Jutisi: Jurnal Ilmiah Teknik Informatika dan Sistem Informasi
Jl. Ahmad Yani, K.M. 33,5 - Kampus STMIK Banjarbaru
Loktabat – Banjarbaru (Tlp. 0511 4782881), e-mail: puslit.stmikbjb@gmail.com
e-ISSN: 2685-0893
p-ISSN: 2089-3787

Uji Akurasi Algoritme *K-Nearest Neighbor* Dan *Naïve Bayes* Dalam Klasifikasi Kelayakan Pemberian Kredit Perbankan

Bisma Asyari[1*], Syarifah Putri Agustini Alkadri[2], Putri Yuli Utami[3]
[1,2]Teknik Informatika, Universitas Muhammadiyah Pontianak, Pontianak, Indonesia
[3] Sistem Informasi, Universitas Muhammadiyah Pontianak, Pontianak, Indonesia
*e-mail *Corresponding Author:* bismaasyari@gmail.com

Abstract

Banking credit is a process of giving money or debt following an agreement between the borrower and the bank, as well as determining the classification of creditworthiness on housing loans (KPR). This affects the customer's waiting time for the results of a bank's decision, the success of a bank's credit management will greatly affect the fate of many customer funds if the analysis is inaccurate, so technology is needed to find hidden information on prospective borrowers' data to predict a customer's loan repayment ability. This study uses an algorithm K-Nearest Neighbor and Naïve Bayes to determine the eligibility classification for bank lending and determine the accuracy of bank credit eligibility for mortgages, to determine the accuracy of the algorithm through three stages of testing, namely several preprocessing stages starting from checking duplicates, deal with missing value, deal with outliers, do label encoding, deal with data imbalance use method SMOTE, and standardize using scaler standard. The results of the Naïve Bayes and KNN algorithms as well as the model stages are evaluated to examine each stage in the data for the model's ability to predict, the evaluation matrix used is in the form of a results confusion matrix. There is the best result, namely the KNN algorithm in the third test with a value of K = 10 with a performance of 80.92% training data accuracy and 78.86% testing data and getting a score confusion matrix TP 76 and TN 21.

Keywords: *Banking Credit; Data Mining; Machine Learning; K-Nearest Neighbors; Naïve Bayes.*

Abstrak

Kredit perbankan suatu proses pemberian uang atau hutang sesuai dengan kesepakatan antara peminjam dengan bank, serta menentukan klasifikasi kelayakan kredit pada Kredit Pemilikan Rumah (KPR). Hal ini mempengaruhi waktu tunggu nasabah atas hasil keputusan bank, keberhasilan pengelolaan kredit suatu bank akan sangat mempengaruhi nasib banyak dana nasabah jika analisisnya tidak akurat, sehingga dibutuhkan teknologi untuk menemukan informasi tersembunyi data calon peminjam untuk memprediksi kemampuan pembayaran pinjaman nasabah. Penelitian ini menggunakan algoritme *K-Nearest Neighbor* dan *Naïve Bayes* untuk menentukan klasifikasi kelayakan pemberian kredit perbankan dan mengetahui tingkat akurasi kelayakan pemberian kredit perbankan pada KPR, untuk mengetahui tingkat akurasi algoritme melalui tiga tahap pengujian, yaitu dilakukan beberapa tahapan preprocessing mulai dari pengecekan *duplicate,* menangani *missing value*, menangani *outliers,* melakukan *label encoding,* mengatasi data *imbalance* menggunakan metode *SMOTE,* dan melakukan standarisasi menggunakan *standar scaler.* Hasil dari algoritme *Naïve Bayes* dan KNN serta tahapan model di evaluasi untuk memeriksa setiap tahap pada data terhadap kemampuan model dalam memprediksi, matrik evaluasi yang digunakan berupa hasil *confusion matrix*. Terdapat hasil terbaik yaitu pada algoritme KNN di pengujian ketiga dengan nilai K=10 dengan performa akurasi data training 80.92% dan data testing 78.86% dan mendapatkan *score confusion matrix* TP 76 dan TN 21.

Kata Kunci: *Kredit Perbankan; Data Mining; Machine Learning; K-Nearest Neighbors; Naïve Bayes.*

1. Pendahuluan

Kredit perbankan merupakan proses memberikan uang atau hutang berdasarkan perjanjian atau kesepakatan pinjam meminjam antara bank, lembaga keuangan dan pihak lain, yang mewajibkan peminjam untuk melunasi hutangnya setelah jangka waktu tertentu dengan disertai bunga. Bank tentunya harus teliti dalam memilih calon debitur untuk meminimalisir risiko kredit. Dalam risiko kredit, jangka waktu merupakan masalah utama kredit dan masalah yang kompleks. Faktor utamanya dari masalah ini yaitu kurangnya penilaian awal yang akurat dari calon debitur [1].

Perkembangan perbankan dinilai sebagai pilar yang sangat penting bagi perekonomian dunia, khususnya Indonesia, Bank juga menawarkan kemudahan menyediakan dana pinjaman yang digunakan nasabah dalam bentuk investasi, yaitu melalui KPR di Indonesia. Selain itu, klasifikasi kelayakan kredit masih dilakukan secara manual, dan dalam beberapa kasus, manajemen bank berupa proses kredit dan perhitungan subsidi pinjaman menggunakan cara konvensional [2]. Hal ini mempengaruhi waktu tunggu nasabah atas hasil keputusan bank sehingga kurang efisien dalam pelaksanaanya. Keberhasilan pengelolaan kredit suatu bank akan sangat mempengaruhi nasib banyak dana nasabah jika analisisnya tidak akurat, sehingga dibutuhkan teknologi untuk menemukan informasi tersembunyi dalam data calon peminjam dan membantu kreditur guna memprediksi kemampuan pembayaran pinjaman KPR nasabah. Syarat atau kriteria sebagai penguat bank dalam proses peminjaman KPR pada BUMN khususnya Bank Mandiri dan BNI, yakni jenis kelamin berpengaruh terhadap peminjaman, status perkawinan mempengaruhi kepercayaan terhadap bank, tanggungan berupa banyaknya jumlah keluarga inti yang ditanggung, pekerjaan sebagai karyawan atau pemilik bisnis berpengaruh dalam jumlah pinjaman yang akan diberikan oleh bank, pendapatan merupakan kepercayaan bank dalam menentukan besarnya peminjaman untuk karyawan atau pemilik bisnis, jumlah pinjaman juga berpengaruh terhadap karyawan dan pemilik bisnis, durasi pinjaman yang dibebankan kepada calon nasabah memiliki rentang waktu, riwayat kredit calon nasabah sangat menentukan dalam proses kelayakan peminjaman uang ke bank, jenis properti yaitu jaminan calon nasabah kepada bank jika calon nasabah tidak mampu membayar maka properti tersebut akan disita oleh bank.

