PENERAPAN METODE NAÏVE BAYES UNTUK KLASIFIKASI KONDISI KEUANGAN UMKM BERDASARKAN LAPORAN KEUANGAN

Ridzkyan Buti Pratama[1*], Pipin Widyaningsih[2], Joni Maulindar[3]

[1]Teknik Informatika/Ilmu Komputer
Universitas Duta Bangsa
[1*]220103245@mhs.udb.ac.id

[2]Sistem Informasi/Ilmu Komputer
Universitas Duta Bangsa
[2]pipin_widyaningsih@udb.ac.id

[3]Teknik Informatika/Ilmu Komputer
Universitas Duta Bangsa
[3]joni_maulindar@udb.ac.id

Abstrak*— Usaha Mikro, Kecil, dan Menengah (UMKM) mempunyai peranan krusial pada perekonomian Indonesia, tetapi masih banyak pelaku usaha yang mengalami kesulitan pada menganalisis kondisi keuangan usahanya. Permasalahan utama terletak pada pencatatan keuangan yang sederhana sehingga sulit menentukan apakah usaha berada dalam kondisi sehat atau tidak sehat. Penelitian ini bertujuan untuk mengklasifikasikan kondisi keuangan UMKM menggunakan metode *Gaussian Naïve Bayes* berdasarkan data laporan keuangan yang terdiri dari pendapatan bulanan, margin keuntungan bersih, rasio pembakaran modal, jumlah transaksi, tingkat pemesanan ulang, dan lama usaha. Dataset yang digunakan berjumlah 150.000 data sintetis yang diperoleh dari platform *Kaggle* dengan empat kategori kelas: *Elite, Growth, Struggling,* dan *Critical*. Label klasifikasi ditentukan dengan menggabungkan kelas *Elite* dan *Growth* sebagai kondisi Sehat, serta *Struggling* dan *Critical* sebagai kondisi Tidak Sehat. Temuan analisis memperlihatkan bahwasanya metode *Gaussian Naïve Bayes* bisa melakukan klasifikasi kondisi keuangan UMKM dengan tingkat akurasi sebesar 94,70%, *F1-Score* sebesar 94,72%, dan nilai *ROC-AUC* sebesar 0,9859 yang menandakan model yang stabil dan konsisten. Penelitian ini menunjukkan bahwasanya metode Naïve Bayes bisa dipakai selaku alat bantu klasifikasi kondisi keuangan UMKM yang sederhana, efisien, sehingga bisa dipakai selaku alat bantu pada pengambilan keputusan untuk pelaku usaha.

Kata kunci— Gaussian Naive Bayes, klasifikasi, UMKM, laporan keuangan, data mining.

Abstract*— Micro, Small, and Medium Enterprises (MSMEs) play a crucial role in Indonesia's economy, yet many business owners continue to struggle with analyzing their financial conditions. The core challenge stems from rudimentary bookkeeping practices, which make it difficult to determine whether a business is financially sound or not. This study aims to classify the financial health of MSMEs using the Gaussian Naïve Bayes method, drawing on financial report data comprising monthly revenue, net profit margin, capital burn ratio, transaction volume, reorder rate, and years in operation. The dataset consists of 150,000 synthetic records sourced from the Kaggle platform, spanning four class categories: Elite, Growth, Struggling, and Critical. Classification labels were established by merging the Elite and Growth classes into a "Healthy" condition, while Struggling and Critical were grouped under "Unhealthy." The analysis reveals that the Gaussian Naïve Bayes method is capable of classifying MSME financial conditions with an accuracy rate of 94.70%, an F1-Score of 94.72%, and a ROC-AUC value of 0.9859, reflecting a stable and consistent model performance. This research demonstrates that the Naïve Bayes method can serve as a straightforward and efficient classification tool for assessing MSME financial conditions, and may consequently be adopted as a decision-support instrument for business practitioners.

Keywords— Gaussian Naïve Bayes, classification, MSMEs, financial statements, data mining.

I. PENDAHULUAN

1.1 Latar Belakang

Usaha Mikro, Kecil, dan Menengah (UMKM) merupakan sektor ekonomi yang memiliki kontribusi besar terhadap pertumbuhan ekonomi nasional. Selain berperan pada menghasilkan lapangan kerja, UMKM juga menjadi salah satu penggerak utama aktivitas ekonomi masyarakat di berbagai daerah. Seiring dengan meningkatnya jumlah UMKM, kemampuan pelaku usaha dalam mengelola aspek keuangan menjadi faktor penting yang menentukan keberlangsungan dan perkembangan usaha [1].

