasifikasi yang tidak dapat diselesaikan menggunakan *linear function* pada regresi. Pembentukan *logistic function* dilakukan dengan mengganti nilai Y pada *linear function* dengan nilai Y pada *sigmoid function*. *Linear function* merupakan suatu fungsi yang digunakan dalam regresi untuk membentuk garis lurus pada grafik.

Sementara *Sigmoid function* merupakan fungsi yang merubah bentuk *odds* menjadi suatu logaritma. *Odds* merupakan bentuk lain dari peluang yang membandingkan proporsi antara peluang suatu kejadian terjadi/ kejadian tidak terjadi. Nilai pada peluang berkisar pada angka 0 sampai 1. Sementara dalam *odds* batasan nilai yang mungkin terjadi berkisar dari angka 0 sampai tak hingga (*infinite*).

$$Odds = \frac{p}{1-p} \tag{2.3}$$

Untuk mendapatkan *sigmoid function*, bentuk *odds* perlu diubah menjadi bentuk logaritma. Proses perubahan ini disebut dengan *log of odds*. Pada log of odds, nilai yang dihasilkan berada pada kisaran angka $-\infty$ sampai ∞.

$$logit\ (p) = \log\left(\frac{p}{1-p}\right) \tag{2.4}$$

Keterangan :

p : Peluang terjadinya suatu kejadian.

Odds dan *log of odds* merupakan penghubung yang dapat merubah hasil regresi linier ke bentuk peluang. *Logistic regression* menghasilkan *log of odds*. Namun, nilai yang ada pada *log of odss* tidak dapat diinterpretasikan sehingga nilai tersebut perlu ditransformasikan kembali dalam bentuk *odds* (untuk interpretasi variabel) atau peluang (untuk klasifikasi). Oleh karena itu, pada *logistic regression* nilai *log of odds* dikembalikan ke bentuk peluang dengan melakukan *inverse*. Hasil *inverse log of odds*/*logit* inilah yang disebut dengan *sigmoid function.*

$$P = \frac{1}{1+e^{-Y}} \tag{2.5}$$

Setelah *sigmoid function* didapatkan maka penyetaraan nilai dilakukan dengan menggunakan *sigmoid function* dan *linear function*. Proses penyetaraan *sigmoid function* dan *linear function* dilakukan dengan melakukan subtitusi nilai Y pada *linear function* ke dalam *sigmoid function* sehingga didapatkan hasil sebagai berikut:

Binary logistic :

$$P = \frac{1}{1+e^{-(b_0+b_1x)}} \tag{2.6}$$

Multinomial logistic :

$$P = \frac{1}{1+e^{-(b_0+b_1x+b_2x_2+\cdots+b_px_p)}} \tag{2.7}$$

Dari rumus 2.7 dan 2.8 dapat diketahui bahwa nilai *sigmoid function* dipengaruhi oleh persamaan $b_0 + b_1x + b_2x_2 + \cdots + b_px_p$. Hal tersebut akan mempengaruhi nilai *likelihood* pada *logistic function.* Peran *maximum likelihood* diperlukan pada *logistic function* untuk menentukan posisi sigmoid yang dapat menjadi model terbaik.

Penggunaan *R-Squared* dapat diimplementasikan untuk mengetahui model manakah dengan nilai *maximum likelihood* terbaik yang dapat digunakan dalam representasi data. Untuk mencari nilai *R-Squared* terdapat dua variabel yang perlu diketahui, yaitu *maximum likelihood* dan *Badfit likelihood*. Variabel *Badfit likelihood* didapatkan dari rumus dibawah ini:

$$Badfit\ likelihood = Log(Y) + Log(Y) + \cdots + Log(1-Y) + Log(1-Y) \quad (2.8)$$

Keterangan :

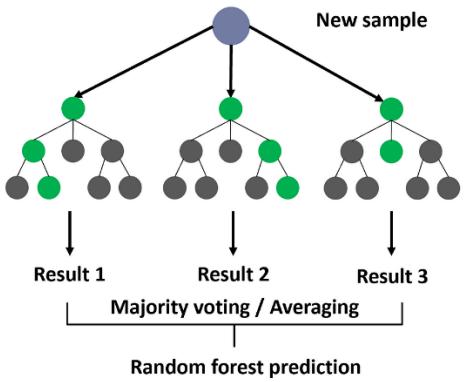
Y = $\frac{banyak\ data\ kelas\ 1}{banyak\ data\ keseluruhan}$

Berikut adalah rumus *R-Squared* dengan *maximum likelihood* dan *Badfit likelihood* :

$$R - Squared = \frac{\text{Badfit Likelihood} - \text{Maximum Likelihood}}{\text{Badfit Likelihood}} \quad (2.9)$$

2.5.1.2 *Random Forest*

Random Forest merupakan algoritma *supervised learning* yang termasuk kedalam model *ensemble*. Model *ensemble* merupakan model yang menggabungkan beberapa metode lain untuk meningkatkan nilai akurasi hasil proses klasifikasi. *Random Forest* terdiri dari beberapa kumpulan *decision tree* yang membentuk sebuah hutan klasifikasi. Setiap pohon keputusan bergantung pada nilai *random vector* yang diambil menjadi sampel secara bebas dan merata pada semua pohon yang terdapat dalam hutan. Saat klasifikasi, setiap pohon keputusan akan memberikan suara untuk kelas yang paling popular [3].



Gambar 3. Algoritma klasifikasi *random forest* [22].

Kelebihan utama dari algoritma *random forest* adalah dapat mengatasi *overfitting*. Selain itu, *random forest* mampu menangani ketidakseimbangan kelas pada suatu data yang dapat menyebabkan nilai akurasi menurun [6]. Namun, algoritma ini memiliki kekurangan, yaitu memakan waktu pemodelan lebih lama karena proses kompleksitas komputasi yang digunakan tinggi. Algoritma *random forest* menggunakan banyak jumlah pohon sehingga dapat menyebabkan kesulitan dalam menafsirkan model secara jelas [10].

Random Forest merupakan hasil pengembangan metode *Classification and Regression Tree* (*CART*) yang menerapkan metode *bagging* atau *bootstrap aggregating* dan *random feature selection*. *Bagging* adalah metode yang dapat memperbaiki hasil dari algoritma klasifikasi. Proses dalam melakukan klasifikasi dengan metode *random forest* terbagi menjadi dua tahapan. Tahap pertama adalah pembentukan ‘k’ pohon untuk membuat hutan yang acak. Tahap kedua adalah melakukan klasifikasi dengan hutan acak (*random forest*) yang telah terbentuk [3]. Berikut adalah penjelasan dari tahapan pertama metode *random forest*:

1. Pembuatan data sampel dengan cara pengambilan acak dataset dengan pengembalian.
2. Penggunaan sampel data untuk membangun pohon ke i dengan nilai i adalah 1, 2, 3, ..., sampai k.
3. Perulangan pada langkah 1 dan 2 sebanyak k.

Pada metode *random forest*, setelah melakukan pemilihan sampel dataset secara acak dilakukan proses pembangunan pohon keputusan menggunakan metode *CART*. Perhitungan dalam membangun pohon keputusan dengan metode *CART* dapat dilakukan dengan menggunakan *information gain*. *Information gain* menggambarkan ukuran dalam pemilihan atribut yang digunakan pada setiap *node* dari pohon-pohon klasifikasi. Namun, perhitungan *entropy* perlu dilakukan terlebih dahulu sebelum menghitung *information gain*. Rumus perhitungan *entropy* dapat dilihat pada persamaan berikut :

$$Entropy\ (S)\ =\ \sum_{i=1}^{n} p_i \log_2(p_i) \tag{2.10}$$

Keterangan :

S : Himpunan Kasus

n : Jumlah Partisi S

p_i : Proporsi Si terhadap S

Setelah menemukan nilai *entropy* maka subtitusi nilai *entropy* ke dalam perhitungan *information gain* pada persamaan berikut :

$$Gain(S, A) = Entropy(S) - \sum_{i=1}^{n} \frac{|Si|}{|S|} * Entropy(Si) \qquad (2.11)$$

Keterangan :

S : Himpunan Kasus

A : Fitur

n : Jumlah Partisi Atribut A

$|S_i|$: Proporsi Si terhadap S

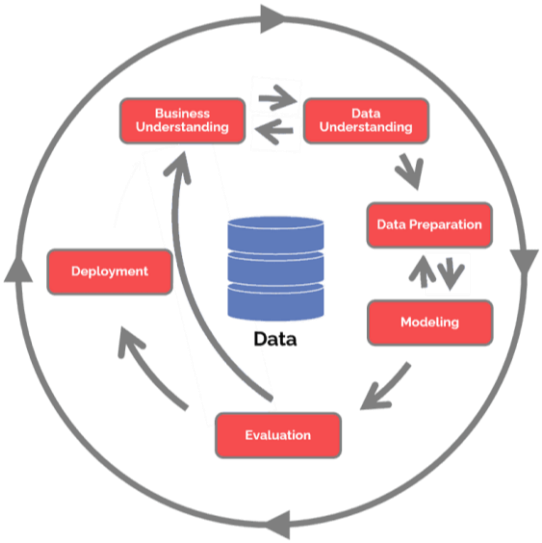
$|S|$: Jumlah kasus dalam S

Setelah pembuatan pohon keputusan telah dibentuk, langkah selanjutnya adalah melakukan klasifikasi dengan data test yang telah terbentuk. Berikut adalah tahapan kedua metode *Random Forest* :

1. Pengambilan data test dan penggunaan *rule* pada setiap pohon keputusan untuk melakukan klasfikasi dari data dan menyimpan hasil klasifikasi yang diperoleh.
2. Penghitungan suara untuk setiap hasil klasifikasi yang diperoleh dari setiap pohon keputusan.
3. Pemilihan hasil prediksi akhir dengan memilih kelas yang paling banyak diprediksi dari metode *random forest*.

2.6 *CRISP-DM*

CRISP-DM atau *Cross Industry Standard Process for Data-Mining* merupakan suatu standar baku untuk *data mining* dalam menganalisis strategi pemecahan masalah pada suatu industri. Penerapan *CRISP-DM* terdiri dari enam tahapan, mulai dari *business understanding*, *data understanding*, *data preparation, modeling*, *evaluation*, dan berakhir pada tahapan *deployment*.



Gambar 4. Proses metode *CRISP-DM* [23].

Tahapan yang terjadi dalam metode *CRISP-DM* tertera pada gambar 4. Pada gambar tersebut terlihat bahwa siklus *CRISP-DM* merupakan proses yang tidak kaku dimana dalam setiap tahapan memungkinkan terjadinya pergerakan maju-mundur sehingga data dapat terus berevolusi melalui iterasi yang dilakukan. Berikut adalah penjelasan terkait proses pada tiap tahapan di dalam *CRISP-DM* :

1. *Business Understanding*

   Pada tahap ini terjadi proses pemahaman dari tujuan kegiatan *data mining* yang akan dilaksanakan berdasarkan sudut pandang bisnis. Kegiatan menilai situasi bisnis, terkait batasan, asumsi, biaya, dan risiko juga dibahas pada *business understanding*.
2. *Data Understanding*

   Pada tahap ini terjadi proses pemahaman data yang diteliti, pengumpulan data, dan pengintegrasian data dengan tujuan untuk memperoleh informasi awal, mengevaluasi kualitas data, serta mengertahui pola permasalahan yang ada pada data.
3. *Data Preparation*

   Pada tahap ini dilakukan pemilihan data sesuai dengan analisis yang akan dilakukan melalui *data reduction*, membersihkan *raw data* melalui *data cleaning*, dan mentransformasikan data kedalam bentuk yang siap dimodelkan.

4. *Modeling*

   Pada tahap *modelling* data akan diolah menggunakan teknik pemodelan yang ada pada *data mining*. Pada tahapan ini dapat terjadi proses pengolahan data lagi untuk mengoptimalkan teknik pemodelan.

5. *Evaluation*

   Pada evaluasi dilakukan uji kelayakan atau uji validitas dari hasil pemodelan *data mining* yang mendekati kriteria kesuksesan bisnis. Setelah itu, pada tahap ini juga dilakukan pengambilan keputusan terkait penggunaan hasil *data mining*.