Ilmu dalam bidang komputer yang terbukti mampu memecahkan sejumlah masalah dan salah satu teknologi yang dapat membantu adalah data mining sebagai referensi data mining dalam pengambilan keputusan, *clustering*, dan *forecasting* serta dapat menggunakan algoritme *C4.5, k-means, dan k-nearest neighbor* serta menggunakan teknik dan algoritme aturan asosiasi [3]. Algoritme K-NN merupakan salah satu algoritme populer serta memiliki kelebihan dan keuntungan yaitu memiliki nilai akurasi yang tinggi, dapat diproses dengan mudah dan sederhana, kelemahan sensitif terhadap data outliers [4]. Naïve Bayes adalah suatu algoritme klasifikasi yang baik dalam tingkat akurasi tetapi memiliki kelemahan dalam proses penyeleksian atribut [4]. Berdasarkan penelitian terdahulu Penerapan Algoritme K-Nearest Neighbor Untuk Penentuan Resiko Kredit Kepemilikan Kendaraan Bemotor dengan akurasi 81.46% dan sebagai algoritme klasifikasi yang baik dan memiliki nilai AUC antara 0.90-1.00 [5]. Kemudian topik penelitian tentang Analisis Kelayakan Kredit Berbasis Algoritme K-Nearst Neighboar (Studi Kasus: Koperasi AKU) akurasi mencapai 79,45% [6]. Selanjutnya penelitian Perbandingan Metode Klasifikasi Naïve Bayes Dan K-Nearest Neighbor Pada Analisis Data Status Kerja Di Kabupaten Demak Tahun 2012 Pada hasil penelitian tersebut menunjukkan bahwa rata-rata akurasi KNN mencapai 96% dan Naïve Bayes mencapai 94% [7].

Berdasarkan permasalahan yang telah dijelaskan maka penelitian ini mengimplementasikan algoritme *K-Nearest Neighbor* dan *Naïve Bayes*, untuk klasifikasi kelayakan pemberian kredit perbankan pada KPR dari dataset kaggle dan untuk mengetahui tingkat akurasi terbaik pada algoritme *K-Nearest Neighbor* dan *Naïve Bayes*, sehingga diharapkan dengan adanya model klasifikasi ini dapat membantu perusahaan menganalisis kemungkinan pemberian pinjaman kepada nasabah untuk mencegah terjadinya kredit macet oleh nasabah.

2. Tinjauan Pustaka

Penelitian Nugraha Listiana Hanun dengan judul “Penerapan Algoritme Klasifikasi Random Forest Untuk Penentuan Kelayakan Pemberian Kredit di Koperasi Mitra Sejahtera. Dari hasil pengujian dengan algoritme klasifikasi *Random Forest* mampu menganalisis kredit yang bermasalah dan yang debitur yang tidak bermasalah dengan nilai akurasi sebesar

87,88%. Selanjutnya model pohon keputusan ternyata mampu meningkatkan akurasi dalam menganalisis kelayakan kredit yang diajukan calon debitur.

Siti Masripah dengan judul Komparasi Algoritme Klasifikasi Data Mining untuk Evaluasi Pemberian Kredit. Hasil yang didapatkan dari perbandingan kedua algoritme C 4.5 dan Naïve Bayes bahwa tingkat akurasi yang lebih baik adalah menganalisa menggunakan algoritme klasifikasi C4.5 yaitu 88.90 % sedangkan untuk tingkat akurasi menggunakan algortima klasifikasi *Naïve Bayes* yaitu 80.00%.

Putri Kurnia Handayani Model Klasifikasi Kelayakan Kredit Koperasi Karyawan Dengan Algoritme Decision Tree. Hasil penelitian menunjukkan akurasi dari algoritme Decision Tree sebesar 92,28% untuk memodelkan kelayakan kredit sebuah koperasi karyawan.

Penelitian Amat Damuri dengan Implementasi Data Mining dengan Algoritme Naïve Bayes Untuk Klasifikasi Kelayakan Penerima Bantuan Sembako. Hasil yang didapatkan dari penelitian menggunakan Algoritme Naïve Bayes, evaluasi menggunakan confusion matrix didapatkan akurasi yang dihasilkan dari 135 data training dengan 40 data testing dan tujuh atribut yang digunakan menghasilkan akurasi sebesar 86%, recall 85%, dan presisi 88%. Akurasi dapat dipengaruhi oleh beberapa faktor, diantaranya: jumlah data training, data testing dan atribut yang digunakan.

Penelitian Wildan Muhollad Habibulloh dengan judul Klasifikasi Kelayakan Kredit Menggunakan Algoritme Naïve Bayes Pada KSP Mekar Jaya Maleber. Dari hasil penelitian ini menggunakan algoritme Naïve Bayes dan menghasilkan nilai akurasi sebesar 76,76% dengan nilai AUC sebesar 0,824 yang berarti merupakan good classification.

Penelitian Diah Nurul Chasanah dengan judul Klasifikasi Kelayakan Siswa dalam Menentukan Kelas Unggulan Menggunakan Algoritme K-Nearest Neighbor. Dari hasil akurasi tertinggi dengan algoritme *K-Nearest Neighbor* untuk mengklasifikasi kelayakan siswa dalam menentukan kelas unggulan terdapat pada nilai K=25. Hasil perhitungan manual terdapat pada kategori Layak, sedangkan hasil perhitungan dengan RapidMiner pada nilai K=25 diperoleh akurasi sebesar 95.05%, dan hasil dari google colaboratory pada nilai K=25 dengan nilai akurasi sebesar 97,00%.

