SCIENCE ELECTRO , Vol. 19, No. 1, Sept 2025

Implementasi Metode *Naïve Bayes Classifier* Pada *Machine Learning* Untuk Sistem Alternatif *Credit Scoring*

Ainina Miftahul Jannah[1], Anang Habibi[2], Bambang Minto Basuki[3]
[1]Universitas Islam Malang
[2]Universitas Islam Malang
22101053015@unisma.ac.id, ananghabibi@unisma.ac.id, bambang.minto@unisma.ac.id

Abstract

Conventional credit systems often limit access to credit evaluation, especially for individuals without a formal credit history. This study aims to develop an alternative credit scoring system using the Naïve Bayes Classifier (NBC) method, leveraging socioeconomic variables derived from the National Health Insurance (JKN) participant data in Malang City. The variables used include age, total income, days employed, and monthly total bill. The research process involved data preprocessing, normalization, scoring on a scale of 1–5, and implementation of the NBC model. The model was evaluated using accuracy, precision, recall, and F1-score metrics. Experimental results showed that the Naïve Bayes model could classify credit eligibility with an accuracy of 84%, precision of 85%, recall of 84%, and an F1-score of 84%. This study demonstrates that a machine learning approach using the Naïve Bayes method has the potential to serve as an alternative system for credit eligibility assessment, particularly for individuals underserved by conventional credit systems. The proposed system may also be adapted by socially inclusive financial institutions as a decision-support tool.

Keywords— *Machine Learning*, *Naïve Bayes*, *Credit Scoring*, Financial Instituion, JKN.

Abstraksi

Permasalahan dalam sistem kredit konvensional sering kali terletak pada keterbatasan akses masyarakat terhadap penilaian kelayakan kredit, khususnya bagi individu yang tidak memiliki riwayat kredit formal. Penelitian ini bertujuan untuk mengembangkan sistem alternatif credit scoring menggunakan metode Naïve Bayes Classifier (NBC) yang mampu mengevaluasi kelayakan kredit berdasarkan variabel sosial ekonomi dari data peserta Jaminan Kesehatan Nasional (JKN) Kota Malang. Variabel yang digunakan meliputi usia, total pendapatan, lama bekerja (days employed), dan total tagihan bulanan. Tahapan penelitian dimulai dengan proses preprocessing data, normalisasi, pembobotan skor berdasarkan skala 1–5, hingga implementasi model NBC. Evaluasi model dilakukan menggunakan metrik akurasi, presisi, recall, dan f1-score. Hasil pengujian menunjukkan bahwa model Naïve Bayes dapat mengklasifikasikan kelayakan kredit dengan akurasi sebesar 84%, presisi 85%, recall 84%, dan f1-score 84%. Penelitian ini menunjukkan bahwa pendekatan machine learning dengan metode Naïve Bayes memiliki potensi untuk dijadikan sebagai sistem alternatif dalam penilaian kelayakan kredit, terutama untuk masyarakat yang tidak terjangkau oleh sistem kredit konvensional. Sistem ini juga dapat diadaptasi oleh lembaga keuangan berbasis inklusi sosial sebagai alat bantu pengambilan keputusan.

Kata Kunci— *Machine Learning*, *Naïve Bayes*, *Credit Scoring*, Lembaga Keuangan, JKN.

I. PENDAHULUAN

Pemberian kredit merupakan praktik umum dan memiliki peran penting dalam operasional lembaga keuangan. Salah satu contoh utama lembaga keuangan yang sering menawarkan kredit adalah bank [1]. Kredit menjadi salah satu sumber pendapatan bagi bank, perusahaan swasta, maupun lembaga lain yang menyediakan layanan pembiayaan kepada konsumen. Oleh karena itu, lembaga keuangan harus mampu mengidentifikasi calon nasabah yang memiliki kemampuan untuk melunasi kredit yang diberikan. Salah satu metode yang digunakan untuk menilai kelayakan nasabah dalam menerima pinjaman adalah *credit scoring* [2].

Saat ini, sistem credit scoring tradisional masih menjadi standar utama dalam menilai kelayakan kredit, baik bagi individu maupun perusahaan. Evaluasi kredit secara konvensional umumnya mempertimbangkan berbagai faktor seperti riwayat kredit, tingkat penghasilan, kepemilikan aset, dan rekam jejak pekerjaan. Metode ini sering kali membatasi akses bagi calon peminjam yang belum memiliki riwayat kredit atau yang baru pertama kali mengajukan pinjaman [3].

Untuk mengatasi kendala tersebut, penelitian ini mengusulkan pengembangan Sistem *Alternative Credit Scoring* sebagai solusi inovatif [4]. Sistem ini mengintegrasikan sumber data non-tradisional, seperti riwayat tagihan serta informasi pribadi dari data Jaminan Kesehatan Nasional [5], guna menyusun profil kredit seseorang secara lebih komprehensif. Keunggulan utama dari pendekatan ini adalah kemampuannya dalam memperluas akses kredit bagi individu yang sebelumnya tidak memenuhi kriteria sistem tradisional. Selain itu, metode alternatif ini juga mampu memberikan penilaian kredit yang lebih objektif dan adil dengan mengurangi potensi bias manusia.