6. *Deployment*

   Pada tahap *deployment* terjadi proses pemaparan informasi dan pengetahuan yang didapatkan dari keseluruhan proses *data mining* yang telah dilakukan [23].

2.7 *Data Preparation*

2.7.1 *Data Cleaning*

Data Cleaning merupakan proses yang dilakukan dalam *data preparation* dengan tujuan untuk melakukan pembersihan pada data yang akan diolah. Proses pembersihan tersebut dilakukan dengan menemukan, memperbaiki, atau menghapus permasalahan-permasalahan yang muncul pada data, seperti nilai yang hilang, ketidaksesuaian data, kesalahan format, dan data berulang [24]. Permasalahan tersebut harus diperbaiki karena dapat menurunkan kualitas suatu data sehingga dapat menyebabkan informasi yang dihasilkan *data mining* menjadi kurang akurat.

Permasalahan nilai yang hilang atau *missing values* merupakan hal yang perlu dilakukan dalam pembersihan data karena model *data mining* tidak dapat memproses nilai yang hilang. Oleh karena itu, *missing values* perlu diidentifikasi dalam *data cleaning*. Identifikasi *missing values* dapat menggunakan perintah isna() fungsi dalam library *pandas* yang terdapat dalam *Python*. Permasalahan nilai yang hilang dapat diatasi dengan metode berikut:

1. Pengabaian tupel merupakan salah satu metode yang biasanya dilakukan pada data yang tidak mempunyai label kelas. Metode ini efekif pada data dengan tuple yang memiliki banyak nilai yang hilang.
2. Pengisian nilai yang hilang secara manual digunakan untuk mengatasi kelemahan metode pertama, tetapi memakan banyak waktu sehingga tidak sesuai pada data besar yang mengandung banyak atribut kosong.
3. Penggunaan konstanta global untuk mengisi nilai yang hilang dengan mengganti semua nilai yang hilang dengan konstanta yang sama seperti label seperti 'tidak diketahui' atau $-\infty$.
4. Pengisian dengan ukuran tendensi sentral untuk mengisi nilai yang hilang. Tendensi sentral menunjukkan nilai tengah dari suatu persebaran data.
5. Penggunaan nilai mean atau median untuk semua sampel yang termasuk dalam kelas yang sama dengan tupel yang diberikan, seperti mengisi nilai pendapatan dengan nilai rata-rata atau median saat distribusi kelas tersebut tidak seimbang.
6. Pengisian dengan nilai yang paling mungkin dengan regresi, alat berbasis inferensi menggunakan *Bayesian Formalism* atau *Decision Tree Induction* [3].

2.7.2 *Data Reduction*

Data reduction merupakan proses terjadinya pengurangan dimensi pada data, tetapi tetap mempertahankan hasil analisis yang optimal. Dalam *data reduction*, terjadi proses penghilangan fitur atau atribut yang memiliki hubungan yang rendah pada akurasi model [24]. Pengurangan dimensi tersebut dapat dilakukan dengan menggunakan *feature selection*.

Feature selection adalah suatu proses penghapusan atribut yang tidak memiliki relevansi dengan dataset sehingga mengurangi waktu yang digunakan dalam eksekusi data dan meningkatkan akurasi. *Feature selection* mengurangi kompleksitas untuk mengetahui atribut yang paling signifikan. Metode *feature selection* dibagi menjadi tiga kelompok, sebagai berikut :

1. *Filters Approach*

   Metode *filters approach* menggunakan teknik pemeringkatan. Setiap atribut akan diberi skor dan peringkat yang sesuai. Atribut dengan skor di bawah nilai

ambang batas yang telah ditetapkan akan dihapus dari dataset. Kelebihan metode ini adalah tidak bergantung pada jenis pengklasifikasi yang digunakan. Sementara kelemahannya adalah mengabaikan ketergantungan atribut karena setiap atribut dianggap terpisah. Contoh *feature selection* yang masuk kedalam kelompok *filters approach* adalah *Correlation*, *Chi-square test*, *Euclidean Distance*, dan *Information Gain*.

2. *Wrappers Approach*

   Dalam metode ini fitur bergantung pada pengklasifikasi yang digunakan, yaitu menggunakan hasil pengklasifikasi untuk menentukan kelebihan fitur atau atribut yang diberikan. Kelebihan dari metode ini adalah menghilangkan kekurangan dari metode filter, yaitu dapat mempertahankan ketergantungan antara fitur. Kekurangan dari metode ini adalah memakan waktu yang lebih lama dari metode *filters approach*. Metode *wrapper* secara garis besar diklasifikasikan kedalam *Sequential Selection Algorithms* dan *Heuristic Search Algorithms.*

3. *Embedded Approach*

   Metode embedded sering disebut juga dengan metode *hybrid* yang merupakan kombinasi metode *filters* dan *wrapper*. Metode tertanam ini mengurangi waktu komputasi yang digunakan untuk mengklasifikasi ulang *subset* berbeda yang dilakukan dalam metode wrapper. *KP-SVM* adalah contoh metode tertanam [25].

2.7.2.1 *Correlation*

Correlation adalah metode statistika yang digunakan untuk menentukan nilai yang menunjukkan bagaimana hubungan antara suatu variabel dengan variabel lain dengan mengabaikan apakah suatu variabel bergantung pada variabel lainnya [26]. *Correlation* dapat digunakan untuk perfoma suatu model. Metode ini dapat digunakan untuk membandingkan variabel mana yang paling relevan dalam menentukan perolehan hasil dari variabel target. *Correlation* dapat difungsikan sebagai bahan pertimbangan dalam melakukan proses pemilihan dengan variabel-variabel yang paling efektif. Metode ini juga dapat mengurangi ukuran data

sehingga proses pengolahan data akan lebih mudah dan peluang terjadinya error menjadi lebih kecil.

Hasil pemilihan *feature* dengan metode *correlation* akan berada pada rentang nilai -1 hingga 1. Nilai positif menunjukkan adanya korelasi positif antara kedua variabel artinya kedua variabel tersebut bergerak searah. Kemudian, nilai negatif menunjukkan korelasi negatif yang artinya kedua variabel tersebut berpengaruh, tetapi bergerak secara berlawanan arah. Sementara nilai 0 pada *correlation* menunjukkan kedua variabel tersebut tidak berpengaruh satu sama lain. Berikut adalah proses perhitungan untuk mendapatkan nilai *correlation* antara dua variabel, X dan Y :

$$r = \frac{\sum_{i=1}^{n}(X_i - \hat{X})(Y_i - \widehat{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \hat{X})^2 (Y_i - \widehat{Y})^2}} \tag{2.12}$$

Keterangan :

r : Koefisien korelasi
X_i : Nilai sampel ke-i dari variabel X
Y_i : Nilai sampel ke-i dari variabel Y
$\hat{X}$: Rata-rata dari variabel X
$\hat{Y}$: Rata-rata dari variabel Y
n : Jumlah sampel

2.7.3 *Data Transformation*

Data transformation merupakan proses melakukan perubahan pada data menjadi bentuk yang dapat diterima oleh algoritma *data mining* sehingga pola yang ditemukan dapat lebih mudah dipahami. Dengan data transformation perubahan pada jenis atau distribusi atribut data dapat dilakukan. Atribut yang terdapat dalam data memiliki beberapa jenis, yaitu numerik atau kategori. Oleh karena itu, teknik yang dipilih untuk melakukan transformasi juga memiliki perbedaan antara jenis yang satu dengan yang lainnya. Dalam *data transformation*, terdapat beberapa teknik yang dapat digunakan pada data dengan jenis numerik, yaitu :

1. *Smoothing*

   Teknik *smoothing* digunakan untuk mengatasi data yang tidak valid untuk proses *mining*. Salah satu metode yang masuk ke dalam teknik *smoothing* adalah *clustering*. *Clustering* merupakan proses mengelompokkan serangkaian pola yang diberikan ke dalam cluster terpisah dan menghilangkan *outliers*.

2. *Generalization*

   Generalization merupakan proses yang menggunakan teknik diskretisasi untuk mereduksi sekumpulan nilai yang terdapat pada atribut *continuous* ke dalam interval. Salah satu metode yang masuk ke dalam teknik *generalization* adalah *histogram analysis*. *Histogram analysis* merupakan proses pengurutan dan pembagian data ke dalam suatu rentang dengan menyimpan nilai rata-rata (total) tiap keranjang.

3. *Normalization*

   Normalization adalah proses mengubah data numerik ke dalam suatu skala yang memiliki rentang dari -1 sampai 1 atau 0 sampai 1. Salah satu metode yang diterapkan dalam normalisasi data adalah *min-max normalization* atau *min-max scaler*. *Min-max normalization* merubah suatu nilai v dari atribut A menjadi v' ke dalam skala $[new\_max_A, new\_min_A]$ dengan perhitungan sebagai berikut :

   $$v_i' = \frac{v_i - min_A}{max_A - min_A} (new\_max_A - new\_min_A) + new\_min_A \quad (2.13)$$

4. *Aggregation*

   Aggregation merupakan proses yang diterapkan pada data numerik dengan pemanfaatan operator *data cube* (operasi *roll up*/meringkas).

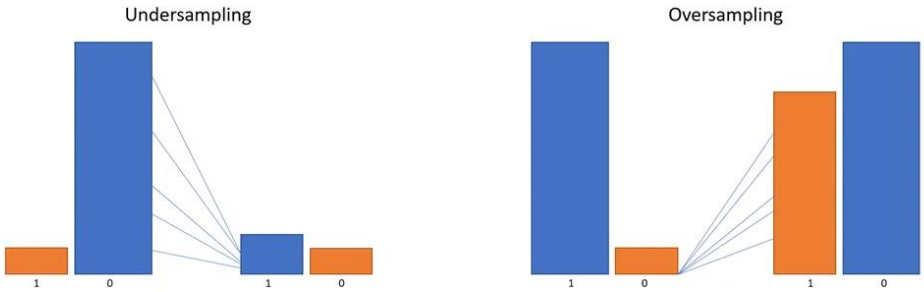
5. *Attribute Construction*

   Attribute construction merupakan proses pembentukan atribut baru dengan memanfaatkan atribut yang telah ada. Atribut yang telah dibuat kemudian ditambah bersama atribut lainnya untuk meningkatkan ketepatan dalam *high-dimensional data* [27].

2.8 Ketidakseimbangan Kelas

Ketidakseimbangan kelas atau *imbalance class* merupakan suatu kondisi yang terjadi ketika penyebaran dataset antarkelas tidak seimbang. *Imbalance class* akan membuat salah satu kelas memiliki jumlah data yang sangat besar (kelas mayoritas)

dibanding kelas lainnya (kelas minoritas). Kondisi ketidakseimbangan ini dapat menjadi masalah dalam klasifikasi karena proses pembelajarannya akan lebih banyak dalam memprediksi ke kelas data mayoritas dibanding dengan kelas minotitas [28]. Hal tersebut dapat mengakibatkan akurasi hasil prediksi yang dilakukan selama klasifikasi terhadap *data training* kelas minoritas akan kurang baik.



Gambar 5. Jenis metode *sampling* [32].

Permasalahan ketidakseimbangan kelas dapat diatasi dengan metode *sampling*. Metode sampling dapat dibagi menjadi dua jenis seperti yang ditampilkan pada gambar 5. Berikut adalah penjelasan dari masing-masing jenis yang ada pada metode *sampling* :

1. *Undersampling*

   Undersampling merupakan teknik penyeimbangan dataset dengan mengurangi ukuran data pada kelas mayor. Metode ini akan tetap menyimpan semua sampel yang ada di kelas minor dan mengurangi jumlah sampel pada kelas mayor secara acak.
2. *Oversampling*

   Oversampling merupakan tenik penyeimbangan dataset untuk menyeimbangkan dataset dengan meningkatkan ukuran data pada kelas minor. Contoh teknik dari *oversampling* adalah *SMOTE* (*Syntetic Minority Oversampling Technique*).

2.8.1 *SMOTE*

SMOTE merupakan salah satu metode jenis *oversampling* yang mengubah ketidakseimbangan data dengan membuat data sintetis baru dari kelas minoritas.