Penelitian Yusfina Susanti Ripka Igo dengan judul Klasifikasi Kelayakan Pemberian Kredit Nasabah Bank Xyz Menggunakan Metode Algoritme C4.5 Dan Naive Bayes". Tujuan dari penelitian ini adalah untuk menentukan metode terbaik dalam klasifikasi kelayakan pemberian kredit kepada nasabah selanjutnya pemilihan metode terbaik yang cocok untuk klasifikasi pemberian kredit dengan akurasi tertinggi. Akurasi terbaik diperoleh dengan membandingkan algoritme c4.5 dan naive Bayes. Hasil akurasi yang diperoleh algoritme C4.5 pada tiga pengujian berturut-turut adalah 65,75%, 67,70%, 64,95%, sedangkan Naive Bayes memberikan akurasi sebesar 64,72%, 66,67% dan 63,40%.

Berdasarkan beberapa penelitian sebelumnya yang telah dijelaskan, terdapat kesamaan dalam klasifikasi kelayakan pemberian kredit perbankan. Penelitian ini bertujuan untuk menentukan klasifikasi kelayakan pemberian kredit perbankan pada KPR dan untuk mengetahui tingkat akurasi terbaik pada algoritme *K-Nearest Neighbor* dan *Naïve Bayes* melalui beberapa tahap pengujian.

3. Metodologi

Metodologi yang digunakan dalam penelitian ini yaitu CRISP-DM merupakan standar untuk proses analitik perusahaan atau industri sebagai strategi pemecahan masalah dari departemen industri [10]. CRISP-DM merupakan sebuah metode standar untuk mengembangkan proyek data mining karena banyak yang menggunakan dalam pengembangan data mining serta mempunyai gambaran beberapa siklus hidup [14].

1. Pemahaman Bisnis (*Business Understanding*)

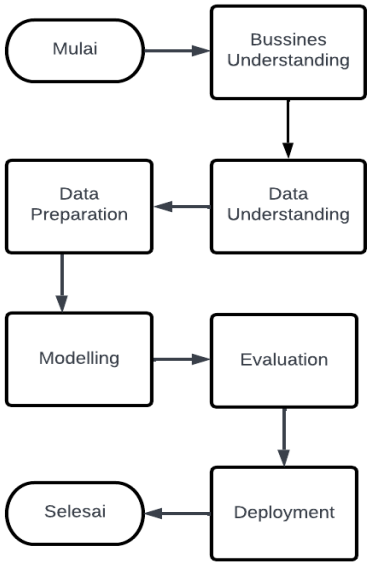
Pada tahap pemahaman bisnis, akan mencari orientasi awal, tujuan dan strategi awal dalam penelitian ini, dengan menentukan masalah bisnis dari penelitian klasifikasi kelayakan pemberian kredit perbankan ini adalah menerapkan algoritme K-Nearest Neighbor dan Naïve Bayes untuk klasifikasi kelayakan pemberian kredit perbankan dan mengetahui tingkat akurasi yang dihasilkan oleh algoritme *K-Nearest Neighbor* dan *Naïve Bayes* dalam klasifikasi kelayakan pemberian kredit perbankan.

2. Pemahaman Data (Data Understanding)

Pada tahap data understanding ini bertujuan untuk menganalisis data yang sudah dikumpulkan. Sumber data pada penelitian ini diambil dari website Kaggle didalam data ini terdapat 491 data training 123 data testing dan 10 atribut yang digunakan.

Bank Mandiri merupakan sebuah badan usaha milik negara yang memberikan pinjaman uang kepada masyarakat bersifat produktif dalam menumbuhkan pertumbuhan ekonomi rakyat untuk produksi dan sekaligus dalam membuka lowongan tenaga kerja atau mengurangi pengangguran. Variabel kelayakan dalam proses penentuan calon nasabah dalam pemberian kredit yaitu jenis usaha yang dilakukan minimal usaha sudah berjalan 2 tahun, pendapatan atau omzet rata-rata per bulan dikurangi beban atau pengeluaran keseluruhan, barang agunan atau jaminan yang di agunankan layak sertifikat tanah atau BPKB mobil atau motor, tenor atau jangka waktu pinjaman kredit antara 1 tahun atau 12 – 60 bulan, limit pinjaman kredit minmal 10 juta – 500 juta disesuaikan dengan jenis agunan, riwayat kredit yang sudah dicek IDEB BI atau rekam jejak transaksi keuangan di semua lembaga perbankan maupun non bank koperasi atau CU yg di ambil data keuangan dari bank indonesia pernah kredit macet atau lancar [18].

Selanjutnya di Bank BNI salah satu badan usaha milik negara untuk membiayai pinjaman kepada masyarakat wirausaha dan swasta dengan kriteria pemohon yakni KTP suami atau istri, surat nikah, kartu keluarga, pekerjaan, surat keterangan kerja dan slip gaji, rekening gaji 3 bulan terakhir, rekening koran 6 bulan terakhir, pinjaman, jangka waktu, legalitas usaha, kualitas kredit bank, jaminan, pengalaman usaha, NPWP [19].



Gambar 1 Metode CRISP-DM

Selanjutnya dataset tersebut diupload pada tahun 2018 berjudul Loan Prediction, dataset berupa data training, testing dan 10 atribut yang digunakan serta kumpulan data cocok untuk dipakai dalam penelitian ini karena atribut tersebut dalam dataset ini lengkap sehingga memudahkan dalam klasifikasi kelayakan pemberian kredit perbankan pada KPR.

3. Pengolahan Data (*Data Preparation*)

Pada tahap pengolahan data memiliki tujuan untuk memperoleh data yang bagus sehingga dapat menghasilkan hasil yang terbaik. Oleh karena itu, bagaimana mempersiapkan data untuk langkah selanjutnya, yaitu data preprocessing. Preprocessing data dalam penelitian ini adalah membersihkan data dengan mengecek missing value, mengecek drop duplicate, handling outlier, menangani imbalance dan standarisasi data yang terdapat dalam dataset Loan Prediction.