Dalam praktiknya, masih banyak pelaku UMKM yang menghadapi kendala dalam pengelolaan dan analisis keuangan. Proses pencatatan transaksi usaha sering kali dilakukan secara sederhana sehingga informasi mengenai pendapatan, pengeluaran, dan keuntungan belum terdokumentasi dengan baik [2]. Keadaan ini membuat pelaku usaha kesulitan pada mencermati situasi keuangan usahanya secara jelas, apakah berada dalam kondisi sehat atau tidak sehat. Padahal, informasi tersebut sangat penting sebagai dasar dalam pengambilan keputusan bisnis [3].

Berdasarkan permasalahan tersebut, diperlukan suatu metode yang sederhana, efisien, dan mudah diimplementasikan untuk membantu pelaku UMKM dalam menganalisis kondisi keuangan usahanya. Dalam penelitian ini digunakan enam indikator laporan keuangan, yaitu omzet bulanan, margin keuntungan bersih, rasio pengeluaran modal, jumlah transaksi, tingkat pemesanan ulang, dan lama usaha. Indikator tersebut digunakan karena dapat menggambarkan kondisi operasional dan kestabilan.

Termasuk metode yang dipakai yakni *Naïve Bayes*, khususnya *Gaussian Naïve Bayes* yaitu algoritma klasifikasi berbasis probabilitas yang mampu menangani data numerik dengan proses komputasi yang cepat dan efisien [4]. Selain itu, metode ini memiliki performa klasifikasi yang baik serta bisa mengolah data pada jumlah besar dengan kompleksitas perhitungan yang relatif sederhana. Karakteristik tersebut

membuat Gaussian Naïve Bayes sesuai digunakan untuk menganalisis data laporan keuangan UMKM yang terdiri dari beberapa indikator numerik seperti pemasukan, pengeluaran, dan laba usaha

Berdasarkan latar belakang tersebut, penelitian ini bertujuan untuk menerapkan metode Gaussian Naïve Bayes dalam mengklasifikasikan kondisi keuangan UMKM berdasarkan indikator laporan keuangan ke dalam kategori Sehat dan Tidak Sehat.

1.2 Tinjauan Pustaka

Sebelum melaksanakan kajian ini, penulis lebih dulu studi literatur dengan menelusuri berbagai referensi yang relevan, baik berupa buku, jurnal ilmiah nasional maupun internasional, serta sumber daring yang berkaitan dengan klasifikasi dan pengelolaan data keuangan UMKM. Penelitian terkait klasifikasi dan pengelolaan data keuangan UMKM telah banyak dilakukan dengan berbagai macam pendekatan. Penelitian oleh Haris et al. menerapkan algoritma Naïve Bayes guna melakukan klasifikasi sentimen terhadap ulasan pemakai aplikasi Pidi UMKM dengan pembobotan fitur TF-IDF [5]. Penelitian tersebut memperlihatkan bahwasanya metode Naïve Bayes bisa mengategorikan data teks ulasan ke kategori sentimen positif, negatif, serta netral dengan tingkat akurasi sebesar 65%. Hasil ini memperlihatkan bahwasanya Naïve Bayes efektif digunakan dalam proses klasifikasi data berbasis teks dengan pendekatan yang relatif sederhana dan efisien. Namun, penelitian tersebut berfokus pada analisis sentimen ulasan pengguna dan belum membahas kondisi keuangan UMKM secara langsung.

Selanjutnya, penelitian oleh Erkamim et al. melakukan klasifikasi performa UMKM menggunakan algoritma *Random Forest* dan *XGBoost* yang menghasilkan akurasi yang tinggi dalam mengelompokkan data keuangan [6]. Meskipun demikian, metode tersebut memiliki kompleksitas yang cukup tinggi sehingga kurang efisien untuk diterapkan pada UMKM dengan keterbatasan sumber daya dan pemahaman teknologi.

Penelitian lain yang dilakukan oleh Huriah dan Nuris menerapkan metode *Naïve Bayes* untuk mengklasifikasikan penerima bantuan sosial UMKM. Temuan analisis memperlihatkan bahwasanya metode Naïve Bayes bisa memberi hasil klasifikasi yang lumayan baik dengan proses yang sederhana dan efisien [7]. Namun, penelitian tersebut berfokus pada kelayakan penerima bantuan dan belum membahas analisis kondisi keuangan usaha.