Dengan *SMOTE*, vektor dari fitur yang ada di kelas minoritas akan dikurangi dengan nilai *nearest neighbor* yang ada di kelas minoritas. Kemudian hasil selisih yang telah didapatkan dikali dengan angka 0 sampai 1 secara random. Hasil perkalian tersebut ditambah dengan vektor fitur dan didapatlah hasil nilai vektor yang baru. Berikut adalah proses penulisan perhitungan *SMOTE* [28]:

$$X_{new} = X_i + \left(\hat{X}_l - X_i\right) \delta \tag{2.14}$$

Keterangan :

X_i : vektor dari fitur pada kelas minoritas

$\hat{X}_l$: k-*nearest* neighbors untuk Xi

δ : angka acak antara 0 sampai 1

2.9 Evaluasi Perfoma

Model klasifikasi yang dibangun menggunakan *machine learning* diharapkan dapat melakukan klasifikasi semua data dengan benar. Namun, hasil kinerja model klasifikasi tidak dapat dipastikan seluruhnya benar. Oleh karena itu, evaluasi perfoma diperlukan untuk mengukur hasil kinerja metode klasifikasi yang dibangun dengan hasil sebenarnya. Salah satu evaluasi perfoma yang dapat dilakukan untuk mengukur kinerja klasifikasi, yaitu *confusion matrix*.

2.9.1 *Confusion Matrix*

Confusion matrix merupakan matriks digunakan untuk menggambarkan hubungan antara variabel sebenarnya dengan variabel prediksinya. *Confusion matrix* akan menampilkan matriks berukuran $n \; x \; n$ yang berkaitan dengan *classifier* dimana variabel n adalah jumlah kelas yang ada pada data. Ada beberapa istilah yang digunakan untuk merepresentasikan hubungan antara variabel sebenarnya dengan variabel prediksi dalam matriks evaluasi, yaitu [3]:

True Positive (*TP*) : data positif yang diberi label dengan tepat pada klasifikasi.

True Negative (*TN*) : data negatif yang diberi label dengan tepat pada klasifikasi.

False Positive (*FP*) : data positif yang diberi label dengan tidak tepat pada klasifikasi.

False Negative (*FN*) : data negatif yang diberi label dengan tidak tepat pada klasifikasi.

Dari representasi nilai yang ditampilkan pada *confusion matrix* ini, perhitungan nilai lain, seperti *accuracy, precision, recall*, dan *F-Measure* juga dapat dilakukan sebagai parameter penilaian kinerja motede evaluasi. Penjelasan dari tiap-tiap parameter evaluasi tersebut dapat dilihat dibawah ini :

1. *Accuracy* merupakan parameter yang didapatkan dengan membandingkan banyaknya data yang terklasifikasi tepat dengan jumlah total data. Semakin banyak data yang terklasifikasikan dengan benar membuktikan semakin baik akurasi metode klasifikasi yang telah dibuat. Rumus perhitungan *accuracy* dapat dilihat pada persamaan :

$$Accuracy = \frac{TP+TN}{P+N} \tag{2.15}$$

2. *Precision* merupakan parameter yang membandingkan banyak data positif yang terklasifikasi benar dengan jumlah total data positif yang ada. Semakin banyak jumlah data positif yang terklarifikasi benar akan meningkatkan nilai *precision*. Rumus perhitungan *precision* dapat dilihat pada persamaan :

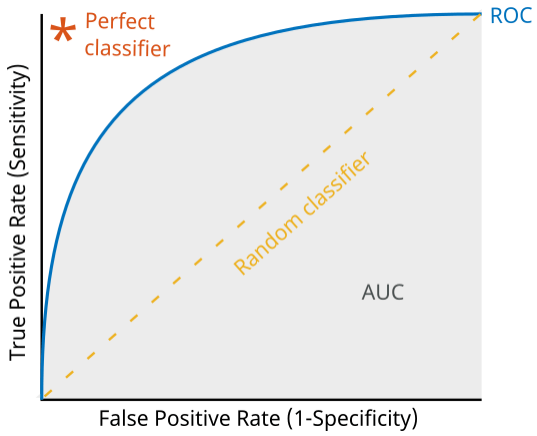
$$Precision = \frac{TP}{TP+FP} \tag{2.16}$$

3. *Recall* atau *sensitivity* merupakan parameter yang membandingkan jumlah data positif yang terklasifikasi benar dengan jumlah total data benar. Rumus perhitungan *recall* dapat dilihat pada persamaan :

$$Recall = \frac{TP}{TP+FN} = \frac{TP}{P} \tag{2.17}$$

4. *F-Measure* atau *F1-Score* merupakan parameter yang dapat digunakan untuk menangani masalah *imbalance class*. *F1-Score* mengkombinasi *recall* dan untuk mencari informasi dalam himpunan yang mengandung ketidakseimbangan kelas. Rumus perhitungan *F1-Score* dapat dilihat pada persamaan :

$$F1 - Score = \frac{2\ x\ recall\ x\ precision}{recall+precision} \tag{2.18}$$



Gambar 6. Contoh *Area Under the ROC Curve* [29].

Selain dari perhitungan parameter diatas, evaluasi perfoma klasifikasi juga dapat dilakukan melalui visualisasi. *Area Under the ROC* (*Receiver Operating Characteristic*) *Curve* atau *AUC* merupakan representasi hasil kinerja model klasifikasi yang menunjukkan tingkat keberhasilan dengan memisahkan data positif. Nilai *AUC* berada diantara angka 0 sampai 1. Apabila nilai perfoma *AUC* yang dihasilkan sebesar 0.5 maka model klasifikasi termasuk kedalam model buruk karena klasifikasi data dilakukan secara acak. Rumus perhitungan *AUC* dapat dilihat pada persamaan :

$$AUC = \frac{1+TP_{Rate}-FP_{Rate}}{2} \tag{2.19}$$

Berikut adalah pembagian kategori nilai AUC [30]:

0.90 – 1.00 : *Excellent Classification*
0.80 – 0.90 : *Good Classification*
0.70 – 0.80 : *Fair Classification*
0.60 – 0.70 : *Poor Classification*
0.50 – 0.60 : *Failure*

2.10 Python

Python merupakan penulisan kode bahasa pemrograman yang diciptakan oleh Guido van Rossumpada tahun 1990 di Belanda. Python memiliki fitur *library* yang luas dan *syntax* yang mudah dan ringkas. Python adalah salah satu bahasa

pemrograman yang multiguna dan interpretatif karena *syntax* pada Python dirancang memiliki tingkat keterbacaan yang tinggi sehingga mudah dipelajari. Beberapa library yang terdapat pada python adalah sebagai berikut [31]:

1. Pandas

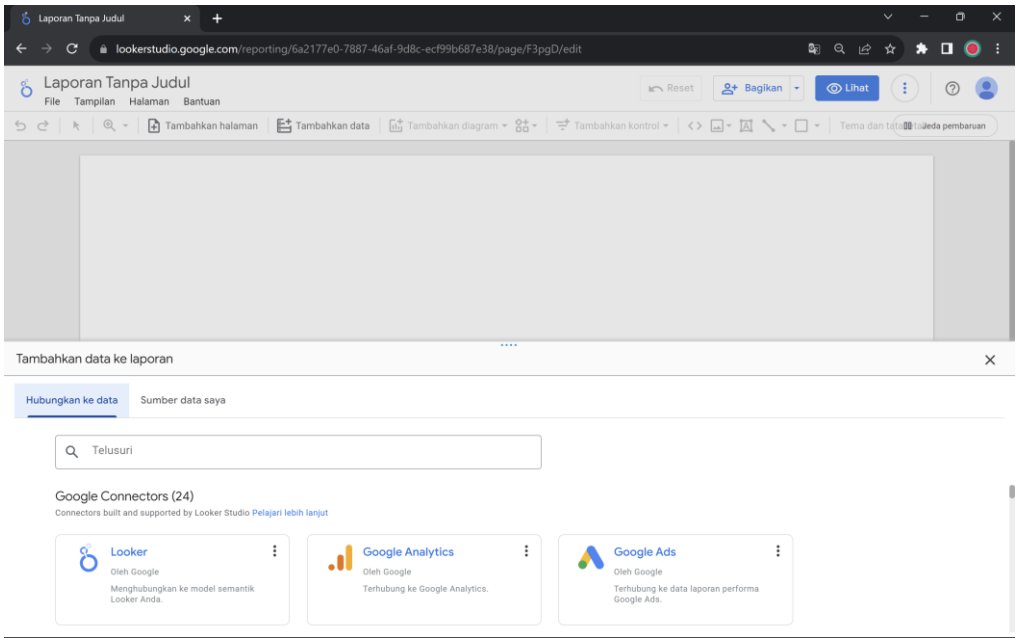
   Pandas merupakan *library open source* yang menyediakan analisis data tingkat tinggi secara terstruktur. Pandas dapat memudahkan proses pengolahan data, seperti analisis data, manipulasi data, dan pembersihan data. Beberapa fungsi yang disediakan oleh *library* ini adalah penyortiran, pengindeksan ulang, perulangan, penggabungan, penggambaran, pengelompokan data, dan lain lain.
2. NumPy

   NumPy atau *Numerical* Python merupakan *library open source* yang digunakan dalam mengolah *array*, aljabar linier, dan matriks. Dengan NumPy proses pengolahan objek array akan lebih cepat dibandingkan dengan penggunaan daftar secara manual.
3. Matplotlib

   Matplotlib merupakan *library open source* yang digunakan untuk merepresentasikan data numerik. Matplotlib digunakan dalam analisis untuk melakukan visualisasi data, seperti melakukan *plotting* angka berdefinisi tinggi dalam bentuk diagram lingkaran, histogram, scatterplot, grafik, dan lain-lain.
4. Sckit-learn

   Scikit-learn merupakan *library open source* yang digunakan pada data kompleks. *Library* ini sering digunakan dalam *machine learning* karena mendukung berbagai algoritma baik *supervised* maupun *unsupervised*, seperti regresi linier, klasifikasi, pengelompokan, dan lain.

2.11 Looker Studio



Gambar 7. Tampilan Google Looker Studio.

Looker Studio sebelumnya dikenal sebagai Google Data Studio merupakan sebuah *tools online* yang disediakan Google secara gratis untuk membuat laporan yang infromatif dan interaktif dengan *dashboard*. Looker Studio dapat menampilkan visualisasi dari data secara menarik sehingga akan membatu dalam pembuatan keputusan bisnis yang lebih cerdas. Looker Studio menyediakan beragam jenis grafik yang dapat digunakan untuk melakukan visualisasi, seperti tabel, diagram garis, diagram batang, diagram lingkaran, diagram sebar, peta area, peta hierarki, dan lain-lain. Kemudian, pada *tools* ini juga terdapat fitur *drag and drop*, *searching*, *checklist box* dan fitur lain yang dapat mebuat dashboard semakin interaktif. Selain itu, Looker Studio juga menyediakan fitur pembuatan *dashboard* kolaboratif dan pembagian *dashboard* yang dapat digunakan secara gratis berbeda dengan perusahaan lain yang menyediakan fitur tersebut secara berbayar.

2.12 Penelitian Terkait

Dalam penelitian ini, penulis mengacu kepada enam penelitian lain sebagai referensi. Salah satu penelitian sejenis dilakukan oleh Amrin dan Pahlevi [32] dengan tujuan untuk melakukan klasifikasi diagnosis penyakit hepatitis. Penelitian

ini memprediksi kondisi pasien positif memiliki penyakit hepatitis atau tidak berdasarkan sembilan belas variabel *predictor* dengan jumlah keseluruhan dataset yang digunakan sebanyak 155. Penelitian dibangun dengan menggunakan model klasifikasi *machine learning* *Logistic Regression* dan *Naive Bayes.* Hasil dari penelitian ini menunjukkan bahwa model yang dibangun dengan algoritma *Logistic Regression* memperoleh nilai akurasi dan AUC lebih tinggi dalam melakukan diagnosis penyakit hepatitis.

Penelitian yang dilakukan oleh Simanjuntak [33] adalah mengelompokkan *Tweet* ke dalam kelas emosi yang telah ditentukan sebelumnya. Data yang digunakan diperoleh menggunakan *TwitterStreaming API* mulai dari tanggal 1 Juni hingga 14 Juni 2018. Pada penelitian ini membandingkan bebarapa algoritma *machine learning*, yaitu *Logistic Regression* dan *Random Forest* dengan *SMOTE*. Tujuannya adalah untuk mengatasi permasalahan ketika dataset memiliki dimensi yang tinggi dan ketidakseimbangan kelas. Hasil penelitian ini menyimpulkan model yang dibangun menggunakan *Logistic Regression* dengan penerapan metode *SMOTE* dianggap lebih efektif dalam melakukan klasifikasi emosi dari data *Tweet* karena menghasilkan nilai akurasi, *Precision, Recall,* dan *F1-measure* lebih besar dari *Random Forest*.