Tabel 1 Sample Dataset

Gender	Married	Dependents	Self Employed	Applicant Income	Loan Amount	Loan Amount Term	Credit History	Property Area	Loan Status
Male	Yes	2	No	3073	200	360	1	Urban	Yes
Male	Yes	0	No	2583	120	360	1	Urban	Yes
Male	No	0	No	6000	141	360	1	Urban	Yes
Male	Yes	3+	No	3036	158	360	1	Semi Urban	No
Male	No	0	No	1853	114	360	1	Rural	No
..	..	..	..	..	..	..	..	..	..

4. Pemodelan (*Modeling*)

Pada tahap pemodelan, penelitian juga melibatkan salah satu dari beberapa teknik data mining dan teknik preprocessing. Teknik data mining yang digunakan adalah melakukan proses algoritme K-Nearest Neighbors (K-NN) dan Naïve Bayes. Perhitungan KNN dituliskan dengan persamaan 1:

$$d_{ij} = \sqrt{\sum_{i=1}^{p} (X_{ik} - X_{jk})^2} \quad (1)$$

Keterangan:
X_{ik} = Sample Data / Data Training
X_{jk} = Data Uji/ Testing
Ij = Variabel Data
d = Jarak
p = Dimensi Data

Perhitungan Naïve Bayes dituliskan dengan persamaan dibawah ini (2):

$$P\ (H|X) = \frac{P(X|H)P(H)}{P(X)} \quad (2)$$

Keterangan:
X = data dengan *class* yang belum diketahui (feature)
H = hipotesis data X merupakan suatu *class* spesifik (kelas feature)
P(H|X) = probabilitas hipotesis H berdasarkan kondisi X (*posteriori probability*)
P(H) = probabilitas hipotesis H kelas target (*prior probability*)
P(X|H) = probabilitas X berdasarkan kondisi hipotesis H
P(X) = probabilitas dari total data X

5. Evaluasi (*Evaluation*)

Pada tahap ini dilakukan mengevaluasi model yang diimplementasikan membahas *Business Understanding* agar tidak ada proses yang tertinggal. Fase ini menggunakan data uji (testing data) sebagai proses pengujian dan mengevaluasi model dengan menggunakan *confusion matrix*. Selanjutnya jika model sudah sesuai dengan tujuan Business Understanding yang dijelaskan di atas, maka tahap selanjutnya adalah *deployment*.

6. Penyebaran (*Deployment*)

Dalam tahapan ini, deployment dapat dilakukan dengan cara menyebarkan model sehingga bisa dibaca dan diintegrasikan [18]. Dalam penelitian ini, tahap deployment yaitu mengimplementasikan model yang telah dibangun kedalam sebuah website. Tahapannya model akan disimpan menggunakan modul bernama Pickle, untuk melakukan integrasi antara

model dan aplikasi berbasis web sebagai interface unuk model tersebut serta penulis akan mengintegrasikan model yang telah disimpan kedalam bentuk format file .pkl menggunakan Pickle, dan mengintegrasikannya kedalam aplikasi berbasis web menggunakan flask sebagai *micro web framework*.

4. Hasil dan Pembahasan

Pada penelitian ini hasil dari implementasi sistem klasifikasi menggunakan metode K-Nearest Neighbor (K-NN) Dan Naïve Bayes pengujian ini menggunakan tingkat akurasi dan evaluasi menggunakan confusion matrix.

4.1 Hasil Preprocessing

Tahapan *preprocessing* merupakan langkah setelah tahapan EDA, *preprocessing* adalah proses untuk menyiapkan data sehingga bisa dilatih oleh model, tahap dari hasil preporcessing dapat dilihat pada sub bab berikut ini.

1) Label Encode

Label Encode merupakan suatu proses dari *machine learning* data yang awalnya dari numerik menjadi kata kategorik sehingga dapat diproses dalam data mining sehingga *Label Encode* menjadi suatu proses langkah yang sangat penting. Jika pada tahapan *Label encode* telah dilakukan maka tahap selanjutnya bisa ke modeling dan evaluasi model.

```
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   Gender            599 non-null    object
 1   Married           611 non-null    object
 2   Dependents        599 non-null    object
 3   Self_Employed     582 non-null    object
 4   ApplicantIncome   612 non-null    float64
 5   LoanAmount        592 non-null    float64
 6   Loan_Amount_Term  600 non-null    float64
 7   Credit_History    564 non-null    float64
 8   property_Area     614 non-null    object
 9   Loan_Status       614 non-null    object
```

Gambar 2 Label Encode

Diatas adalah gambar tipe data yang harus dilakukan label encode sehingga bisa diproses oleh *machine learning*, dalam penelitian ini penulis melakukan label encode pada tipe data object.

```
Data columns (total 10 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   Gender            614 non-null    float64
 1   Married           614 non-null    float64
 2   Dependents        614 non-null    float64
 3   Self_Employed     614 non-null    float64
 4   ApplicantIncome   614 non-null    float64
 5   LoanAmount        614 non-null    float64
 6   Loan_Amount_Term  614 non-null    float64
 7   Credit_History    614 non-null    float64
 8   property_Area     614 non-null    int32
 9   Loan_Status       614 non-null    int32
```

Gambar 3 Setelah Label *Encode*

Gambar 3 merupakan tipe data yang sudah dilakukan label encode sehingga bisa diproses oleh *machine learning*.

2. Missing Value

Missing Value merupakan hilangnya beberapa data yang telah diperoleh dan saling berkaitan dalam proses pembersihan sebelum nantinya akan dilakukan proses analisis dan prediksi data sehingga data tersebut tehindar dari data yang kotor dan siap diolah.