Berdasarkan kajian literatur tersebut, terdapat celah penelitian pada penerapan *Gaussian Naïve Bayes* untuk klasifikasi kondisi keuangan UMKM. Maka karenanya, kajian ini mengembangkan penerapan metode itu untuk mengklasifikasikan kondisi keuangan UMKM berdasarkan laporan keuangan sederhana yang terdiri dari pendapatan, beban, dan laba.

1.2.1. Usaha Mikro, Kecil, dan Menengah (UMKM)

Usaha Mikro, Kecil, dan Menengah (UMKM) merupakan salah satu sektor yang memiliki peran penting dalam perekonomian di Indonesia, terutama dalam menciptakan lapangan kerja dan meningkatkan kesejahteraan masyarakat [3]. Dalam praktiknya, kebanyakan UMKM masih mengalami hambatan pada pengelolaan keuangan, terkhusus pada pencatatan serta analisis laporan keuangan. Keterbatasan pengetahuan serta minimnya penggunaan teknologi menyebabkan tahapan pengelolaan keuangan masih dilaksanakan dengan sederhana serta manual, maka berpeluang memunculkan kesalahan dalam mengambil keputusan [2].

1.2.2. Laporan Keuangan UMKM

Laporan keuangan merupakan informasi yang menjabarkan situasi keuangan suatu usaha dalam periode tertentu. Pada UMKM, laporan keuangan umumnya terdiri dari beberapa komponen utama yaitu pendapatan, beban, dan laba. Informasi tersebut digunakan untuk mengetahui kinerja usaha serta sebagai dasar dalam pengambilan keputusan. Namun, keterbatasan dalam mencatat dan menganalisis menyebabkan laporan keuangan sering kali tidak optimal [8]. Maka karenanya, dibutuhkan suatu metode yang bisa membantu dalam mengolah serta menganalisis data keuangan dengan otomatis.

1.2.3. Data Mining

Data mining merupakan tahap pengolahan data guna mengetahui pola ataupun informasi yang berguna dari sejumlah data yang besar. Tahapan ini dilakukan dengan menggunakan teknik statistik, matematika, serta machine learning guna menghasilkan wawasan baru yang bisa dipakai di pengambilan keputusan. Proses data mining terdiri dari beberapa proses, yakni pengumpulan data, pembersihan data (data cleaning), transformasi data, pemilihan metode, serta proses evaluasi hasil. Salah satu Teknik pada data mining adalah klasifikasi, yaitu tahap pengategorian data ke dalam kategori khusus berlandaskan karakteristik yang dimiliki [9]. Dalam penelitian penulis, teknik klasifikasi digunakan untuk menentukan kondisi keuangan UMKM ke dalam kategori sehat dan tidak sehat.

1.2.4. Metode Gaussian Naïve Bayes

Naïve Bayes yakni termasuk algoritma klasifikasi yang mempunyai basis kepada *Teorema Bayes* dengan asumsi bahwasanya tiap fitur mempunyai sifat independent satu sama lain [4]. Varian *Gaussian Naïve Bayes* digunakan demi data numerik kontinu, di mana likelihood setiap fitur dihitung menggunakan fungsi distribusi Gaussian sebagai berikut [10]:

$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} e^{-\frac{(x_i-\mu_y)^2}{2\sigma_y^2}} \quad (1)$$

Keterangan:

a. x_i= nilai fitur
b. μ_y= rata-rata fitur pada kelas
c. σ_y^2= variansi fitur pada kelas
d. y= kelas (Sehat / Tidak Sehat)

Untuk beberapa fitur, probabilitas dihitung dengan [4]:

$$P(y \mid X) = P(y)\prod P(x_i \mid y) \quad (2)$$

Model memilih kelas dengan nilai posterior tertinggi selaku hasil klasifikasi.