Sementara itu, penelitian yang dilakukan oleh Mujaddid dkk. [34] bertujuan untuk melakukan klasifikasi kemungkinan pelanggan *churn* dan pelanggan bertahan. Algoritma *machine learning* yang digunakan dalam penelitian ini adalah *Logistic Regression* dengan menerapkan teknik *SMOTE* untuk mengatasi masalah *imbalance data*. Hasil penelitian menunjukkan bahwa nilai akurasi model yang dibangun menggunakan *Logistic Regression* dengan teknik *SMOTE* lebih baik dibandingkan model yang dibangun tanpa menggunakan teknik *SMOTE*.

Penelitian lain oleh Zhu dkk. [35] juga dilakukan untuk melakukan prediksi kegagalan pembayaran dalam pengembalian pinjaman. Pada penelitian ini membandingkan bebarapa algoritma *machine learning*, yaitu *Logistic Regression, SVM, Decision Tree* dan *Random Forest*. Teknik *SMOTE* diterapkan pada penelitian untuk mengatasi masalah ketidakseimbangan kelas dalam kumpulan data. Hasil penelitian ini menyimpulkan bahwa masalah ketidakseimbangan data

dapat diatasi dengan *SMOTE* dan model dengan algoritma *Random Forest* memiliki nilai akurasi terbesar dari algoritma lainnya.

Selain itu, penelitian untuk klasifikasi juga dilakukan oleh Khadija dan Setiawan [36] dengan tujuan mengidentifikasi diagnosis pasien penyakit hati. Penelitian ini menggunakan sepuluh variabel *predictor* dengan jumlah keseluruhan dataset yang digunakan sebanyak 583. Teknik *SMOTE* digunakan pada penelitian ini karena dataset yang digunakan memiliki ketidakseimbangan kelas dengan proporsi 70,36% kelas mayor dan 28,64% kelas minor. Penelitian dibangun dengan menggunakan model klasifikasi *Naïve Bayes, KNN, Random Forest,* dan *SVM.* Hasil dari penelitian ini menunjukkan bahwa model yang dibangun dengan algoritma *Random Forest* memperoleh nilai akurasi dan *AUC* lebih tinggi dalam melakukan diagnosis penyakit hati.

Penelitian lain dilakukan Rachmatullah [37] dengan tujuan untuk membuat model terbaik yang dapat digunakan dalam menilai pengajuan kredit dan mengklasifikasikan calon nasabah sebagai pelamar bagus atau buruk. Penentuan model terbaik dilakukan dengan membandingkan algoritma klasifikasi *KNN, random forest, SVM,* dan *multilayer perceptron* (*MLP).* Teknik *SMOTE* digunakan pada penelitian ini karena dataset yang digunakan memiliki ketidakseimbangan kelas dengan proporsi 70% kelas mayor dan 30% kelas minor. Hasil dari penelitian ini menunjukkan bahwa model yang dibangun dengan algoritma *Random Forest* lebih baik dibandingkan algoritma lainnya.

Sementara itu, penelitian untuk klasifikasi juga dilakukan oleh Septama dkk. [38] dengan tujuan menilai pengajuan kredit dan mengklasifikasikan calon nasabah Home Credit. Penelitian ini menggunakan 22 variabel *predictor* yang didapatkan dengan melakukan pemilihan *feature* menggunakan metode *Chi-Square Test* dari dataset *application train/test, bureau,* dan *previous application.* Penelitian dibangun dengan menggunakan model *Logistic Regression, Decision Tree, Random Forest,* dan *Naive Bayes.* Hasil dari penelitian ini menunjukkan bahwa model yang dibangun dengan algoritma *Random Forest* memperoleh nilai akurasi yang lebih tinggi dalam menilai pengajuan kredit nasabah.

Tabel 1. Penelitian terdahulu

No	Peneliti	Data	Algoritma	Hasil
1.	Amrin dan Pahlevi (2019) [32]	Data pasien hepatitis *UCI Machine Learning Repository*	*Logistic Regression* dan *Naive Bayes*	Model terbaik : *Logistic Regression* dengan akurasi sebesar 84,62% dan nilai AUC sebesar 0,84.
2.	Simanjuntak, dkk. (2023) [33]	Data tweet yang diperoleh menggunakan TwitterStreaming API mulai dari 1 Juni hingga 14 Juni 2018	*Logistic Regression* dan *Random Forest* dengan *SMOTE*	Model terbaik : *Logistic Regression* dengan akurasi sebesar 78.22%, *Precision* sebesar 79.06%, *Recall* sebesar 78.21%, dan *F1-score* sebesar 78.48%
3.	Mujaddid, dkk. (2017) [34]	Data WITEL PT. Telekomunikasi	*Logistic Regression* dan *Logistic Regression* dengan *SMOTE*	Model terbaik : *Logistic Regression* dengan *SMOTE* yang memiliki nilai akurasi sebesar 92,4% dan *F1-score* sebesar 31,27%
4.	Zhu, dkk. (2019) [35]	Data Klub Peminjaman kuartal pertama tahun 2019	*Logistic Regression, SVM, Decision Tree* dan *Random Forest*.	Model terbaik : *Random Forest* dengan nilai akurasi sebesar 98%, *AUC* sebesar 0.98, *F1-score* sebesar 0.98, dan *Recall* sebesar 0.98

No	Peneliti	Data	Algoritma	Hasil
5.	Khadija dan Setiawan (2020) [36]	Data *Indian Liver Patient UCI Machine Learning Repository*	*Naïve Bayes, KNN, Random Forest,* dan *SVM*	Model terbaik : *Random Forest* yang menggunakan *SMOTE* dan *Feature Selection* memperoleh nilai akurasi sebesar 77% serta *F1-score, Recall,* dan *Precision* sebesar 77%
6.	Rachmatullah (2023) [37]	Dataset Statlog (German Credit Data) *UCI Machine Learning Repository*	*KNN, Random Forest, SVM,* dan *MLP*	Model terbaik : *Random Forest* dengan SMOTE memperoleh AUC sebesar 0.83, *accuracy* sebesar 72%, *Recall* sebesar 82%, *F1-score* sebesar 75.1% dan *Precision* sebesar 69.3%
7.	Septama dkk. (2023) [38]	Data pengajuan kredit Home Credit	*Feature selection Chi-Square* pada *Logistic Regression, Decision Tree, Random Forest,* dan *Naive Bayes.*	Model terbaik : *Random Forest* dengan akurasi sebesar 89%, *Precision* sebesar 93%, *Recall* sebesar 96%, *F1-score* sebesar 94%, dan AUC 0.68.

Berdasarkan penelitian terdahulu yang telah dijelaskan pada tabel 1 didapatkan kesimpulan bahwa pembangunan model klasifikasi untuk memprediksi data dapat dilakukan menggunakan algoritma *logistic regression* dan *random forest*. Penelitian yang akan dilakukan adalah membangun model klasifikasi untuk

menentukan kemampuan pembayaran kredit dengan menggunakan *logistic regression* dan *random forest*. Hasil klasifikasi yang telah diprediksi dengan kedua model tersebut akan dibandingkan berdasarkan nilai evaluasi perfoma yang dihasilkan. Sumber data yang digunakan pada penelitian ini adalah data pengajuan kredit nasabah disertai data historis kredit nasabah pada pengajuan sebelumnya yang ada pada Home Credit. Penelitian ini menggunakan seluruh dataset disediakan Home Credit, yaitu *application train/test, bureau, bureau balance, previous application, POS cash balance, credit card balance*, dan *instalment payments*. Kemudian, penelitian ini menerapkan pemilihan *feature* dengan menggunakan metode *correlation* untuk memilih atribut yang akan digunakan pada proses pemodelan. Penelitian ini juga menerapkan teknik *SMOTE* untuk mengatasi masalah ketidakseimbangan data. Teknik tersebut digunakan untuk mengurangi pengaruh ketidakseimbangan data dalam nilai evaluasi pada model yang dihasilkan.

BAB III
METODE PENELITIAN

3.1 Waktu dan Tempat Penelitian

Penelitian ini dilaksanakan dari awal bulan September 2023 sampai dengan Januari 2024 bertempatan di Laboratorium Teknik Digital Jurusan Teknik Elektro Fakultas Teknik Universitas Lampung.

Tabel 2. Jadwal penelitian

No	Kegiatan	2023				2024
		September	Oktober	November	Desember	Januari
1.	Studi Pendahuluan					
2.	Persiapan Alat dan Bahan					
3.	Penerapan Metode *CRISP-DM*					
4.	Penulisan Laporan					

3.2 Alat dan Bahan Penelitian

3.2.1 Alat

Adapun penelitian ini menggunakan perangkat keras (*hardware*) dan perangkat lunak (*software*) dengan spesifikasi berikut.

Tabel 3. Alat penelitian

No	Perangkat	Spesifikasi	Deskripsi
1.	Laptop	Asus Vivobook Pro 14 OLED K3400PH, Intel Core i5-11300H, RAM 8 GB, SSD 512 GB	Perangkat keras yang digunakan untuk membangun dan menguji model klasifikasi *machine learning*.
2.	Python	Python 3.10.5	Python yang digukanan sebagai bahasa pemrograman dalam membangun model *machine learning.*
3.	Jupyter Notebook	Jupyter Notebook 6.4.8	Perangkat lunak yang digunakan untuk pengolahan data sebagai tempat pengintegrasian kode pemrograman menjadi sebuah *output*.
4.	Google Looker Studio		Perangkat lunak yang digunakan untuk pengolahan data menjadi *dashboard* interaktif dan laporan yang informatif melalui beragam visualisasi.

3.2.2 Bahan

Bahan yang digunakan dalam penelitian adalah data telekomunikasi dan transaksional Home Credit yang didapatkan saat mengikuti Studi Independen *Data Analytics* Batch 3 di Zenius. Data tersebut dapat diakses melalui website Kaggle dengan kata kunci Home Credit Default Risk [11]. Terdapat tujuh dataset yang diberikan, yaitu (1) application_{train|test}; (2) bureau; (3) bureau_balance; (4) POS_CASH_balance; (5) credit_card_balance; (6) previous_application; (7) installments_payments. Tabel 4 menampilkan deskripsi atribut-atribut yang ada pada setiap dataset.