Gender	15
Married	3
Dependents	15
Self_Employed	32
ApplicantIncome	2
LoanAmount	22
Loan_Amount_Term	14
Credit_History	50
property_Area	0
Loan_Status	0

Gambar 4. Missing *Value*

Gambar diatas adalah daftar data yang terdapat *Missing value*, setelah melakukan analisis terdapat *Missing value* pada data *Gender, Married, Dependents, Self employed, AplicantIncome, Loan Amount, Loan Amout Term,* dan *Credit History* kemudian dilakukan penanganan pada penelitian ini merubah data yang hilang pada data *Dependents, Self employed, Credit History* menjadi data *modus* karena mengisi data dengan hasil yang sering muncul, dan *AplicantIncome, Loan Amount, Loan Amout Term* menjadi data *median* agar data tidak terpengaruh dengan nilai outlier, sedangkan *Gender, Married* menjadi data modus agar tidak mempengaruhi hasil model.

Gender	0
Married	0
Dependents	0
Self_Employed	0
ApplicantIncome	0
LoanAmount	0
Loan_Amount_Term	0
Credit_History	0
property_Area	0
Loan_Status	0

Gambar 5 Menangani missing value

3. Data Duplicate

Data Duplicate seperti namanya yaitu data yang mirip atau set data mungkin terdiri dari objek data yang ganda (*duplikat*), dalam mengolah hampir selalu terjadi duplikasi antara satu data dengan yang lainnya.

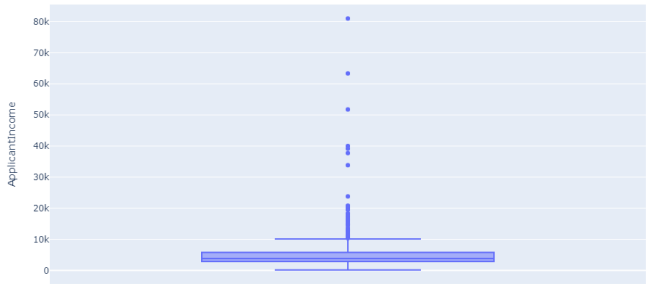
Gender	Married	Dependents	Self_Employed	ApplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	property_Area	Loan_Status

Gambar 6 Data I

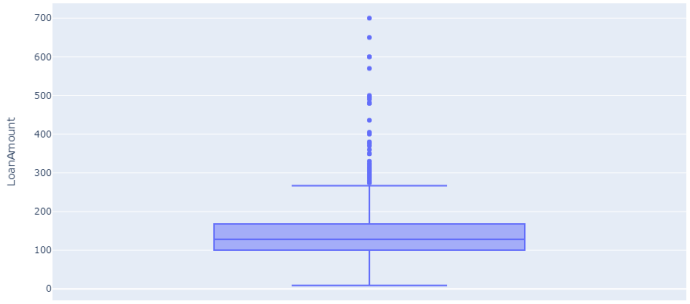
Gambar diatas adalah data yang tidak memiliki *Data Duplicate*, jika terdapat *data duplicate* maka harus menangani dengan menghapus data yang duplikat dengan cara *drop duplicate*.

4. Outlier

Outlier Merupakan suatu nilai yang signifikan yang frekuensinya tiba tiba tinggi dari yang lain sehingga dapat mempengaruhi hasil dari sebuah model machine learning, biasanya untuk menentukan suatu titik objek yang memiliki outlier merupakan khas individual tersendiri dalam melakukan proses studi pada data.

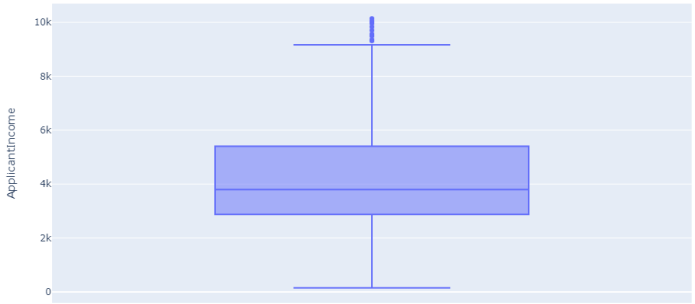


Gambar 7. Cek Outlier Aplicant Income



Gambar 8 Cek Outlier Loan Amount

Gambar 8 merupakan proses pengecekan pada outlier di data Aplicant Income dan Loan Amount menggunakan *box plot*, dari data tersebut bahwa banyak data *outlier* di fitur Aplicant Income dan Loan Amount sehingga penulis melakukan penanganan pada Aplicant Income dengan merubah data *outlier* menjadi nilai mean dari batas atas.



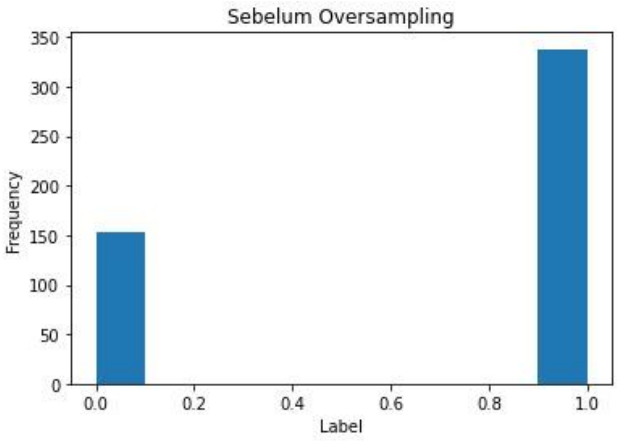
Gambar 9 Mengatasi *Outlier Aplicant Income*

Gambar 9 merupakan hasil dari penanganan menggunakan nilai mean dari data atas pada feature Aplicant Income dan untuk Loan Amount penulis tidak merubah apapun agar tgidak merusak data.