Dalam penelitian ini, analisis data dilakukan menggunakan metode Machine Learning Naïve Bayes Classifier, yang efektif dalam mengidentifikasi pola

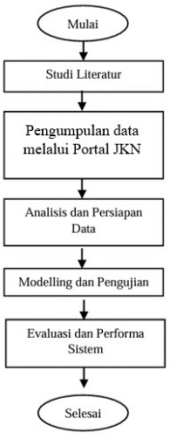
dan hubungan kompleks dalam data. Naïve Bayes Classifier dianggap mampu menyelesaikan permasalahan probabilitas dengan mempertimbangkan kriteria yang relevan bagi penerima kredit [6]. Dengan demikian, penelitian ini bertujuan untuk mengembangkan sistem
yang dapat memberikan penilaian akurat terkait kelayakan konsumen dalam mengajukan kredit berdasarkan data Jaminan Kesehatan Nasional [7].

II. METODE PENELITIAN

Sistem ini dikembangkan melalui penelitian yang dilakukan secara bertahap dan terencana dengan pendekatan kuantitatif, yang berfokus pada pengumpulan serta analisis data dalam bentuk angka atau nilai numerik.

Alternative credit scoring, metode kuantitatif memungkinkan evaluasi berbasis data yang lebih akurat untuk menentukan kelayakan kredit dapat dilakukan secara lebih presisi dengan bantuan teknik statistik dan *machine learning* [8].

A. Flowchart Sistem



Gambar 1. Flowchart Sistem

Pada bagian ini, langkah-langkah penelitian yang akan dilakukan seperti yang ditunjukkan pada **Gambar 1.** Proses klasifikasi dalam penelitian ini bertujuan untuk mengembangkan sistem credit scoring yang dapat membantu lembaga jasa keuangan dalam menentukan kelayakan kredit seseorang dengan menggunakan data sekunder dari Jaminan Kesehatan Nasional (JKN). Setiap tahapan penelitian dalam klasifikasi ini divisualisasikan dalam flowchart yang menunjukkan proses dari awal hingga akhir. Tahapan ini meliputi preprocessing data, pemilihan fitur, pemodelan dengan algoritma machine learning, serta evaluasi hasil untuk memastikan akurasi prediksi kredit

B. Tahap Studi Literatur

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. Pada tahapan ini juga dilakukan pengumpulkan data dari sumber yang sudah disetujui.

Berdasarkan sampel data JKN 2020-2023 yang sudah didapatkan. Dilakukan eksplorasi dan pemeriksaan antara kualitas dan tujuan data dan kemudian didapatkan data yang dipakai untuk menentukan variabel independen yang nantinya dapat digunakan sebagai acuan untuk variabel dependen.

C. Teknik Pengumpulan Data

Data yang digunakan merupakan data sample yang didapatkan langsung melalui permohonan data sample dari website Portal Data Jaminan Kesehatan Nasional pada tahun 2020-2023 di wilayah Kota Malang, Jawa Timur [9]. Jenis pengambilan data yang digunakan adalah web scraping atau data crawling yang berarti proses otomatisasi pengambilan data dari situs web berupa teks dan tabel yang kemudian diolah dan dianalisis lebih lanjut untuk mengekstrasi informasi yang relevan atau sesuai dengan tujuan penelitian. Data yang didapatkan dari permohonan sample berupa 2 file .ZIP yaitu (1) Data Sampel Kontekstual DM-20220328T123037Z-001.zip; (2) Data Sampel Reguler 2020 -2023-20220328T103541Z-001.zip. Masing masing file memiliki 4 datasheet yang berbeda. Tabel 2.1 – 2.3 menampilkan deskripsi atribut-atribut yang ada pada setiap datasheet.