Tabel 4. Deskripsi data

No.	Dataset	Nama Atribut	Deskripsi
1.	Application Train/Test	SK_ID_CURR	ID pinjaman di Home Credit saat ini
		TARGET	Variabel target yang diprediksi
		NAME_CONTRACT_TYPE	Jenis kontrak pinjaman saat ini
		CODE_GENDER	Jenis kelamin nasabah
		FLAG_OWN_CAR	Kepemilikkan mobil
		FLAG_OWN_REALTY	Kepemilikkan properti
		CNT_CHILDREN	Jumlah anak
		AMT_INCOME_TOTAL	Jumlah pendapatan
		AMT_CREDIT	Jumlah kredit dari pinjaman
		AMT_ANNUITY	Anuitan pinjaman
		AMT_GOODS_PRICE	Harga barang yang diberikan pinjaman
		NAME_TYPE_SUITE	Pendamping saat mengajukan pinjaman
		NAME_INCOME_TYPE	Jenis pendapatan
		NAME_EDUCATION_TYPE	Tingkat pendidikan tertinggi yang dicapai
		NAME_FAMILY_STATUS	Status keluarga
		NAME_HOUSING_TYPE	Status tempat tinggal
		REGION_POPULATION _RELATIVE	Jumlah populasi wilayah tempat tinggal yang dinormalisasi
		DAYS_BIRTH	Usia
		DAYS_EMPLOYED	Lama waktu bekerja
		DAYS_REGISTRATION	Rentang waktu mengubah pendaftarannya
		DAYS_ID_PUBLISH	Rentang waktu mengubah dokumen identitas yang digunakannya

No.	Dataset	Nama Atribut	Deskripsi
		OWN_CAR_AGE	Usia mobil nasabah
		FLAG_MOBIL	Apakah nasabah memiliki telepon seluler
		FLAG_EMP_PHONE	Apakah nasabah memiliki telepon kantor
		FLAG_WORK_PHONE	Apakah nasabah memiliki telepon rumah
		FLAG_CONT_MOBILE	Apakah pnasabah memiliki ponsel
		FLAG_PHONE	Apakah nasabah memiliki telepon rumah
		FLAG_EMAIL	Status kepemilikkan email
		OCCUPATION_TYPE	Pekerjaan apa yang dimiliki nasabah
		CNT_FAM_MEMBERS	Berapa banyak anggota keluarga yang dimiliki nasabah
		REGION_RATING_CLIENT	Peringkat untuk wilayah tempat nasabah tinggal (1,2,3)
		REGION_RATING _CLIENT_W_CITY	Peringkat untuk wilayah tempat tinggal nasabah dengan mempertimbangkan kota (1,2,3)
		WEEKDAY_APPR _PROCESS_START	Pada hari apa nasabah mengajukan pinjaman
		HOUR_APPR _PROCESS_START	Kira-kira pada jam berapa nasabah mengajukan pinjaman
		REG_REGION_NOT _LIVE_REGION	Penanda alamat permanen nasabah tidak cocok dengan alamat kontak
		REG_REGION_NOT_WORK _REGION	Penanda alamat permanen nasabah tidak cocok dengan alamat kantor

No.	Dataset	Nama Atribut	Deskripsi
		LIVE_REGION_NOT _WORK_REGION	Penanda alamat kontak nasabah tidak cocok dengan alamat kantor
		REG_CITY_NOT_LIVE_CITY	Penanda alamat permanen nasabah tidak cocok dengan alamat kontak
		REG_CITY_NOT_WORK _CITY	Penanda alamat tetap nasabah tidak sesuai dengan alamat kantor
		LIVE_CITY_NOT_WORK _CITY	Penanda alamat kontak nasabah tidak sesuai dengan alamat kantor
		ORGANIZATION_TYPE	Jenis organisasi tempat nasabah bekerja
		EXT_SOURCE_1	Skor yang dinormalisasi dari sumber data eksternal
		EXT_SOURCE_2	Skor yang dinormalisasi dari sumber data eksternal
		EXT_SOURCE_3	Skor yang dinormalisasi dari sumber data eksternal
		APARTMENTS_AVG	Informasi yang dinormalisasi tentang bangunan tempat tinggal nasabah
		BASEMENTAREA_AVG	Informasi yang dinormalisasi tentang bangunan tempat tinggal nasabah
		YEARS _BEGINEXPLUATATION _AVG	Informasi yang dinormalisasi tentang bangunan tempat tinggal nasabah
		YEARS_BUILD_AVG	Informasi yang dinormalisasi tentang bangunan tempat tinggal nasabah

No.	Dataset	Nama Atribut	Deskripsi
		COMMONAREA_AVG	Informasi yang dinormalisasi tentang bangunan tempat tinggal nasabah
		ELEVATORS_AVG	Informasi yang dinormalisasi tentang bangunan tempat tinggal nasabah
		ENTRANCES_AVG	Informasi yang dinormalisasi tentang bangunan tempat tinggal nasabah
		FLOORSMAX_AVG	Informasi yang dinormalisasi tentang bangunan tempat tinggal nasabah
		FLOORSMIN_AVG	Informasi yang dinormalisasi tentang bangunan tempat tinggal nasabah
		LANDAREA_AVG	Informasi yang dinormalisasi tentang bangunan tempat tinggal nasabah
		LIVINGAPARTMENTS_AVG	Informasi yang dinormalisasi tentang bangunan tempat tinggal nasabah
		LIVINGAREA_AVG	Informasi yang dinormalisasi tentang bangunan tempat tinggal nasabah
		NONLIVINGAPARTMENTS _AVG	Informasi yang dinormalisasi tentang bangunan tempat tinggal nasabah
		NONLIVINGAREA_AVG	Informasi yang dinormalisasi tentang bangunan tempat tinggal nasabah
		APARTMENTS_MODE	Informasi yang dinormalisasi tentang bangunan tempat tinggal nasabah

No.	Dataset	Nama Atribut	Deskripsi
		BASEMENTAREA_MODE	Informasi yang dinormalisasi tentang bangunan tempat tinggal nasabah
		YEARS _BEGINEXPLUATATION _MODE	Informasi yang dinormalisasi tentang bangunan tempat tinggal nasabah
		YEARS_BUILD_MODE	Informasi yang dinormalisasi tentang bangunan tempat tinggal nasabah
		COMMONAREA_MODE	Informasi yang dinormalisasi tentang bangunan tempat tinggal nasabah
		ELEVATORS_MODE	Informasi yang dinormalisasi tentang bangunan tempat tinggal nasabah
		ENTRANCES_MODE	Informasi yang dinormalisasi tentang bangunan tempat tinggal nasabah
		FLOORSMAX_MODE	Informasi yang dinormalisasi tentang bangunan tempat tinggal nasabah
		FLOORSMIN_MODE	Informasi yang dinormalisasi tentang bangunan tempat tinggal nasabah
		LANDAREA_MODE	Informasi yang dinormalisasi tentang bangunan tempat tinggal nasabah
		LIVINGAPARTMENTS _MODE	Informasi yang dinormalisasi tentang bangunan tempat tinggal nasabah
		LIVINGAREA_MODE	Informasi yang dinormalisasi tentang bangunan tempat tinggal nasabah

No.	Dataset	Nama Atribut	Deskripsi
		NONLIVINGAPARTMENTS _MODE	Informasi yang dinormalisasi tentang bangunan tempat tinggal nasabah
		NONLIVINGAREA _MODE	Informasi yang dinormalisasi tentang bangunan tempat tinggal nasabah
		APARTMENTS_MEDI	Informasi yang dinormalisasi tentang bangunan tempat tinggal nasabah
		BASEMENTAREA_MEDI	Informasi yang dinormalisasi tentang bangunan tempat tinggal nasabah
		YEARS _BEGINEXPLUATATION _MEDI	Informasi yang dinormalisasi tentang bangunan tempat tinggal nasabah
		YEARS_BUILD_MEDI	Informasi yang dinormalisasi tentang bangunan tempat tinggal nasabah
		COMMONAREA_MEDI	Informasi yang dinormalisasi tentang bangunan tempat tinggal nasabah
		ELEVATORS_MEDI	Informasi yang dinormalisasi tentang bangunan tempat tinggal nasabah
		ENTRANCES_MEDI	Informasi yang dinormalisasi tentang bangunan tempat tinggal nasabah
		FLOORSMAX_MEDI	Informasi yang dinormalisasi tentang bangunan tempat tinggal nasabah
		FLOORSMIN_MEDI	Informasi yang dinormalisasi tentang bangunan tempat tinggal nasabah

No.	Dataset	Nama Atribut	Deskripsi
		LANDAREA_MEDI	Informasi yang dinormalisasi tentang bangunan tempat tinggal nasabah
		LIVINGAPARTMENTS _MEDI	Informasi yang dinormalisasi tentang bangunan tempat tinggal nasabah
		LIVINGAREA_MEDI	Informasi yang dinormalisasi tentang bangunan tempat tinggal nasabah
		NONLIVINGAPARTMENTS _MEDI	Informasi yang dinormalisasi tentang bangunan tempat tinggal nasabah
		NONLIVINGAREA _MEDI	Informasi yang dinormalisasi tentang bangunan tempat tinggal nasabah
		FONDKAPREMONT _MODE	Informasi yang dinormalisasi tentang bangunan tempat tinggal nasabah
		HOUSETYPE_MODE	Informasi yang dinormalisasi tentang bangunan tempat tinggal nasabah
		TOTALAREA_MODE	Informasi yang dinormalisasi tentang bangunan tempat tinggal nasabah
		WALLSMATERIAL _MODE	Informasi yang dinormalisasi tentang bangunan tempat tinggal nasabah
		EMERGENCYSTATE _MODE	Informasi yang dinormalisasi tentang bangunan tempat tinggal nasabah
		OBS_30_CNT_SOCIAL _CIRCLE	Jumlah observasi lingkungan sosial nasabah pada 30 hari

No.	Dataset	Nama Atribut	Deskripsi
		DEF_30_CNT_SOCIAL _CIRCLE	Jumlah observasi lingkungan sosial nasabah yang gagal pada 30 hari
		OBS_60_CNT_SOCIAL _CIRCLE	Jumlah observasi lingkungan sosial nasabah pada 60 hari
		DEF_60_CNT_SOCIAL _CIRCLE	Jumlah observasi lingkungan sosial nasabah yang gagal pada 60 hari
		DAYS_LAST_PHONE _CHANGE	Berapa hari sebelum pengajuan, nasabah mengganti telepon
		FLAG_DOCUMENT_2	Status dokumen 2
		FLAG_DOCUMENT_3	Status dokumen 3
		FLAG_DOCUMENT_4	Status dokumen 4
		FLAG_DOCUMENT_5	Status dokumen 5
		FLAG_DOCUMENT_6	Status dokumen 6
		FLAG_DOCUMENT_7	Status dokumen 7
		FLAG_DOCUMENT_8	Status dokumen 8
		FLAG_DOCUMENT_9	Status dokumen 9
		FLAG_DOCUMENT_10	Status dokumen 10
		FLAG_DOCUMENT_11	Status dokumen 11
		FLAG_DOCUMENT_12	Status dokumen 12
		FLAG_DOCUMENT_13	Status dokumen 13
		FLAG_DOCUMENT_14	Status dokumen 14
		FLAG_DOCUMENT_15	Status dokumen 15
		FLAG_DOCUMENT_16	Status dokumen 16
		FLAG_DOCUMENT_17	Status dokumen 17
		FLAG_DOCUMENT_18	Status dokumen 18
		FLAG_DOCUMENT_19	Status dokumen 19
		FLAG_DOCUMENT_20	Status dokumen 20
		FLAG_DOCUMENT_21	Status dokumen 21

No.	Dataset	Nama Atribut	Deskripsi
		AMT_REQ_CREDIT _BUREAU_HOUR	Jumlah pertanyaan ke Biro Kredit tentang nasabah satu jam sebelum permohonan
		AMT_REQ_CREDIT _BUREAU_DAY	Jumlah pertanyaan ke Biro Kredit tentang nasabah satu hari sebelum permohonan
		AMT_REQ_CREDIT _BUREAU_WEEK	Jumlah pertanyaan ke Biro Kredit tentang nasabah satu minggu sebelum permohonan
		AMT_REQ_CREDIT _BUREAU_MON	Jumlah pertanyaan ke Biro Kredit tentang nasabah satu bulan sebelum permohonan
		AMT_REQ_CREDIT _BUREAU_QRT	Jumlah pertanyaan ke Biro Kredit tentang nasabah 3 bulan sebelum permohonan
		AMT_REQ_CREDIT _BUREAU_YEAR	Jumlah pertanyaan ke Biro Kredit tentang nasabah satu tahun sebelum permohonan
2.	Bureau	SK_ID_CURR	ID pinjaman di Home Credit saat ini saat ini
		SK_BUREAU_ID	ID kredit yang ada pada Biro Kredit
		CREDIT_ACTIVE	Status Biro Kredit melaporkan kredit
		CREDIT_CURRENCY	Mata uang kredit Biro Kredit yang dikodekan ulang
		DAYS_CREDIT	Berapa hari sebelum permohonan saat ini nasabah mengajukan permohonan kredit Biro Kredit
		CREDIT_DAY_OVERDUE	Jumlah hari jatuh tempo kredit pada saat mengajukan