1.2.5. Evaluasi Model Klasifikasi

Evaluasi model klasifikasi dilaksanakan guna mencermati tingkat kinerja dari model yang dibangun. Termasuk metode yang umum dipakai yaitu confusion matrix, yang dipakai guna memperbandingkan temuan prediksi dengan data asli. Dari confusion matrix, dapat dihitung beberapa metrik evaluasi seperti akurasi, *precision*, *recall*, serta *F1-score* [11]. Metrik tersebut dipakai guna mengukur seberapa baik model pada melaksanakan klasifikasi data, Adapun rumusnya bisa dilihat dibawah ini [10].

a. Akurasi

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

b. Precision

$$Precision = \frac{TP}{TP+FP} \quad (4)$$

c. Recall

$$Recall = \frac{TP}{TP+FN} \quad (5)$$

d. F1-Score

$$F1 = \frac{2TP}{2TP+FP+FN} \quad (6)$$

II. METODOLOGI PENELITIAN

2.1 Pengumpulan Data

Data yang dipakai di analisis ini yakni dataset sintetis yang berasal dari *Kaggle* yang merepresentasikan kondisi keuangan Usaha Mikro, Kecil, dan Menengah (UMKM). Dataset tersusun atas 150.000 data dengan 15 atribut yang mencakup indikator keuangan dan operasional usaha. Penggunaan data sintetis dalam penelitian ini mempunyai tujuan untuk menjamin ketersediaan data dalam jumlah yang memadai guna tahap pelatihan dan validasi model machine learning.

2.2 Eksplorasi Data (EDA)

Sebelum tahap preprocessing dilakukan, terlebih dahulu dilakukan proses eksplorasi data untuk memahami karakteristik dari dataset yang digunakan. Tahap ini mencakup beberapa analisis berikut:

a. Statistik Deskriptif, yaitu pemeriksaan nilai minimum, maksimum, rata-rata, serta standar deviasi pada tiap atribut numerik. Analisis ini bertujuan untuk mengidentifikasi adanya nilai yang tidak wajar atau indikasi anomali pada data.
b. Analisis Distribusi Kelas, yaitu evaluasi proporsi masing-masing kelas dalam dataset untuk mengetahui apakah terdapat ketidakseimbangan kelas (class imbalance) yang berpotensi memengaruhi kinerja model.
c. Analisis Korelasi, yang dilakukan dengan menghitung koefisien korelasi Pearson antar fitur keuangan dan memvisualisasikannya dalam bentuk heatmap. Tujuannya adalah untuk melihat hubungan antar variabel serta mendeteksi kemungkinan *multikolinearitas*.
d. Analisis Distribusi Fitur terhadap Label, yaitu visualisasi distribusi setiap fitur keuangan berdasarkan label menggunakan metode *Kernel Density Estimation* (KDE). Analisis ini digunakan untuk menilai tingkat kemampuan diskriminasi masing-masing fitur terhadap kelas target.

2.3 Preprocessing Data

Tahap preprocessing mempunyai tujuan mempersiapkan data supaya bersih, konsisten, serta siap dipakai pada tahap pelatihan model. Preprocessing dilakukan melalui empat langkah berikut:

a. Seleksi Fitur

Tabel 1. Fitur model

No	Nama Fitur	Keterangan
1	Monthly_revenue	Pendapatan bulanan usaha
2	net_profit_margin (%)	Persentase laba bersih terhadap pendapatan
3	Burn_rate_ratio	Rasio pengeluaran terhadap pendapatan
4	Transaction_count	Jumlah transaksi per bulan
5	Repeat_order_rate (%)	Persentase pelanggan yang melakukan pemesanan ulang
6	Business_tenure_months	Lama usaha berjalan dalam bulan

Pada Tabel 1, dipilih enam fitur yang secara langsung merepresentasikan indikator laporan keuangan UMKM. Pemilihan fitur didasarkan pada relevansi terhadap kondisi keuangan usaha dan kesesuaian dengan judul penelitian.

b. Pemberian Label

Tabel 2. Distribusi label

Label	Kategori	Jumlah	Persentase
1	Sehat	95.868	63,9%
0	Tidak Sehat	54.132	36,1%
	Total	150.000	100%

Label klasifikasi tidak diambil langsung dari kolom Class, melainkan ditentukan melalui proses binarisasi berdasarkan kategori kelas asli. Kelas Elite dan Growth yang merepresentasikan kondisi keuangan positif dikategorikan sebagai Sehat (1), sedangkan kelas Struggling dan Critical yang merepresentasikan kondisi keuangan bermasalah dikategorikan sebagai Tidak Sehat (0). Proses binarisasi ini menghasilkan distribusi sebagaimana ditunjukkan pada Tabel 2.

c. Normalisasi Data

Seluruh fitur numerik dilakukan proses normalisasi ke dalam rentang [0, 1] memakai metode *Min-Max Normalization* dengan x merupakan nilai asli, x_min adalah nilai minimum, x_max adalah nilai maksimum, dan x' merupakan nilai temuan normalisasi. Proses normalisasi ini diperlukan karena setiap fitur memiliki skala yang berbeda-beda, misalnya *Monthly_Revenue* yang berada pada skala jutaan Rupiah, sedangkan *Burn_Rate_Ratio* hanya berada pada rentang 0 hingga 2. Perbedaan skala tersebut dapat memengaruhi kinerja model apabila tidak diseragamkan.