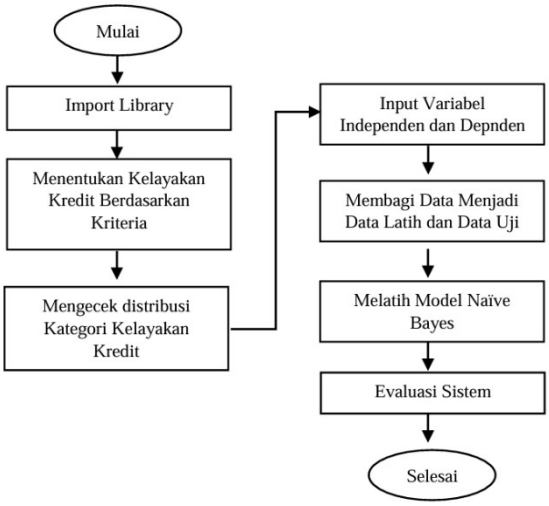
Gambar 2. Portal Data JKN

D. Teknik Analisa Data

Data dari Portal Data JKN diunduh dalam format terstruktur (CSV atau Excel) untuk kemudian diolah lebih lanjut. Data tersebut diintegrasikan ke dalam Python menggunakan library Pandas, dengan penyesuaian encoding ke UTF-8 agar format teks tetap konsisten [10]. Selanjutnya dilakukan pembersihan data dengan memeriksa dan memperbaiki kesalahan pada atribut numerik maupun kategorikal, serta menghapus atribut yang tidak relevan. Proses identifikasi dan penanganan missing value dilakukan dengan fungsi `.isnull()` dan teknik imputasi menggunakan nilai mean, median, atau modus melalui `.fillna()`. Setelah data bersih, dilakukan *Exploratory Data Analysis* (EDA) untuk mengeksplorasi distribusi variabel independen menggunakan visualisasi seperti histogram, boxplot, dan scatter plot [11]. Tahap ini bertujuan untuk menemukan pola dalam data, memahami faktor utama yang memengaruhi credit scoring, serta menentukan algoritma yang tepat untuk sistem yang akan dikembangkan.

Gambar 3. Teknik Analisa Data

E. Tahapan Modelling



Gambar 4. Teknik Modelling

Tahapan modeling merupakan tahap utama dalam pembuatan sistem alternative credit scoring, dimana data yang sudah siap akan dilatih dan diuji dengan sesuai dengan kriteria dan parameter yang sudah dibuat [12].

Dalam proses pembuatan sistem klasifikasi kelayakan kredit, langkah pertama adalah mengimpor library penting seperti pandas untuk pengolahan data, numpy untuk manipulasi numerik, serta train_test_split dan GaussianNB dari scikit-learn untuk membagi data dan membangun model klasifikasi. Kelayakan kredit ditentukan berdasarkan kriteria seperti usia (18–55 tahun), lama bekerja (minimal 180 hari), tagihan bulanan (maksimal 10% dari pendapatan), serta penalti berupa pengurangan skor jika tidak memenuhi kriteria, yang kemudian menghasilkan skor kelayakan kredit dalam rentang 1–5. Selanjutnya dilakukan analisis distribusi data untuk mengetahui apakah dataset seimbang atau tidak, karena distribusi yang tidak merata dapat memengaruhi akurasi model.

Variabel independen (fitur) seperti usia, pendapatan, dan lama bekerja diinputkan, sedangkan variabel dependen adalah skor kelayakan kredit, lalu data dibagi menjadi 80% data latih dan 20% data uji untuk mencegah overfitting dan meningkatkan generalisasi model. Model dilatih menggunakan GaussianNB, yang menghitung probabilitas setiap kelas berdasarkan fitur dengan mengaplikasikan Teorema Bayes, sehingga sistem mampu memprediksi skor kelayakan kredit berdasarkan data masukan secara otomatis dan akurat.

III. Hasil dan Pembahasan

A. Gambaran Umum Datasheet

Penelitian ini menggunakan data sekunder dari Portal Data Jaminan Kesehatan Nasional (JKN) yang dikelola oleh BPJS Kesehatan. Data berupa sampel reguler dan kontekstual JKN tahun 2020–2023 untuk wilayah Kota Malang, Jawa Timur, diperoleh melalui permohonan resmi dan diunduh dalam format ZIP. Dari beberapa datasheet yang tersedia, digunakan datasheet bpjs_dm_01 yang paling relevan dengan tujuan penelitian. Datasheet ini memuat 1.030 data peserta JKN yang mencakup informasi demografis, status keanggotaan, dan kondisi sosial ekonomi.

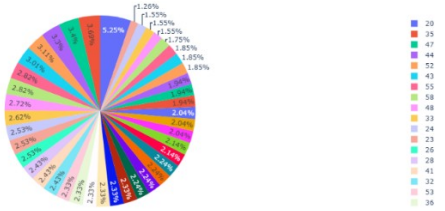
1. Visualisasi Data Jenis Kelamin Peserta

Sebagai bagian dari validasi awal terhadap kualitas dan representasi data, dilakukan analisis terhadap sebaran jenis kelaminpeserta JKN dalam dataset bpjs_dm_01. Tujuan dari validasi ini adalah untuk memastikan bahwa data yang digunakan dalam pemodelan Gambar berikut menunjukkan visualisasi distribusi Jenis Kelamin peserta dalam bentuk diagram pie.

Berdasarkan hasil visualisasi, terlihat bahwa data peserta JKN Kota Malang menunjukkan proporsi yang sangat seimbang antara jenis kelamin perempuan dan laki-laki. Tercatat bahwa 50,9% dari total peserta merupakan perempuan. 49,1% dari total peserta merupakan laki-laki.

2. Visualisasi Data Usia Peserta

Komposisi ini menunjukkan bahwa dataset memiliki distribusi gender yang merata, sehingga analisis dan pemodelan yang dilakukan dalam penelitian ini tidak bias terhadap salah satu kelompok jenis kelamin.