No.	Dataset	Nama Atribut	Deskripsi
			pinjaman terkait dalam sampel kami
		DAYS_CREDIT_ENDDATE	Sisa durasi kredit pada saat pengajuan di Home Credit
		DAYS_ENDDATE_FACT	Hari sejak kredit berakhir pada saat pengajuan di Home Credit
		AMT_CREDIT_MAX _OVERDUE	Jumlah maksimal yang telah jatuh tempo pada kredit Biro Kredit sejauh ini
		CNT_CREDIT_PROLONG	Berapa kali kredit Biro Kredit diperpanjang
		AMT_CREDIT_SUM	Jumlah kredit saat ini untuk kredit Biro Kredit
		AMT_CREDIT_SUM _DEBT	Hutang saat ini pada kredit Biro Kredit
		AMT_CREDIT_SUM_LIMIT	Batas kredit kartu kredit saat ini dilaporkan di Biro Kredit
		AMT_CREDIT_SUM _OVERDUE	Jumlah saat ini yang telah jatuh tempo pada kredit Biro Kredit
		CREDIT_TYPE	Jenis kredit pada Biro Kredit
		DAYS_CREDIT_UPDATE	Berapa hari sebelum permohonan pinjaman, informasi terakhir tentang kredit Biro Kredit datang
		AMT_ANNUITY	Anuitas kredit Biro Kredit
3.	Bureau Balance	SK_BUREAU_ID	ID kredit pada Biro Kredit
		MONTHS_BALANCE	Rincian bulanan
		STATUS	Status pinjaman Biro Kredit selama sebulan
4.	POS Cash Balance	SK_ID_PREV	ID kredit sebelumnya di Home Credit
		SK_ID_CURR	ID pinjaman Home Credit

No.	Dataset	Nama Atribut	Deskripsi
		MONTHS_BALANCE	Rincian bulanan
		CNT_INSTALMENT	Jangka waktu kredit sebelumnya
		CNT_INSTALMENT_FUTURE	Cicilan tersisa untuk dibayar pada kredit sebelumnya
		NAME_CONTRACT _STATUS	Status kontrak selama sebulan
		SK_DPD	Hari lewat jatuh tempo pada bulan kredit sebelumnya
		SK_DPD_DEF	Jatuh tempo selama bulan tersebut dengan toleransi
5.	Credit Card Balance	SK_ID_PREV	ID kredit sebelumnya di Home Credit
		SK_ID_CURR	ID pinjaman Home Credit
		MONTHS_BALANCE	Rincian bulanan
		AMT_BALANCE	Saldo selama bulan kredit sebelumnya
		AMT_CREDIT_LIMIT _ACTUAL	Limit kartu kredit selama bulan kredit sebelumnya
		AMT_DRAWINGS_ATM _CURRENT	Jumlah penarikan di ATM pada bulan kredit sebelumnya
		AMT_DRAWINGS_CURRENT	Jumlah penarikan selama bulan kredit sebelumnya
		AMT_DRAWINGS_OTHER _CURRENT	Jumlah penarikan lainnya selama bulan kredit sebelumnya
		AMT_DRAWINGS_POS _CURRENT	Jumlah penarikan atau pembelian barang selama bulan kredit sebelumnya
		AMT_INST_MIN _REGULARITY	Minimal cicilan bulan ini dari kredit sebelumnya
		AMT_PAYMENT _CURRENT	Berapa nasabah membayar selama sebulan pada kredit sebelumnya

No.	Dataset	Nama Atribut	Deskripsi
		AMT_PAYMENT_TOTAL _CURRENT	Berapa total yang dibayar nasabah selama bulan tersebut pada kredit sebelumnya
		AMT_RECEIVABLE _PRINCIPAL	Jumlah piutang pokok pada kredit sebelumnya
		AMT_RECIVABLE	Jumlah piutang pada kredit sebelumnya
		AMT_TOTAL _RECEIVABLE	Jumlah total piutang pada kredit sebelumnya
		CNT_DRAWINGS_ATM _CURRENT	Jumlah penarikan di ATM selama bulan ini pada kredit sebelumnya
		CNT_DRAWINGS _CURRENT	Jumlah penarikan selama bulan ini pada kredit sebelumnya
		CNT_DRAWINGS_OTHER _CURRENT	Jumlah penarikan lainnya selama bulan ini pada kredit sebelumnya
		CNT_DRAWINGS_POS _CURRENT	Jumlah penarikan barang selama bulan ini pada kredit sebelumnya
		CNT_INSTALMENT _MATURE_CUM	Jumlah angsuran yang dibayarkan pada kredit sebelumnya
		NAME_CONTRACT_STATUS	Status kontrak pada kredit sebelumnya
		SK_DPD	Hari Lewat Jatuh Tempo selama sebulan pada kredit sebelumnya
		SK_DPD_DEF	Hari Lewat Jatuh Tempo selama sebulan dengan toleransi (utang dengan jumlah pinjaman rendah

No.	Dataset	Nama Atribut	Deskripsi
			diabaikan) dari kredit sebelumnya
6.	Previous Application	SK_ID_PREV	ID kredit sebelumnya di Home Credit
		SK_ID_CURR	ID pinjaman Home Credit
		NAME_CONTRACT_TYPE	Jenis produk kontrak dari pengajuan sebelumnya
		AMT_ANNUITY	Anuitas pengajuan sebelumnya
		AMT_APPLICATION	Berapa jumlah pengajuan yang diminta nasabah pada permohonan sebelumnya
		AMT_CREDIT	Jumlah kredit akhir pada permohonan sebelumnya.
		AMT_DOWN_PAYMENT	Uang muka pada permohonan sebelumnya
		AMT_GOODS_PRICE	Harga barang yang diminta nasabah pada pengajuan sebelumnya
		WEEKDAY_APPR_PROCESS _START	Pada hari apa dalam seminggu nasabah mengajukan permohonan sebelumnya
		HOUR_APPR_PROCESS _START	Kira-kira pada jam berapa nasabah mengajukan permohonan sebelumnya
		FLAG_LAST_APPL_PER _CONTRACT	Tandai jika itu adalah permohonan terakhir untuk kontrak sebelumnya.
		NFLAG_LAST_APPL_IN _DAY	Tandai jika pengajuan tersebut merupakan pengajuan terakhir nasabah per hari.
		NFLAG_MICRO_CASH	Tandai Pinjaman keuangan mikro

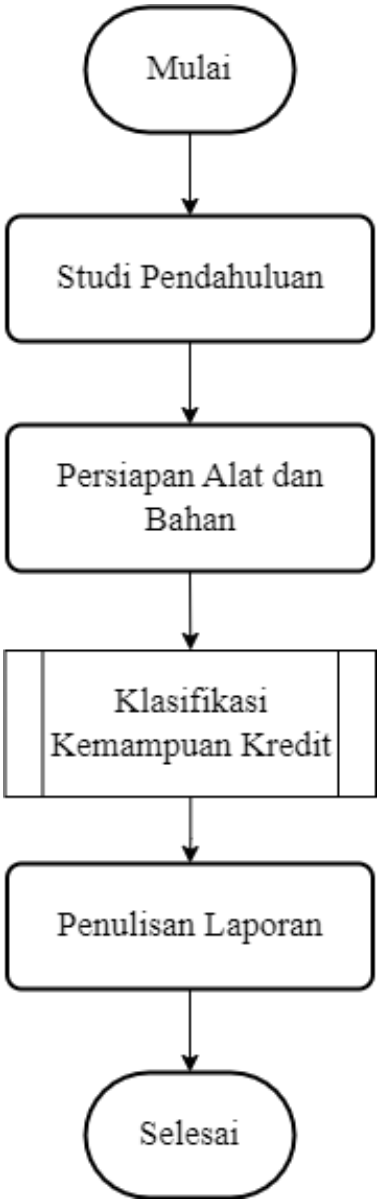
No.	Dataset	Nama Atribut	Deskripsi
		RATE_DOW _PAYMENT	Tingkat uang muka dinormalisasi pada kredit sebelumnya
		RATE_INTEREST_PRIMARY	Suku bunga dinormalisasi pada kredit sebelumnya
		RATE_INTEREST _PRIVILEGED	Suku bunga dinormalisasi pada kredit sebelumnya
		NAME_CASH_LOAN _PURPOSE	Tujuan pinjaman tunai
		NAME_CONTRACT _STATUS	Status kontrak dari permohonan sebelumnya
		DAYS_DECISION	Sehubungan dengan permohonan saat ini, kapan keputusan tentang permohonan sebelumnya dibuat
		NAME_PAYMENT_TYPE	Metode pembayaran yang dipilih nasabah untuk membayar pengajuan sebelumnya
		CODE_REJECT_REASON	Mengapa lamaran sebelumnya ditolak
		NAME_TYPE_SUITE	Pendamping nasabah saat melamar lamaran sebelumnya
		NAME_CLIENT_TYPE	Jenis nasabah pada pengajuan sebelumnya
		NAME_GOODS_CATEGORY	Barang apa saja yang diminta nasabah pada permohonan sebelumnya
		NAME_PORTFOLIO	Jenis portofolio untuk pengajuan
		NAME_PRODUCT_TYPE	Jenis produk yang diambil

No.	Dataset	Nama Atribut	Deskripsi
		CHANNEL_TYPE	Saluran tempat memperoleh nasabah pada pengajuan sebelumnya
		SELLERPLACE_AREA	Area penjualan tempat penjual pengajuan sebelumnya
		NAME_SELLER_INDUSTRY	Industri penjual
		CNT_PAYMENT	Jangka waktu kredit sebelumnya pada saat pengajuan permohonan sebelumnya
		NAME_YIELD_GROUP	Mengelompokkan suku bunga menjadi kecil menengah dan tinggi dari pengajuan sebelumnya
		PRODUCT_COMBINATION	Kombinasi produk terperinci dari pengajuan sebelumnya
		DAYS_FIRST_DRAWING	Waktu pencairan pertama permohonan sebelumnya
		DAYS_FIRST_DUE	Haei jatuh tempo pertama seharusnya dari permohonan sebelumnya
		DAYS_LAST_DUE_1ST _VERSION	Hari jatuh tempo pertama permohonan sebelumnya
		DAYS_LAST_DUE	Tanggal jatuh tempo terakhir permohonan sebelumnya
		DAYS_TERMINATION	Perkiraan penghentian permohonan sebelumnya
		NFLAG_INSURED_ON _APPROVAL	Apakah nasabah meminta asuransi pada pengajuan sebelumnya
7.	Instalment Payments	SK_ID_PREV	ID kredit sebelumnya di Home Credit
		SK_ID_CURR	ID pinjaman Home Credit

No.	Dataset	Nama Atribut	Deskripsi
		NUM_INSTALMENT _VERSION	Perubahan versi angsuran dari bulan ke bulan
		NUM_INSTALMENT _NUMBER	Di cicilan mana kami mengamati pembayarannya
		DAYS_INSTALMENT	Kapan angsuran kredit sebelumnya seharusnya dibayar
		DAYS_ENTRY _PAYMENT	Kapan sebenarnya angsuran kredit sebelumnya dibayarkan
		AMT_INSTALMENT	Jumlah angsuran yang ditentukan dari kredit sebelumnya pada angsuran ini
		AMT_PAYMENT	Jumlah yang sebenarnya dibayar nasabah pada kredit sebelumnya pada cicilan ini