5. Data *Imbalance*

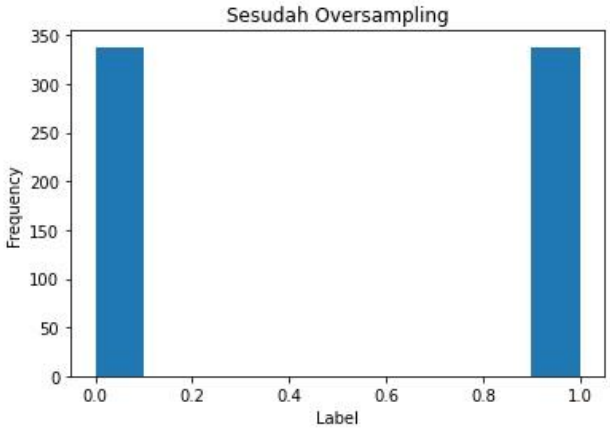
Kelas yang tidak seimbang merupakan masalah umum dalam klasifikasi pembelajaran mesin di mana terdapat rasio yang tidak proporsional di setiap kelas. Ketidakseimbangan kelas

dapat ditemukan di berbagai bidang termasuk klasifikasi kelayakan pemberian kredit perbankan.



Gambar 10 Sebelum mengatasi imbalance

Diatas merupakan plot label yang berisi tentang data layak atau tidak dari data diatas bisa dilihat bahwa ada ketidakseimbangan data antara Label 0 dan 1 yang mana lebih banyak data yang layak daripada data yang tidak layak.



Gambar 11 Sesudah mengatasi imbalance

Diatas adalah plot persebaran data label sesudah dilakukan penanganan data imbalance dan penelitian ini menggunakan metode SMOTE yang termasuk dalam teknik oversampling yaitu memperbanyak data minoritas sehingga data sintetis yang baru berdekatan dengan data asli.

6. Standarisasi

Standarisasi menggunakan standar scaler merupakan suatu proses merubah nilai asal data suatu variabel agar mudah dimengerti dan dibandingkan dengan data dari variabel lain berdasarkan mean dan standar deviasi dan proses ini dilakukan untuk mencegah adanya data yang memiliki nilai terlalu besar dibanding dengan nilai yang lain yang akan dapat mengakibatkan proses training tidak sesuai keinginan.

```
array([[ 0.48816023,  0.77559662,  1.33471037, ...,  0.30501141,
        -2.01933146,  0.08844545],
       [-2.21165568, -1.42569569, -0.77914328, ...,  0.30501141,
         0.57839502,  1.36055451],
       [-2.21165568,  0.77559662, -0.77914328, ...,  0.30501141,
         0.57839502,  1.36055451],
       ...,
       [ 0.48816023,  0.77559662, -0.09646056, ...,  0.30501141,
         0.57839502,  1.36055451],
       [ 0.48816023, -1.42569569, -0.77914328, ..., -1.98183304,
         0.04634477,  0.08844545],
       [ 0.48816023,  0.77559662,  0.76221063, ...,  0.30501141,
        -1.31578335,  1.36055451]])
```

Gambar 12 Hasil Standar *Scaler*

Diatas adalah hasil dari mengskalakan menggunakan standar scaler nilai variabel di kurangi mean dari nilai variabel dan dibagi standar deviasi sehingga didapati hasil dari perhitungan diatas, pada penelitian ini proses melakukan standar scaler yaitu pada pengujian ketiga untuk melihat akurasi terbaik.

4.2 Pengujian Pertama

Pada pengujian pertama Naïve Bayes melakukan preprocessing pada dataset yaitu terdiri dari mengisi missing value, mengecek data duplikat, melakukan label encoding dan mengatasi outlier.

```
Training Accuracy is:- 81.06 %
================================================
Testing Accuracy is:- 80.49 %
================================================
Testing Precision is:- 78.3 %
Testing Recall is:- 98.81 %
```

Gambar 13 Naïve Bayes 1

Pada gambar pengujian pertama akurasi algoritme *Naïve Bayes* cukup baik, data training sebesar 81.06% dan data testing sebesar 80.49%.

Tabel 2 *Confusion Matrix* NB 1

	Label Prediksi		
Label	**0**	**16**	**23**
Aktual	**1**	**1**	**83**
		0	**1**

Selanjutnya Pada pengujian pertama K-Nearest Neighbor dan menggunakan K = 10 melakukan preprocessing pada dataset yaitu terdiri dari mengisi missing value, mengecek data duplikat, melakukan label *encoding* dan mengatasi *outlier*.

Pada gambar 14 pengujian pertama akurasi algoritme KNN cukup baik, data training sebesar 68.02 % dan data testing sebesar 65.85 %.

```
Training Accuracy is:- 68.02 %
================================================
Testing Accuracy is:- 65.85 %
================================================
Testing Precision is:- 71.0 %
Testing Recall is:- 84.52 %
```

Gambar 14 KNN 1

Tabel 3 Confusion Matrix KNN 1

		Label Prediksi	
Label Aktual	**0**	**10**	**29**
	1	**13**	**71**
		0	**1**

4.3 Pengujian Kedua

Pada pengujian kedua Naïve Bayes melakukan praprosesing pada dataset yaitu terdiri dari mengisi missing value, mengecek data duplikat, melakukan label encoding, mengatasi outlier dan mengatasi imbalance pada dataset menggunakan metode SMOTE.

```
Training Accuracy is:- 75.15 %
==================================================
Testing Accuracy is:- 80.49 %
==================================================
Testing Precision is:- 78.3 %
Testing Recall is:- 98.81 %
```

Gambar 15 Naïve Bayes 2

Pada gambar pengujian kedua akurasi pada algoritme *Naïve Bayes* cukup baik, data training sebesar 75.15 % dan data testing sebesar 80.49%.

Tabel 4 *Confusion Matrix* NB 2

		Label Prediksi	
Label Aktual	**0**	**16**	**23**
	1	**1**	**83**
		0	**1**

Selanjutnya Pada pengujian kedua *K-Nearest Neighbor* dan menggunakan K = 10 melakukan *preprocessing* pada dataset yaitu terdiri dari mengisi missing value, mengecek data duplikat, melakukan label encoding, mengatasi outlier dan mengatasi *imbalance* pada dataset menggunakan metode SMOTE.

```
Training Accuracy is:- 67.01 %
==================================================
Testing Accuracy is:- 44.72 %
==================================================
Testing Precision is:- 64.81 %
Testing Recall is:- 41.67 %
```

Gambar 16 KNN 2

Pada gambar pengujian kedua akurasi algoritme KNN kurang baik dikarenakan terjadinya perbedaan akurasi secara signifikan, data training sebesar 67.01 % dan data testing sebesar 44.72 %.