Dari hasil visualisasi dapat dilihat bahwa sebaran usia cukup merata dengan persentase tiap kelompok usia berada di kisaran 1% hingga 5%. Usia dengan proporsi tertinggi adalah usia **20 tahun** (5,25%), diikuti oleh usia **35 tahun** (3,59%) dan **47 tahun** (3,49%). Sebaran usia yang beragam ini memperkuat validitas data sebagai representasi populasi peserta JKN Kota Malang yang relevan untuk dianalisis dalam konteks alternatif credit scoring.

3. Visualisasi Data Status Anggota

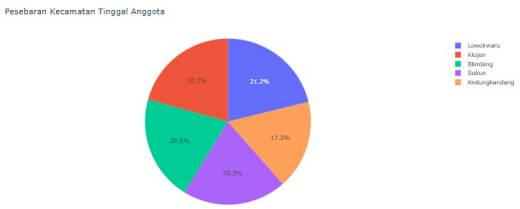
Sebagai bagian dari validasi lebih lanjut terhadap kualitas dataset, dilakukan pula analisis terhadap atribut status pernikahan peserta. Tujuan dari analisis ini adalah untuk mengetahui apakah terdapat keseimbangan dalam distribusi status

peserta

Berdasarkan hasil visualisasi, diketahui bahwa peserta yang sudah menikah memiliki proporsi sebesar 51,1%, sedangkan peserta yang belum menikah sebesar 48,9%. Dengan validasi ini, dapat disimpulkan bahwa dataset memiliki kualitas representasi yang baik dari segi status sosial ekonomi peserta

4. Visualisasi Data Persebaran Kecamatan

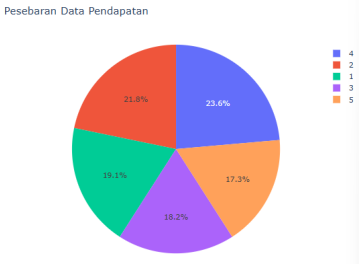
Validasi selanjutnya pada dataset dilakukan terhadap atribut kecamatan tempat tinggal peserta untuk memastikan bahwa data mencakup distribusi geografis yang merata di wilayah Kota Malang



Dari distribusi tersebut, terlihat bahwa empat kecamatan memiliki proporsi yang hampir seimbang, yaitu sekitar 20% masing-masing, sedangkan Kedungkandang sedikit lebih rendah dengan proporsi 17,3%. Sebaran yag relative merata ini menunjukkan bahwa data memiliki representasi geografis yang baik dan tidak terpusat pada satu wilayah tertentu.

5. Visualisasi Data Pendapatan Anggota

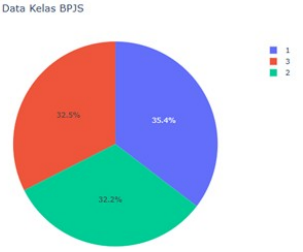
Pada tahapan distribusi data pendapatan dilakukan visualisasi dalam variable untuk menunjukkan data pendapatan dari anggota yang tersebar dalam dataset



Dari persebaran ini dapat disimpulkan bahwa data pendapatan anggota memiliki distribusi yang relatif merata, meskipun terdapat sedikit dominasi pada Kategori 4. Penyebaran yang cukup seimbang ini memberikan gambaran bahwa dataset memiliki representasi yang baik di hampir semua kelas pendapatan.

6. Visualisasi Data Kelas BPJS

Selain data demografis dan ekonomi seperti yang telah dijelaskan diatas variabel kepesertaan dalam program Jaminan Kesehatan Nasional (JKN) juga menjadi bagian penting dari dataset yang digunakan dalam penelitian**.**



Distribusi yang hampir merata antara ketiga kelas ini menunjukkan bahwa dataset memiliki representasi yang seimbang terhadap kelompok masyarakat dengan latar belakang ekonomi yang berbeda

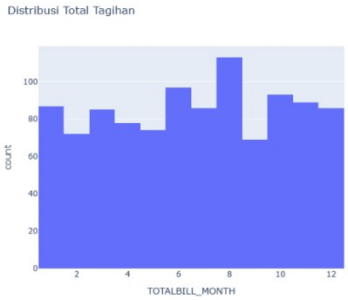
B. Eksplorasi Data (Exploratory Data Analysis)

Tahapan eksplorasi data dilakukan untuk memahami pola, hubungan, serta distribusi dari **masing-masing** variabel dalam dataset sebelum dilakukan pemodelan. Beberapa visualisasi dan analisis statistik digunakan untuk memberikan gambaran awal terhadap karakteristik data. Berikut beberapa hasil eksplorasi data

1. Distribusi Total Tagihan Bulanan

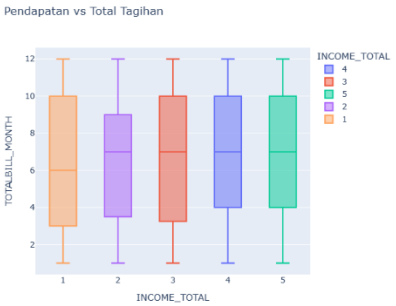
Hasil visualisasi pada Gambar 11 menunjukkan bahwa nilai TOTALBILL_MONTH tersebar dari angka 1 hingga 12. Puncak distribusi terlihat pada nilai 8, yang berarti mayoritas peserta memiliki tagihan bulanan sebesar 8. Sebagian besar data

terkonsentrasi dalam rentang nilai 6 hingga 10, yang menunjukkan bahwa nilai tagihan cenderung berkumpul pada kisaran tersebut.