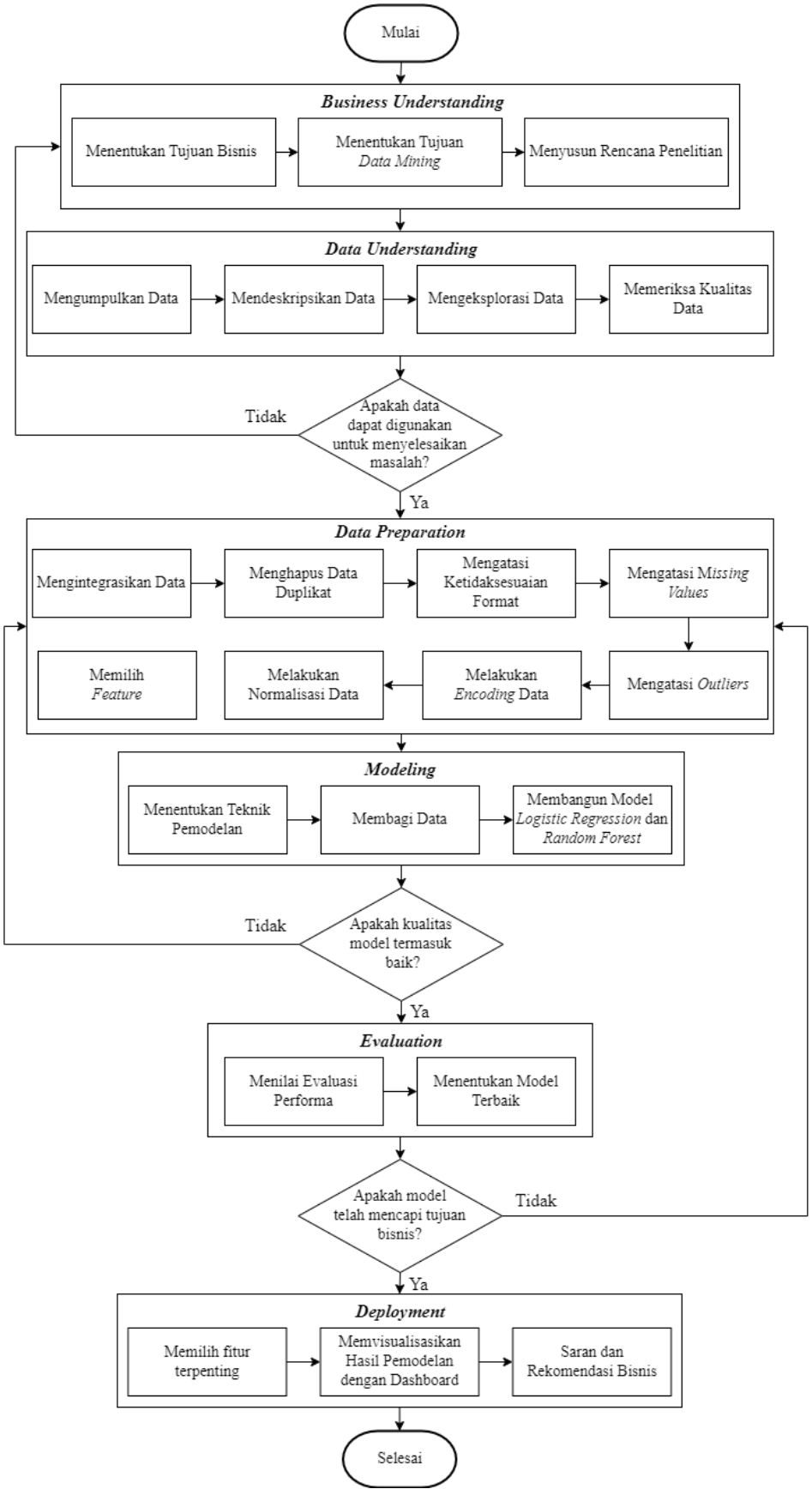
3.3 Tahapan Penelitian

Penelitian ini dilakukan melalui beberapa tahapan yang tertera pada gambar 8. Tahapan pertama dimulai dengan melakukan studi pendahuluan. Studi pendahuluan dilakukan dengan mencari dan mempelajari literasi yang berkaitan dengan penelitian dari beragam sumber, seperti buku dan jurnal. Hal tersebut dilakukan untuk mencari informasi dan teori dasar yang relevan dengan masalah yang akan diteliti. Tahap selanjutnya melakukan persiapan alat dan bahan yang digunakan untuk menunjang penelitian. Kemudian, klasifikasi nasabah dengan metode pengembangan *CRISP-DM* dilakukan untuk penilaian kemampuan pembayaran kredit nasabah menggunakan standar pemrosesan dalam proses *data mining*. Tahapan *CRISP-DM* meliputi beberapa proses, yaitu *Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation,* dan *Deployment*. Setelah semua proses pada *CRISP-DM* dilakukan, penelitian akan masuk ke tahap penulisan laporan untuk memaparkan seluruh proses kegiatan dan hasil penelitian yang telah dilakukan.



Gambar 8. *Flowchart* tahapan penelitian.

Penjelasan proses yang terjadi pada tahapan klasifikasi penilaian kemampuan pembayaran kredit nasabah dengan metode pengembangan *Cross Industry Standard Process for Data Mining* (*CRISP-DM*) menggunakan *machine learning logistic regression* dan *random forest classifier* ditampilkan pada gambar berikut :



Gambar 9. Tahapan klasifikasi dengan metode pengembangan *CRISP-DM*.

3.3.1 Tahapan *Business Understanding*

Tahapan pertama dalam metode *CRISP-DM* adalah tahapan *business understanding*. Pada tahapan ini ada beberapa kegiatan yang harus dilakukan, yaitu menentukan tujuan bisnis, menentukan tujuan *data mining*, dan menyusun rencana penelitian. Penelitian ini memiliki tujuan bisnis untuk mengetahui kemampuan pembayaran kredit pada nasabah yang mengajukan pinjaman. Hal tersebut perlu dilakukan untuk menghindari terjadinya kredit bermasalah atau kredit macet yang dapat mengancam alur pendapatan dan keberlangsungan bisnis. Pemanfaatan data nasabah dapat digunakan untuk menganalisis kemampuan nasabah dalam pembayaran kredit. Setiap informasi yang terdapat dalam data-data tersebut mempengaruhi penilaian penentuan kelayakan nasabah dalam menerima pinjaman. Oleh karena itu dengan data yang tersedia, penggunaan *data mining* dapat diterapkan untuk mencapai tujuan bisnis perusahaan. Dengan *data mining*, perusahaan dapat melakukan prediksi kemampuan membayar nasabah yang layak dan tidak layak untuk diberikan pinjaman. Hasil klasifikasi ini dapat digunakan perusahaan dalam mengambil keputusan pemberian pinjaman untuk menghindari terjadinya kredit bermasalah. Seluruh kegiatan tersebut disusun menjadi rencana penelitian yang digunakan sebagai strategi untuk mencapai tujuan bisnis.

3.3.2 Tahapan *Data Understanding*

Pada tahapan *data understanding* ada beberapa kegiatan yang harus dilakukan yaitu, mengumpulkan data, mendeskripsikan data, mengeksplorasi data, dan memeriksa kualitas data. Kegiatan pertama tahapan ini adalah mendapatkan dan memasukkan data agar dapat terbaca ke dalam *tools* yang digunakan untuk pembuatan model. Setelah itu, pendeskripsian data dilakukan untuk melihat gambaran atau kondisi atribut yang akan digunakan dalam penelitian, seperti melihat ukuran dan tipe data. Kegiatan selanjutnya adalah melakukan eksplorasi data untuk membantu memahami hubungan antar data, menemukan pola dalam data, dan menampilkan data agar dapat lebih mudah diinterpretasikan. Kegiatan terakhir pada tahapan *data understanding* adalah memeriksa kualitas data. Pemeriksaan ini dilakukan untuk menentukan apakah data yang telah ada dapat

digunakan dalam mencapai tujuan bisnis. Kemudian, pencarian solusi dilakukan untuk mengatasi masalah kualitas data yang dapat mengganggu analisis kemampuan pembayaran nasabah.

3.3.3 Tahapan *Data Preparation*

Pada tahap *data preparation* terdapat beberapa kegiatan yang dilalui, yaitu, pengintegrasian data, penghapusan data duplikat, penanganan nilai tidak sesuai dan hilang, pemilihan fitur, perubahan bentuk data, dan pengodean data. Tahapan ini menggunakan bantuan *tools* Jupyter Notebook dan bahasa pemrograman Python untuk mengolah data agar siap dilakukan *modeling*. Proses pertama tahapan ini adalah menggabungkan beberapa dataset menjadi sebuah *DataFram*e. Penghapusan atribut duplikat juga dilakukan untuk mengurangi dimensi data selama proses penggabungan data berlangsung. Selanjutnya juga dilakukan pengolahan pada data dengan format yang tidak sesuai. Untuk mengatasi masalah tersebut, atribut yang memiliki tipe data tidak sesuai akan diubah. Kemudian, persiapan juga dilakukan untuk menangani atribut dengan memiliki *missing values* yang dapat mengurangi potensi kesalahan hasil analisis. Penanganan *missing values* dilakukan melaui penghapusan atribut dan pengisian nilai. Setelah itu, proses penanganan *outliers* dilakukan karena *outliers* menyebabkan nilai pada atribut tidak proporsional. Kemudian, proses normalisasi data dilakukan untuk menyamakan atribut dengan *record* numerik ke dalam skala yang sama sehingga dapat mengurangi redudansi pada data. Proses selanjutnya pada adalah melakukan *encoding*. *Encoding* dilakukan untuk mengubah data kategorikal menjadi data numerikal sehingga dapat dipahami dan diolah dengan *machine learning*. Proses terakhir dalam tahapan *data preparation* adalah melakukan pemilihan *feature* dengan tujuan mengurangi dimensi data. Proses pemilihan *feature* dilakukan dengan menghitung nilai korelasi antara atribut bebas dengan atribut terikat dan menghapus atribut yang memiliki nilai korelasi di bawah ambang batas. Metode pemilihan *feature* yang digunakan pada penelitian ini adalah *correlation* untuk melihat besar hubungan antara semua atribut dengan atribut target yang akan diprediksi.

3.3.4 Tahapan *Modeling*

Pada tahapan *modeling*, proses pertama yang dilakukan adalah menentukan algoritma *machine learning* yang akan digunakan. Penentuan tersebut diperoleh dari pengamatan pada penelitian lain yang *model machine learning*-nya memiliki hasil paling baik. Berdasarkan pengamatan tersebut, penelitian ini menggunakan algoritma *Logistic Regression* dan *Random Forest*. Proses selanjutnya dalam *modeling* adalah melakukan pembagian data. Atribut yang telah dipilih untuk proses *modeling* pada tahapan *data preparation* akan dipisah menjadi dua variabel baru. Variabel pertama adalah variabel latih yang berisi semua atribut bebas. Sementara variabel kedua adalah variabel target yang berisi atribut terikat. Setelah itu, seluruh *record* yang ada pada data akan dibagi menjadi data latih dan data uji secara random. Selanjutnya dilakukan proses penyeimbangan data untuk mengatasi masalah ketidakseimbangan kelas dengan menggunakan teknik *SMOTE*. Setelah pembagian data selesai, tahap selanjutnya adalah melakukan pemodelan dengan menggunakan dua teknik, yaitu *logistic regression* dan *random forest*. *Logistic regression* merepresentasikan pola hubungan antara sekumpulan atribut bebas dan atribut terikat menggunakan fungsi yang dibentuk dengan menyamakan nilai Y pada *linear function* dengan nilai Y pada *sigmoid function.* Sementara, *random forest* merupakan pemodelan yang menggunakan sekumpulan *decision tree* yang membentuk sebuah hutan klasifikasi. Setiap *decision tree* dibangun berdasarkan perhitungan nilai *entropy* dan *gain* dan menghasilkan *rule*. Hasil klasifikasi dengan *random forest* ditentukan berdasarkan hasil *voting* dari setiap suara yang diberikan oleh *decision tree* yang terbentuk.

3.3.5 Tahapan *Evaluation*

Pada tahap *evaluation* dilakukan penilaian perfoma dari hasil pemodelan yang telah dilakukan oleh *logistic regression* dan *random forest*. Penilaian perfoma dilakukan untuk menilai seberapa baik prediksi model yang dibangun dalam melakukan klasifikasi kemampuan pembayaran nasabah. Beberapa parameter yang digunakan untuk mengukur kinerja klasifikasi adalah *confusion matrix, accuracy*, *precision*, *recall*, *F1-score*, dan *AUC* pada algoritma *logistic regression* dan *random forest*.

Hasil evaluasi dengan parameter antara kedua algoritma tersebut akan dibandingkan untuk memilih satu model algoritma terbaik yang dapat diterapkan untuk memprediksi dan menilai kemampuan pembayaran kredit seorang nasabah. Jika hasil model dianggap telah mampu mencapai tujuan bisnis maka proses *deployment* akan dilakukan. Namun, jika model tersebut dirasa belum mampu mencapai tujuan bisnis dalam melakukan klasifikasi dengan baik maka tahapan *data preparation* dapat diulang untuk melakukan persiapan data dengan cara lain hingga didapatkan model dengan perfoma terbaik.

3.3.6 Tahapan *Deployment*

Pada tahap *deployment*, proses penyebaran informasi dan pengetahuan yang telah didapatkan dengan melakukan pembuatan *dashboard* menggunakan Google Looker Studio. *Dashboard* yang dikembangkan akan memvisualisasikan atribut-atribut pada model terbaik. Hasil visualisasi yang ditampilkan pada *dashboard* diharapkan bermanfaat dan dapat digunakan sebagai saran dalam penentuan kriteria nasabah seperti apa yang layak dan tidak layak diberikan pinjaman kredit. *Dashboard* ini juga dapat menampilkan atribut-atribut apa saja yang paling mempengaruhi penilaian pemberian kredit pada nasabah. Selain itu, *dashboard* hasil *modeling* ini juga dapat dimanfaatkan oleh pihak Home Credit untuk melihat persebaran kredit dan informasi konsumen yang dimiliki.