Tabel 5 *Confusion Matrix* KNN 2

		Label Prediksi	
Label Aktual	**0**	**20**	**19**
	1	**49**	**35**
		0	**1**

4.4 Pengujian Ketiga

Pada pengujian ketiga *Naïve Bayes* melakukan preprocessing pada dataset yaitu terdiri dari mengisi missing value, mengecek data duplikat, melakukan label *encoding*, mengatasi outlier dan mengatasi imbalance pada label menggunakan metode SMOTE serta melakukan standarisasi menggunakan standar scaler.

```
Training Accuracy is:- 58.28 %
==================================================
Testing Accuracy is:- 45.53 %
==================================================
Testing Precision is:- 77.42 %
Testing Recall is:- 28.57 %
```

Gambar 17 Naïve Bayes 3

Pada gambar pengujian ketiga akurasi pada algoritme *Naïve Bayes* kurang baik dikarenakan terjadinya perbedaan akurasi secara signifikan, data training sebesar 58.28 % dan data testing sebesar 45.53%.

Tabel 6 Confusion Matrix NB 3

	Label Prediksi		
Label Aktual	**0**	**32**	**7**
	1	**60**	**24**
		0	**1**

Selanjutnya Pada pengujian ketiga K-Nearest Neighbor dan menggunakan K = 10 melakukan preprocessing pada dataset yaitu terdiri dari mengisi missing value, mengecek data duplikat, melakukan label encoding, mengatasi outlier, mengatasi imbalance pada label menggunakan metode SMOTE serta melakukan standarisasi menggunakan standar scaler.

```
Training Accuracy is:- 80.92 %
==================================================
Testing Accuracy is:- 78.86 %
==================================================
Testing Precision is:- 80.85 %
Testing Recall is:- 90.48 %
```

Gambar 18 KNN 3

Pada gambar pengujian ketiga pada akurasi pada algoritme KNN cukup baik, data training sebesar 80.92 % dan data testing sebesar 78.86 %.

Tabel 7 *Confusion Matrix* KNN 3

	Label Prediksi		
Label Aktual	**0**	**21**	**18**
	1	**8**	**76**
		0	**1**

4.5 Pembahasan

Hasil Pengujian tingkat akurasi dapat dilihat perbandingan akurasi dari setiap pengujian dan di pengujian pertama penelitian menggunakan algoritme Naïve Bayes mendapatkan performa akurasi data training 81.06% dan data testing 80.49% dari pengamatan tabel confusion matrix score TP 83 dan TN 16 baik dalam memprediksi kelas 1 tetapi kurang baik dalam memprediksi kelas 0, selanjutnya algoritme KNN K=10 mendapatkan performa akurasi data training 68.02% dan data testing 65.85% serta mengalami penurunan akurasi dari hasil Naïve Bayes dari pengamatan tabel confusion matrix score TP 71 dan TN 10 baik dalam memprediksi kelas 1 tetapi kurang baik dalam memprediksi kelas 0. Pengujian kedua setelah penelitian melakukan penyeimbangan data menggunakan metode SMOTE yaitu memperbanyak data minoritas sehingga data sintetis yang baru berdekatan dengan data asli dan peneliti menggunakan algoritme Naïve Bayes mendapatkan performa akurasi data training 75.15% dan data testing 80.49% dari pengamatan tabel confusion matrix score TP 83 dan TN 16 baik dalam memprediksi kelas 1 tetapi kurang baik dalam memprediksi kelas 0, selanjutnya algoritme KNN K=10 mendapatkan performa akurasi data training 67.01% dan data testing 44.72% pada KNN mengalami penurunan akurasi secara signifikan yang dapat menunjukan adanya overfitting dikarekankan model terlalu menyesuaikan diri pada data training dan tidak

dapat mengeneralisasi data testing dengan baik, dari pengamatan tabel confusion matrix score TP 35 dan TN 20 kurang baik dalam memprediksi kelas 1 tetapi cukup baik dalam memprediksi kelas 0. Pengujian ketiga penelitian melakukan penyeimbangan data menggunakan metode SMOTE yaitu memperbanyak data minoritas sehingga data sintetis yang baru berdekatan dengan data asli. Penelitian ini mencoba melakukan standarisasi menggunakan standar scaler dikarenakan untuk mengubah variabel ke skala yang sama untuk pembelajaran mesin sehingga dapat meningkatkan kinerja algoritme dan peneliti menggunakan algoritme Naïve Bayes mendapatkan performa akurasi data training 58.28% dan data testing 45.53% pada Naïve Bayes mengalami penurunan akurasi secara signifikan yang dapat menunjukan adanya overfitting dikarenakan model terlalu menyesuaikan diri pada data training dan tidak dapat mengeneralisasi data testing dengan baik dari pengamatan tabel confusion matrix score TP 24 dan TN 32 kurang baik dalam memprediksi kelas 1 tetapi cukup baik dalam memprediksi kelas 0, selanjutnya algoritme KNN K=10 mendapatkan performa akurasi data training 80.92% dan data testing 78.86%, dari pengamatan tabel *confusion matrix score* TP 76 dan TN 21 baik dalam memprediksi kelas 1 dan cukup baik dalam memprediksi kelas 0, selanjutnya penelitian ini mengambil pengujian ke 3 dari KNN sebagai akurasi terbaik.

5. Simpulan

Hasil dari penelitian yang dilakukan melalui beberapa tahapan mulai dari menangani *missing value*, menangani *outliers*, mengecek *duplicate* melakukan label *encoding*, standarisasi data menggunakan standar *scaler*, serta menangani data imbalance menggunakan metode SMOTE. Pengujian Pertama dilakukan beberapa tahapan *preprocessing* mulai dari pengecekan *duplicate*, menangani *missing value*, menangani *outliers*, melakukan label *encoding*. Pengujian Kedua dilakukan beberapa tahapan *preprocessing* mulai dari pengecekan *duplicate*, menangani *missing value*, menangani *outliers*, melakukan label *encoding* dan mengatasi data imbalance menggunakan metode SMOTE. Pengujian Ketiga dilakukan beberapa tahapan *preprocessing* mulai dari pengecekan *duplicate*, menangani *missing value*, menangani *outliers*, melakukan label *encoding*, mengatasi data *imbalance* menggunakan metode SMOTE, dan melakukan standarisasi menggunakan standar *scaler*.