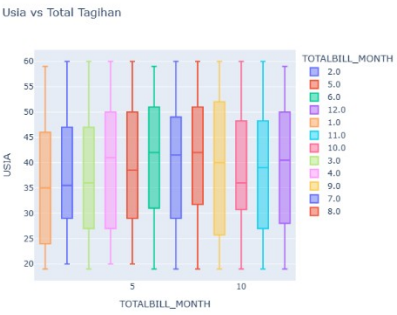
2. Distribusi Tagihan Bulanan Berdasarkan Pendapatan

Hasil visualisasi menunjukkan bahwa distribusi tagihan bulanan relatif seragam di seluruh kategori pendapatan. Median dari masing-masing kategori berada pada kisaran yang hampir sama, menandakan bahwa tidak terdapat perbedaan signifikan pada nilai tengah tagihan bulanan antar kelompok pendapatan. Selain itu, rentang sebaran (interkuartil dan whisker) juga memperlihatkan bahwa variasi tagihan bulanan tidak jauh berbeda, meskipun terdapat sedikit kecenderungan bahwa peserta dengan pendapatan lebih tinggi memiliki sebaran tagihan yang sedikit lebih stabil.



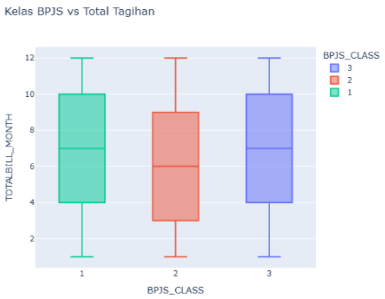
3. Distribusi Usia Berdasarkan Tagihan Bulanan

Hasil visualisasi menunjukkan bahwa distribusi usia cukup merata di hampir semua kategori tagihan bulanan, dengan rentang umum berkisar antara 20 hingga 60 tahun. Median usia pada setiap kelompok tagihan juga relatif seimbang, yang mengindikasikan tidak adanya kecenderungan bahwa usia tertentu lebih dominan pada kategori tagihan tertentu.



4. Distribusi Tagihan Berdasarkan Kelas BPJS

Dari hasil visualisasi terlihat bahwa peserta dengan kelas BPJS 1 dan 3 memiliki sebaran tagihan yang lebih luas dibandingkan dengan kelas 2. Hal ini dapat diinterpretasikan bahwa peserta dari kelas 1 dan 3 memiliki variasi kebutuhan pelayanan kesehatan yang lebih tinggi. Meskipun demikian, median tagihan bulanan pada ketiga kelas berada dalam kisaran nilai yang relatif sama, sehingga tidak ada perbedaan signifikan dari segi nilai tengah.



C. Tahapan Prepocessing

Bertujuan untuk membersihkan, menyesuaikan, serta menyiapkan data agar dapat digunakan secara optimal dalam proses pelatihan dan pengujian algoritma. Pada penelitian ini, proses preprocessing dilakukan terhadap data peserta Jaminan Kesehatan Nasional (JKN) Kota Malang periode 2020–2023, yang sebelumnya telah diperoleh melalui permohonan data dari Portal Data JKN.

Adapun langkah-langkah preprocessing yang dilakukan pada penelitian ini meliputi pembersihan data, seleksi variabel, transformasi data, dan pembuatan variabel target (labeling).

D. Implementasi Model Naïve Bayes dengan Sklearn Gaussian NB

Pada tahap ini, proses implementasi dilakukan menggunakan bahasa pemrograman Python dengan bantuan pustaka scikit-learn yang menyediakan modul GaussianNB, yaitu varian Naïve Bayes yang cocok digunakan untuk data numerik dengan asumsi distribusi normal (Gaussian distribution). Langkah – Langkah implementasi model secara umum sebagai berikut :

1. Pembagian Dataset (Train-Test Split)

   Pembagian ini dilakukan dengan perbandingan 80:20, di mana 80% data digunakan untuk pelatihan, dan 20% sisanya digunakan untuk pengujian [13]. Pembagian ini dilakukan dengan fungsi train_test_split dari pustaka scikit-learn. Dengan pembagian ini, model dapat belajar dari sebagian besar data dan diuji keakuratannya terhadap data baru secara objektif

2. **Training Dataset**

   Setelah dataset dibagi menjadi data latih dan data uji, langkah berikutnya adalah melatih model Naïve Bayes Classifier menggunakan data latih. Pada tahap ini, model mempelajari hubungan antara fitur-fitur input dan label target. Proses ini dikenal sebagai training (pelatihan).