Tahapan normalisasi hanya dilakukan menggunakan data latih (training set) untuk menghindari terjadinya *data leakage*. Nilai minimum dan maksimum yang diperoleh dari data latih kemudian digunakan sebagai parameter transformasi yang sama pada data uji (testing set), sehingga konsistensi proses normalisasi tetap terjaga.

d. Pembagian Data

Tabel 3. Pembagian dataset

Subset	Sehat	Tidak Sehat	Total
Data Latih (80%)	76.694	43.306	120.000
Data Uji (20%)	19.174	10.826	30.000
Total	95.868	54.132	150.000

Berdasarkan Tabel 3, dataset dibagi menjadi data latih serta data uji dengan perbandingan 80:20 memakai metode stratified sampling. Stratified sampling dipilih guna memastikan proporsi setiap kelas di data latih dan data uji tetap representatif terhadap distribusi kelas pada dataset keseluruhan. Pembagian data dilakukan sebelum proses normalisasi untuk memastikan tidak ada informasi dari data uji yang bocor ke proses pelatihan model.

2.4 Pembangunan Model dan Pipeline

Penelitian ini mengimplementasikan model klasifikasi menggunakan *Gaussian Naïve Bayes* yang dipadukan dengan proses normalisasi data menggunakan *Min-Max Scaler*. Seluruh proses preprocessing dan pelatihan model diintegrasikan ke dalam sebuah *machine learning pipeline* untuk menjaga konsistensi alur pemrosesan data serta mencegah terjadinya data leakage.Pipeline terdiri dari dua tahap utama, yaitu transformasi data menggunakan *ColumnTransformer* dengan *MinMaxScaler*, serta proses klasifikasi menggunakan *Gaussian Naïve Bayes*. Model dilatih menggunakan data latih dan kemudian dievaluasi menggunakan data uji. Langkah-langkah pelatihan model adalah sebagai berikut:

1. Pipeline dilatih memakai data latih dengan fungsi fit(X_train, y_train).
2. MinMaxScaler menghitung nilai minimum dan maksimum berdasarkan data latih.
3. Data latih kemudian dinormalisasi dan digunakan untuk melatih model Gaussian Naïve Bayes.
4. Prediksi dilakukan pada data uji menggunakan fungsi predict(X_test) dan predict_proba(X_test).

2.5 Evaluasi Model

Tabel 4. Komponen Confusion Matrix

	Prediksi Positif	Prediksi Negatif
Aktual Positif	True Positive (TP)	False Negative (FN)
Aktual Negatif	False Positive (FP)	True Negative (TN)

Evaluasi model dilaksanakan memakai confusion matrix yang memperbandingkan temuan prediksi model dengan label aktual. Confusion matrix menghasilkan empat nilai sebagaimana ditunjukkan pada Tabel 4.

Berdasarkan nilai confusion matrix, dihitung empat metrik evaluasi utama yakni akurasi, precision, recall, F1-score. Seluruh metrik dihitung menggunakan pendekatan *weighted average* untuk menyesuaikan ketidakseimbangan jumlah data antar kelas.

2.6 Kurva ROC dan AUC

Selain metrik utama, evaluasi model juga dilakukan memakai kurva ROC (*Receiver Operating Characteristic*) serta nilai AUC (*Area Under Curve*). Kurva ROC menguraikan hubungan antara True Positive Rate (TPR) serta False Positive Rate (FPR) di bermacam nilai ambang (threshold). Nilai AUC ada di rentang 0 sampai 1, yang mana nilai yang mendekati 1 memperlihatkan kemampuan model yang kian baik pada membedakan antar kelas.

2.7 Validasi Silang 10-Fold

Untuk memastikan kestabilan dan kemampuan generalisasi model, dijalankan validasi silang memakai metode *10-Fold Stratified Cross Validation*. Proses ini hanya diterapkan pada data latih agar data uji tetap sepenuhnya independen.