Hasil dari pengujian algoritme *Naïve Bayes* dan KNN serta tahapan model dievaluasi untuk memeriksa pengaruh setiap tahap pada data terhadap kemampuan model dalam memprediksi, metrik evaluasi yang digunakan berupa hasil *confusion matrix*. Terdapat hasil terbaik yaitu pada algoritme KNN di pengujian ketiga dengan nilai K=10 dengan performa akurasi data training 80.92% dan data testing 78.86% dan mendapatkan *score confusion matrix* TP 76 dan TN 21.

Daftar Referensi

[1] D. A. Kurniawan and D. Kriestanto, “Penerapan Naive Bayes Untuk Prediksi Kelayakan Kredit,” *JIKO (Jurnal Inform. dan Komputer)*, vol. 1, no. 1, pp. 19–23, 2016, doi: 10.26798/jiko.2016.v1i1.10.

[2] S. Kosasi, “Aplikasi Pemberian Kredit Pada Bank Kalbar Pemangkat Menggunakan Metode Case Based Reasoning,” *Semin. Lokal Inf.*, pp. 30–37, 2013.

[3] F. Hadi, “Penerapan Metode Algoritme C4.5 dalam Menganalisa Pegajuan Kredit pada Koperasi Jasa Keuangan Syariah Kelurahan Limau Manis Selatan,” *Indones. J. Comput. Sci.*, vol. 7, no. 1, pp. 28–42, 2018, doi: 10.33022/ijcs.v7i1.58.

[4] A. U. Budi Santosa, “Data Mining dan Big Data Analytics Teori dan Implementasi menggunakan Python & Apache Spark Edisi 2,” in *Data Mining dan Big Data Analytics Teori dan Implementasi menggunakan Python & Apache Spark Edisi 2*, Penebar Media Pustaka, 2018.

[5] H. Leidiyana, “Penerapan Algoritme K-Nearest Neighbor Untuk Penentuan Resiko Kredit Kepemilikan Kendaraan Bermotor,” *J. Penelit. Ilmu Komputer, Syst. Embed. Log.*, vol. 1, no. 1, pp. 65–76, 2013.

[6] R. Wajhillah, I. H. Ubaidallah, and S. Bahri, “Analisis Kelayakan Kredit Berbasis Algoritme K-Nearest Neighbor (Studio Kasus: Koperasi AKU),” *InfoTekJar (Jurnal Nas. Inform. dan Teknol. Jaringan)*, vol. 4, no. 1, pp. 121–125, 2019, doi: 10.30743/infotekjar.v4i1.1264.

[7] R. E. Putri, Suparti, and R. Rahmawati, “Perbandingan Metode Klasifikasi Naãve Bayes Dan K-Nearest Neighbor Pada Analisis Data Status Kerja Di Kabupaten Demak,” *J.*

Gaussian, vol. 3, no. 4, pp. 831–838, 2014.
[8] L. Hakim, “Prinsip Kehati-hatian pada Lembaga Perbankan dalam Pemberian Kredit,” *J. Keadilan Progresif*, vol. 9, No 2, no. 2, pp. 164–176, 2018.
[9] Jiawei Han;Micheline Kamber;Jian Pei, “Data mining: Data mining concepts and techniques,” *International Conference on Machine Intelligence Research and Advancement, ICMIRA 2013*. p. 740, 2014, doi: 10.1109/ICMIRA.2013.45.
[10] S. Müller, A., & Guido, *Introduction to Machine Learning with Python: A Guide for Data Scientists 1st Edition.* O’Reilly Media, 2018.
[11] F. Ramadhana, F. Fauziah, and W. Winarsih, “Aplikasi Sistem Pakar untuk Mendiagnosa Penyakit ISPA menggunakan Metode Naive Bayes Berbasis Website,” *STRING (Satuan Tulisan Ris. dan Inov. Teknol.*, vol. 4, no. 3, p. 320, 2020, doi: 10.30998/string.v4i3.5441.
[12] A. Géron, *Hands-On Machine Learning with with Scikit-Learn & TensorFlow : Concepts, Tools, and Techniques to Build Intelligent Systems*. O Reilly, 2017.
[13] W. Y. Ayele, “Adapting CRISP-DM for idea mining a data mining process for generating ideas using a textual dataset,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 6, pp. 20–32, 2020, doi: 10.14569/IJACSA.2020.0110603.
[14] F. Martinez-Plumed *et al.*, “CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories,” *IEEE Trans. Knowl. Data Eng.*, vol. 33, no. 8, pp. 1–14, 2021, doi: 10.1109/TKDE.2019.2962680.
[15] R. Rozaq, “Klasifikasi Penyakit Dengue Menggunakan Algoritme K-Nearest Neighbors Berbasis Flask,” *Remik (Riset dan E-Jurnal Manaj. Inform. Komputer)*, vol. 6, no. 3, pp. 359–369, 2022, doi: 10.33395/remik.v6i3.11501.
[16] R. Abdulloh, *Web Programing is Easy*. PT Elex Media Komputindo, 2015.
[17] Provost & Fawcett, *Data Science for Business What You Need to Know About Data Mining and Data-Analytic Thinking*. O’Reilly Media, 2013.
[18] S. Aisyah, Interviewee, *Wawancara Variabel Dominan dalam penentuan kelayakan pemberian kredit.* [Interview]. 07 November 2022.
[19] "BNI Kredit Digital e-Form," PT. Bank Negara Indonesia (Persero) Tbk, [Online]. Available: https://eform.bni.co.id/BNI_eForm/index.html. [Accessed 06 December 2022].