   Dalam penelitian ini digunakan model Gaussian Naïve Bayes, yang mengasumsikan bahwa data numerik pada masing-masing fitur terdistribusi secara normal (distribusi Gaussian).

Pada tahap pelatihan model, objek GaussianNB digunakan untuk mempelajari hubungan antara fitur-fitur numerik dan label target kelayakan kredit. Model menghitung nilai rata-rata dan varians dari setiap fitur pada masing-masing kelas (Layak, Kurang Layak, Tidak Layak), serta menghitung probabilitas awal (prior) tiap kelas. Proses ini dilakukan menggunakan fungsi.fit (x_train, y_train), yang secara internal membentuk model probabilistik berbasis distribusi Gaussian dari data pelatihan.

3. **Prediksi dan Evaluasi Dataset**

Setelah proses pelatihan model selesai dilakukan pada data latih, tahap selanjutnya dalam implementasi sistem klasifikasi adalah melakukan prediksi terhadap data uji. Proses prediksi bertujuan untuk mengukur kemampuan model dalam mengklasifikasikan data baru yang belum pernah dilihat sebelumnya berdasarkan pola yang telah dipelajari.

Dalam penelitian ini, prediksi dilakukan menggunakan metode supervised learning dengan pendekatan Gaussian Naïve Bayes, di mana model yang telah dilatih digunakan untuk mengestimasi kelas kelayakan kredit peserta berdasarkan input fitur numerik seperti usia, penghasilan, lama bekerja, dan total tagihan bulanan. Selain melakukan prediksi terhadap keseluruhan data uji untuk mengevaluasi performa model secara umum

E. Implementasi Model Naïve Bayes Manual dengan Gaussian NB

Model Gaussian Naïve Bayes secara khusus digunakan ketika fitur (variabel input) bersifat numerik dan diasumsikan mengikuti distribusi normal (Gaussian) [14]. Oleh karena itu, probabilitas kemunculan nilai fitur terhadap suatu kelas dihitung menggunakan fungsi distribusi normal.

Model ini menghitung probabilitas posterior dari sebuah kelas (Ck) berdasarkan data fitur x = [x_1, x_2, ..., xn] dengan menggunakan Teorema Bayes [15].

$$P(C = k \mid x) \propto P(C = k) \times \prod_{i=1}^{n} P(x_i \mid C = k)$$

Dengan Keterangan:

- P(Ck) = probabilitas awal (prior) kelas ke-k
- P(xi | Ck) = probabilitas kemunculan nilai fitur ke-i pada kelas ke-k

Untuk fitur numerik, distribusi Gaussian (normal) digunakan untuk menghitung likelihood P(xi | Ck). Rumus distribusi normal adalah:

$$P(x_i \mid C_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \cdot \exp\left(-\frac{(x_i - \mu_k)^2}{2\sigma_k^2}\right)$$

Dengan Keterangan:

- μk adalah rata-rata (mean) fitur xi pada kelas Ck
- σk adalah standar deviasi fitur xi pada kelas Ck

Maka perhitungan Prior Probabilities (P(Ck)) untuk setiap kelas k = 1, 2, 3, 4, 5 dengan rumus seperti diatas akan menghasilkan prior P(C=1) = 0.261, P(C=2) = 0.579, P(C=3) = 0.159, P(C=4) = 0.001.

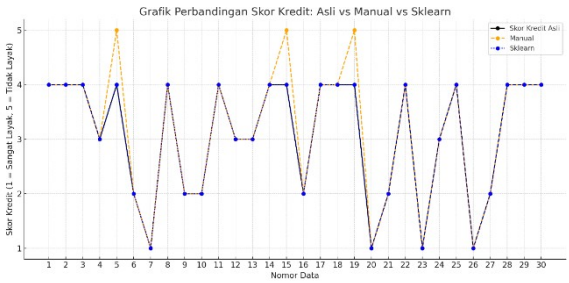
Kemudian Ambil nilai rata-rata (μ) dan standar deviasi (σ) dari masing-masing fitur per kelas misalkan C=2 dengan hasil mean dan standar deviasi pada variabel USIA: μ=39.62, σ=10.56, INCOME_TOTAL: μ=2.59, σ=1.26, DAYS_EMPLOYED: μ=1511.4, σ=1706.53, TOTALBILL_MONTH: μ=5.12, σ=2.77. Maka hasilnya ditampilkan pada table dibawah ini.

Tabel 1. Perhitungan Manual Naïve Bayes

Variable	Nilai
USIA	35
INCOME_TOTAL	4
DAYS_EMPLOYED	3202
TOTALBILL_MONTH	2
P(C=1)	2.19 X 10-19
P(C=2)	3.75 x 10-14
P(C=3)	2.38 x 10-9
P(C=4)	1.28 x 10-8
P(C=5)	-
KELAYAKAN KREDIT SKOR	4

F. Hasil Perbandingan Pengujian Naïve Bayes dengan Scklearn dengan Manual Gaussian NB

Terdapat dua kolom hasil klasifikasi Skor Kredit Manual (Naive Bayes Manual) dan Skor Kredit Sklearn (GaussianNB dari Scikit-Learn). Untuk membuat grafik perbandingan akan divisualisasikan akurasi hasil prediksi dari masing-masing metode (manual dan sklearn) terhadap label asli.