BAB V
KESIMPULAN DAN SARAN

5.1 Kesimpulan

Berdasarkan hasil dan analisis dari penelitian yang telah dilakukan dapat disimpulkan beberapa hal sebagai berikut.

1. Berdasarkan hasil membandingkan metode klasifikasi tanpa *SMOTE* dan dengan *SMOTE* dalam mengatasi ketidakseimbangan kelas pada data diperoleh hasil bahwa metode *SMOTE* berpengaruh dalam meningkatkan hasil penilaian kemampuan pembayaran kredit dengan memperoleh nilai evaluasi *AUC* lebih besar di kedua algoritma, yaitu *Logistic Regression* sebesar 0.74 dan *Random Forest* sebesar 0.97.
2. Berdasarkan hasil dari evaluasi perfoma yang diperoleh antara dua algoritma *machine learning* yang digunakan, algoritma dengan kinerja terbaik untuk melakukan penilaian kemampuan pembayaran kredit adalah model *random forest classifier* dengan *SMOTE* yang mana diperoleh nilai *accuracy* 90%, *precision* sebesar 92%, *recall* sebesar 88%, *F1-score* sebesar 90%, dan nilai *AUC* sebesar 0.97.
3. Berdasarkan *importance feature* dari model *random forest classifier* didapatkan sepuluh atribut berpengaruh, yaitu atribut skor yang dinormalisasi dari sumber data eksternal ke-3, skor yang dinormalisasi dari sumber data eksternal ke-2, rentang waktu nasabah mengganti nomor, jumlah pembayaran cicilan pada kredit sebelumnya di Home Credit, usia nasabah, waktu registrasi, rentang waktu mengajukan kredit di biro kredit, rentang waktu nasabah mengganti

dokumen identitas, waktu informasi nasabah diperbarui oleh biro kredit, dan lama waktu nasabah bekerja.

4. Proses *deployment* berhasil dilakukan dengan menampilkan hasil prediksi *application test* menggunakan algoritma *random forest classifier* secara visual melalui pembuatan *dashboard* dengan menggunakan Google Looker Studio.

5.2 Saran

Saran yang dapat diberikan untuk penelitian selanjutnya berdasarkan penelitian yang telah dilakukan adalah sebagai berikut.

1. Melakukan *feature construction* dengan mengombinasikan atribut yang telah ada untuk menghasilkan atribut baru yang memiliki kontribusi signifikan terhadap keakuratan prediksi.
2. Melakukan pengujian pengaruh pemilihan *feature* dengan metode *correlation* pada pemodelan.
3. Melakukan pengembangan model dengan menerapkan *hyperparameter tuning* untuk mencari nilai-nilai yang optimal pada parameter-parameter yang digunakan dalam setiap model.

DAFTAR PUSTAKA

[1] M. Fuady, *Hukum Tentang Pembiayaan dalam Teori dan Praktek*. Bandung: Citra Aditya Bakti, 1999.

[2] Kasmir, *Manajemen Perbankan*, Revised Edition. Jakarta: Rajawali Pers, 2015.

[3] J. Han, M. Kamber, and P. Jian, *Data Mining: Concepts and Techniques*, Third Edition. Waltham: Morgan Kaufmann, 2012.

[4] Y. Ramadhani, "Rancang Bangun Aplikasi Untuk Memprediksi Kelas Resiko Pemberian Kredit dengan Menggunakan Metode Naïve Bayes Classifier," Sekolah Tinggi Manajemen Informatika dan Teknik Komputer, Surabaya, Skripsi, 2010.

[5] Amna *et al.*, *Data Mining*, First Edition. Padang: PT Global Eksekutif Teknologi, 2023.

[6] B. Bawono and R. Wasono, "Perbandingan Metode Random Forest dan Naïve Bayes Untuk Klasifikasi Debitur Berdasarkan Kualitas Kredit," in *Seminar Nasional Edusaintek*, Semarang: Universitas Muhammadiyah Semarang, 2019, pp. 343–348.

[7] A. P. Wibawa, M. G. A. Purnama, M. F. Akbar, and F. A. Dwiyanto, "Metode-metode Klasifikasi," in *Seminar Ilmu Komputer dan Teknologi Informasi*, in 1, vol. 3. Malang: Universitas Negeri Malang, Mar. 2018, pp. 134–138.

[8] Sukarna and St. R. A. Nirwana, "Aplikasi Regresi Logistik Multinomial dalam Menentukan Faktor-Faktor yang Mempengaruhi Pemilihan Program Studi di Jurusan Matematika FMIPA UNM," *J. Math. Stat.*, vol. 01, pp. 65–72, 2015.

[9] H. Rianto and R. S. Wahono, "Resampling Logistic Regression untuk Penanganan Ketidakseimbangan Class pada Prediksi Cacat Software," *J. Softw. Eng.*, vol. 1, pp. 46–53, 2015.

[10] R. M. Awangga and N. H. Khonsa, "Analisis Performa Algoritma Random Forest dan Naive Bayes Multinomial pada Dataset Ulasan Obat dan Ulasan Film," *J. Telekomun. Dan Komput.*, vol. 12, pp. 60–70, Apr. 2022, doi: 10.22441/incomtech.v12i1.14770.

[11] A. Montoya, KrillOdintsov, and M. Kotek, “Home Credit Default Risk.” Kaggle, 2018. Accessed: Dec. 01, 2022. [Online]. Available: https://kaggle.com/competitions/home-credit-default-risk

[12] Home Credit contributors, “Home Credit,” Home Credit. Accessed: Oct. 31, 2023. [Online]. Available: https://www.homecredit.co.id/tentang-perusahaan

[13] Pemerintah Indonesia, “Undang-Undang Republik Indonesia Nomor 10 Tahun 1998 Tentang Perbankan.” Sekertariat Negara, 1998. Accessed: Oct. 20, 2023. [Online]. Available: https://www.bphn.go.id/data/documents/98uu010.pdf

[14] M. D. Badrulzaman, *Perjanjian Kredit Bank*. Bandung: Citra Aditya Bakti, 1983.

[15] Ismail, *Manajemen Perbankan*, Edisi Pertama. Jakarta: Kencana, 2010.

[16] S. P. Robbins and M. Coulter, *Management*, Sixth Edition. Jakarta: PT Prenhallindo, 1999.

[17] F. A. Hermawati, *Data Mining*, Ed. 1. Yogyakarta: Andi, 2013.

[18] D. E. Goldberg, Holland, and J. Henry, “Genetic Algorithms and Machine Learning,” vol. 3, no. 2, pp. 95–99, 1988.

[19] K. P. Murphy, *Machine Learning : A Probabilistic Perspective*. The MIT Press, 2012.

[20] R. Maulid, “Variasi Jenis Algoritma Machine Learning, Sudah Tahu?,” DQLab. Accessed: Oct. 27, 2023. [Online]. Available: https://dqlab.id/jenis-metode-regresi-algoritma-supervised-learning

[21] K. Wibowo, “Classification 1 : In-class Materials,” RPubs. Accessed: Oct. 31, 2023. [Online]. Available: https://rpubs.com/algokev/c_one

[22] R. Yehoshua, “Random Forest,” Medium. Accessed: Nov. 03, 2023. [Online]. Available: https://medium.com/@roiyeho/random-forests-98892261dc49

[23] R. Fitriana, A. N. Habyba, and E. Febrianti, *Data Mining dan Aplikasinya Contoh Kasus di Industri Manufaktur dan Jasa*, Edisi Pertama. Banyumas: Wawasan Ilmu, 2022.

[24] D. Barapatre and V. A, “Data Preparation On Large Datasets For Data Science,” *Asian J. Pharm. Clin. Res.*, vol. 10, pp. 458–488, 2017.

[25] K. Pavya and B. Srinivasan, “Feature Selection Techniques in Data Mining: A Study,” *Int. J. Sci. Dev. Res.*, vol. 2, no. 6, pp. 594–598, 2017.

[26] Michael J de Smith, *A Comprehensive Handbook of Statistical Concepts, Techniques and Software Tools*. London: The Winchelsea Press, 2018.

[27] H. Junaedi, H. Budianto, and Y. Melani, "Data Transformation pada Data Mining," in *Konferensi Nasional Inovasi dalam Desain dan Teknologi*, Surabaya: Sekolah Tinggi Teknik Surabaya, 2011, pp. 93–99.

[28] E. Sutoyo and M. A. Fadlurrahman, "Penerapan SMOTE untuk Mengatasi Imbalance Class dalam Klasifikasi Television Advertisement Performance Rating Menggunakan Artificial Neural Network," *J. Edukasi Dan Penelit. Inform.*, vol. 6, pp. 379–385, 2020.

[29] "Compare Deep Learning Models Using ROC Curves," Mathworks. Accessed: Dec. 10, 2023. [Online]. Available: https://www.mathworks.com/help/deeplearning/ug/compare-deep-learning-models-using-ROC-curves

[30] F. Gorunescu, *Data Mining Concepts, Models and Techniques*, First Edition. German: Springer-Verlag Berlin Heidelberg, 2011. [Online]. Available: https://doi.org/10.1007/978-3-642-19721-5

[31] Algoritma Team, "Library Python Populer," Algoritma. Accessed: Oct. 27, 2023. [Online]. Available: https://algorit.ma/blog/library-python/

[32] Amrin and O. Pahlevi, "Implementasi Algoritma Klasifikasi Logistic Regression dan Naïve Bayes untuk Diagnosa Penyakit Hepatitis," *J. Tek. Komput. AMIK BSI*, vol. 8, pp. 162–167, Jul. 2022, doi: 10.31294/jtk.v4i2.

[33] W. O. Simanjuntak, A. B. P. Negara, and R. Septriana, "Perbandingan Algoritma Logistic Regression dan Random Foret (Studi Kasus : Klasifikasi Emosi Tweet)," *J. Apl. Dan Ris. Inform.*, vol. 2, pp. 160–164, Agustus 2023, doi: 10.26418/juara.v2i1.69682.

[34] M. F. Mujaddid, S. Al-Faraby, and Adiwijaya, "Analisis Churn Prediction Menggunakan Metode Logistic Regression dan SMOTE (Synthetic Minority Over-sampling Technique) Pada Perusahaan Telekomunikasi," *Univ. Telkom*, vol. 4, pp. 5046–5054, 2017.

[35] L. Zhu, D. Qiu, D. Ergu, C. Ying, and K. Liu, "A Study on Predicting Loan Default Based on the Random Forest Algorithm," presented at the 7th International Conference on Information Technology and Quantitative Management (ITQM 2019), Granada: Elseiver B.V., 2019, pp. 503–513.

[36] M. A. Khadija and N. A. Setiawan, "Detecting Liver Disease Diagnosis by Combining SMOTE,Information Gain Attribute Evaluation and Ranker," *J. Ilm. Teknol. Dan Inf.*, vol. 9, pp. 13–17, Jun. 2020.

[37] M. I. C. Rachmatullah, "Penerapan SMOTE untuk Meningkatan Kinerja Klasifikasi Penilaian Kredit," *J. Ris. Komput.*, vol. 10, pp. 302–309, Feb. 2023, doi: 10.30865/jurikom.v10i1.5612.

[38] H. D. Septama, T. Yulianti, D. Budiyanto, S. M. Mulyadi, and A. H. Cahyana, "A Comparative Analysis of Machine Learning Algorithms for Credit Risk Scoring using Chi-Square Feature Selection," in *2023 International Conference on Converging Technology in Electrical and Information Engineering (ICCTEIE)*, Bandar Lampung, Indonesia: IEEE, Oct. 2023, pp. 32–37. doi: 10.1109/ICCTEIE60099.2023.10366576.

[39] A. Siddique, M. A. Khan, and Z. Khan, "The effect of credit risk management and bank-specific factors on the financial performance of the South Asian commercial banks," *Asian J. Account. Res.*, vol. 7, no. 2, pp. 182–194, May 2022, doi: 10.1108/AJAR-08-2020-0071.