Gambar di atas menampilkan grafik perbandingan hasil klasifikasi skor kredit kelayakan antara metode Naive Bayes manual dan metode Gaussian Naive Bayes menggunakan library Scikit-Learn (sklearn) terhadap data asli. Skor kredit diklasifikasikan dalam lima kategori dengan skala 1 hingga 5, di mana Skor 1 menunjukkan kategori tidak layak. Skor 5 menunjukkan kategori layak untuk mendapatkan kredit.

Pada grafik tersebut, terdapat tiga garis yang merepresentasikan. Skor Kredit Asli (garis hitam): Merupakan label sebenarnya dari masing-masing data berdasarkan kondisi riil. Skor Kredit Manual (garis oranye): Hasil klasifikasi menggunakan perhitungan Naive Bayes secara manual dengan pendekatan Gaussian. Skor Kredit Sklearn (garis biru): Hasil klasifikasi menggunakan library Scikit-Learn

GaussianNB, yang secara otomatis menghitung parameter probabilistiknya.

Dari hasil visualisasi terlihat bahwa akurasi dari prediksi model klasifikasi Naïve Bayes yang diimplementasikan melalui library Scikit-learn berhasil memprediksi 26 data dengan tepat dari total 30 data uji, sementara 4 data sisanya mengalami ketidaksesuaian antara prediksi dan label asli. Dengan demikian, akurasi klasifikasi yang diperoleh sebesar 84%. Namun, dibandingkan dengan model Sklearn GaussianNB, metode manual masih memiliki beberapa kekeliruan dalam pengklasifikasian, yang kemungkinan disebabkan oleh kesalahan pembulatan atau ketidaktepatan pada perhitungan parameter statistik (mean, standar deviasi, atau probabilitas posterior).

IV. KESIMPULAN

Kesimpulan dari penelitian skripsi yang berjudul "Implementasi Metode Naïve Bayes Classifier Pada Machine Learning Untuk Sistem Alternative Credit Scoring" adalah sebagai berikut :

1. Sistem Alternative Credit Scoring dibangun dengan pendekatan machine learning menggunakan algoritma Naïve Bayes Classifier (GaussianNB) yang diimplementasikan melalui platform Google Colab dengan bahasa pemrograman Python. Data yang digunakan berasal dari Jaminan Kesehatan Nasional (JKN) masyarakat Kota Malang periode 2020–2023. Proses pengembangan sistem mencakup beberapa tahapan utama. Pengumpulan data melalui metode web scraping dari Portal Data JKN. Preprocessing data, termasuk data cleaning, normalisasi, dan penanganan missing value. Penentuan variabel, dengan empat variabel independen utama yaitu: usia (USIA), penghasilan (INCOME_TOTAL), jumlah tagihan bulanan (TOTALBILL_MONTH), dan durasi kerja (DAYS_EMPLOYED). Penentuan kriteria kelayakan kredit, berdasarkan kebijakan OJK dan literatur terkait. Pelatihan model Gaussian Naïve Bayes untuk mengklasifikasikan kelayakan kredit menjadi 5 skor, dari skor 1 (tidak layak) hingga skor 5 (sangat layak).
2. Hasil implementasi sistem menunjukkan bahwa model Naïve Bayes Classifier (GaussianNB) mampu melakukan klasifikasi kelayakan kredit dengan akurasi tinggi. Berdasarkan pengujian terhadap dataset yang berjumlah 30 sampel valid (yang digunakan untuk uji manual dan Sklearn). Model Sklearn GaussianNB menghasilkan akurasi 84%, yaitu seluruh prediksi skor kredit sesuai dengan data label actual. Hasil ini membuktikan bahwa algoritma Naïve Bayes efektif untuk dikembangkan dalam sistem alternative credit scoring berbasis data JKN.

Saran untuk Sistem alternative kredit scoring yang dibuat masih jauh dari kata sempurna dan masih banyak kekurangan. Untuk itu diperlukan penyempurnaan sistem agar dapat menjadi lebih baik. Adapun saran agar sistem ini bisa verjalan dengan lebih optimal sebagai berikut :

1. Penggunaan Dataset yang Lebih Luas dan Beragam
   Penelitian ini menggunakan sampel data dari Jaminan Kesehatan Nasional (JKN) masyarakat Kota Malang. Untuk pengembangan lebih lanjut, disarankan agar sistem diujikan pada dataset yang lebih besar dan mencakup wilayah geografis yang lebih luas agar hasil klasifikasi lebih general dan representatif terhadap kondisi masyarakat Indonesia secara keseluruhan.
2. Pengembangan Antarmuka Pengguna (User Interface). Sistem yang dibangun pada penelitian ini masih berfokus pada pengolahan dan analisis data berbasis kode di Google Colab. Untuk penerapan nyata di lembaga keuangan, disarankan agar sistem dikembangkan lebih lanjut dalam bentuk aplikasi dengan antarmuka pengguna (UI) yang ramah, agar dapat digunakan oleh staf operasional non-teknis maupun pengguna umum

DAFTAR PUSTAKA

[1] H. D. Fata, G. I. Marthasari, and Y. Azhar, "Sistem Pendukung Keputusan Kelayakan Kredit pada PT. BPR Mitra Catur Mandiri Menggunakan Metode Credit Scoring," *J. Repos.*, vol. 2, no. 5, pp. 649–658, 2020, doi: 10.22219/repositor.v2i5.608.

[2] S. A. Nurrizqi Rahmadania and N. Nurismalatri, "Analisis Credit Scoring Dan Faktor-Faktor Yang Mempengaruhi Npf Pembiayaan Murabahah Pada Pt Bank Muamalat Indonesia," *JABI (Jurnal Akunt. Berkelanjutan Indones.*, vol. 3, no. 2, pp. 204–221, 2020, doi: 10.32493/jabi.v3i2.y2020.p204-221.

[3] A. Ramaquita and A. Alamsyah, "Model Credit Scoring Berdasarkan Data Demografi Dan Jejaring Sosial Di Media Sosial," *e-Proceeding Manag.* , vol. 7, no. 2, pp. 2214–2219, 2020.

[4] Fatimah, M. Abdul Mukid, and A. Rusgiyono, "Analisis Credit Scoring Menggunakan Metode Bagging K-Nearest Neighbor," *J. Gaussian*, vol. 6, no. 1, pp. 161–170, 2017, [Online]. Available: http://ejournal-s1.undip.ac.id/index.php/gaussian

[5] DJSN, "Statistik JKN 2016-2021," *Dewan Jaminan Sos. Nas.*, pp. 1–252, 2022.

[6] S. Informasi, U. Darwan Ali, J. Batu Berlian, and K. Tengah, "Penerapan Metode Naive Bayes Dalam Mengklasifikasi Penerima BLT Pada Desa Pelangsian," *J. JUPITER*, vol. 14, no. 2, pp. 64–70, 2022.

[7] A. Nuriksan, T. Hendro Pudjiantoro, and P. Nurul Sabrina, "Klasifikasi Kelayakan Kredit Motor Menggunakan Metode Naïve Bayes Dan Model Credit Scoring," *Semin. Nas. Inform. dan Apl.*, pp. G1–G6, 2021.

[8] R. M. Samhadi1 and S. , Bambang Minto Basuki2, "Implementasi Multiple Linear Regression Pada Machine Learning Untuk Prediksi," pp. 1–7, 2022.

[9] E. T. Kurniawati, "Pengembangan sistem credit scoring pembiayaan UMKM," *Inovasi*, vol. 17, no. 3, pp. 609–616, 2021.

[10] M. R. A. A. S, B. M. Basuki, A. Habibi, I. Malang, and U. I. Malang, "IMPLEMENTASI ALGORITMA MACHINE LEARNING UNTUK PREDIKSI PENJUALAN CROISSANT PADA

CAFE RETAWUDELI BERBASIS HISTORI PENJUALAN TIME SERIES," pp. 1–7.

[11] Syarli and A. A. Muin, "Metode Naive Bayes Untuk Prediksi Kelulusan," *J. Ilm. Ilmu Komput.*, vol. 2, no. 1, pp. 22–26, 2020, [Online]. Available: https://media.neliti.com/media/publications/283828-metode-naive-bayes-untuk-prediksi-kelulu-139fcfea.pdf

[12] D. W. Triscowati and L. D. Jayanti, "Penilaian Kredit Pada Data Tak Seimbang Menggunakan Random Forest Credit Scoring In Imbalance Data Using Random Forest," *JIKOSTIK – J. Ilm. Komputasi dan Stat.*, vol. 1, no. 1, pp. 25–31, 2021.

[13] D. Normawati and S. A. Prayogi, "Implementasi Naïve Bayes Classifier Dan Confusion Matrix Pada Analisis Sentimen Berbasis Teks Pada Twitter," *J. Sains Komput. Inform. (J-SAKTI*, vol. 5, no. 2, pp. 697–711, 2021.

[14] H. F. Putro, R. T. Vulandari, and W. L. Y. Saptomo, "Penerapan Metode Naive Bayes Untuk Klasifikasi Pelanggan," *J. Teknol. Inf. dan Komun.*, vol. 8, no. 2, 2020, doi: 10.30646/tikomsin.v8i2.500.

[15] I. D. S. Tarigan, Roni Habibi, and Rd. Nuraini Siti Fatonah, "Evaluasi Algoritma Klasifikasi Machine Learning Kategori Nilai Akhir Tunjangan Kinerja Pegawai," *J. Sist. Cerdas*, vol. 6, no. 3, pp. 251–261, 2023, doi: 10.37396/jsc.v6i3.246.