Mekanisme validasi dilakukan sebagai berikut:

a. Data latih dibagi menjadi 10 bagian (fold) dengan proporsi kelas yang tetap seimbang.
b. Di tiap iterasi, 9 fold dipakai selaku data latih dan 1 fold sebagai data validasi.
c. Pipeline dilatih ulang pada setiap iterasi, termasuk proses normalisasi data.
d. Proses ini diulang sebanyak 10 kali hingga setiap fold menjadi data validasi satu kali.
e. Hasil evaluasi dari setiap iterasi kemudian dirata-ratakan.

Metrik yang dipakai di validasi ini meliputi akurasi, precision, recall, F1-score, serta ROC-AUC. Nilai rata-rata serta standar deviasi dipakai guna menilai kestabilan model.

2.8 Analisis Kepentingan Fitur

Analisis kepentingan fitur dilakukan untuk mengetahui kontribusi masing-masing variabel dalam proses klasifikasi. Metode yang digunakan adalah pendekatan selisih rata-rata antar kelas pada ruang fitur yang telah dinormalisasi Semakin besar selisih rata-rata tersebut, maka semakin tinggi kemampuan fitur dalam membedakan kedua kelas. Analisis ini hanya dilakukan pada data latih yang telah melalui proses normalisasi untuk menghindari bias dari data

III. HASIL DAN PEMBAHASAN

3.1 Eksplorasi Data

Tahapan awal analisis dilakukan melalui proses eksplorasi data (Exploratory Data Analysis/EDA) guna mencermati karakteristik dataset sebelum digunakan dalam proses pemodelan klasifikasi. Dataset yang dipakai tersusun dari 150.000 data dengan 15 atribut, yakni *ID, Monthly_Revenue, Net_Profit_Margin (%), Burn_Rate_Ratio, Transaction_Count, Avg_Historical_Rating, Review_Text, Review_Volatility, Business_Tenure_Months, Repeat_Order_Rate (%), Digital_Adoption_Score, Peak_Hour_Latency, Location_Competitiveness, Sentiment_Score,* dan *Class*. Namun, pada proses pemodelan penelitian ini hanya digunakan enam atribut utama, yaitu *Monthly_Revenue, Net_Profit_Margin (%), Burn_Rate_Ratio, Transaction_Count, Business_Tenure_Months, dan Repeat_Order_Rate (%).* Keenam atribut tersebut dipilih karena dinilai mampu merepresentasikan kondisi keuangan dan aktivitas operasional UMKM dalam proses klasifikasi kondisi usaha. Di tahapan ini dijalankan analisis statistik deskriptif guna mengetahui distribusi data, nilai rata-rata, standar deviasi, nilai minimum, serta maksimum dari setiap atribut numerik. Temuan ringkasan statistik deskriptif diperlihatkan di Tabel 5.

Tabel 5. Statistik Deskriptif

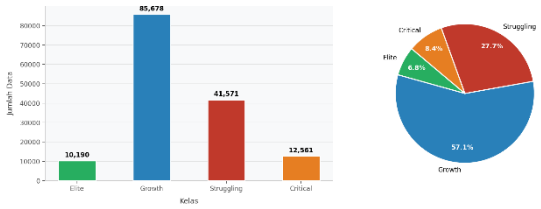
Atribut	Min	Max	Mean	Std Dev
Monthly_Revenue (Rp)	1.500.000	82.067.536	8.451.726	5.291.163
Net_Profit_Margin (%)	-35,00	45,00	1,84	15,00
Burn_Rate_Ratio	0,44	1,55	0,97	0,14
Transaction_Count	9	285	117,77	42,62
Repeat_Order_Rate (%)	2,00	54,06	19,98	8,02
Business_Tenure_Months	3	179	91,01	51,10

Berdasarkan hasil tersebut, terlihat bahwa setiap fitur memiliki variasi nilai yang cukup berbeda. Misalnya, *Net_Profit_Margin* memiliki rentang nilai yang cukup lebar dari -35% hingga 45%, namun dengan rata-rata yang rendah yaitu 1,84%. Hal ini memperlihatkan bahwasanya kebanyakan UMKM ada di kondisi margin keuntungan yang tipis. Sementara itu, *Burn_Rate_Ratio* memiliki nilai rata-rata mendekati 1, yang berarti secara umum pengeluaran UMKM hampir sebanding dengan pendapatannya. Kondisi ini mengindikasikan bahwa banyak UMKM berada pada situasi keuangan yang cukup rentan.Pada fitur *Monthly_Revenue*, terlihat variasi yang sangat besar, yang ditunjukkan oleh standar deviasi yang tinggi. Hal ini mencerminkan adanya perbedaan skala usaha yang signifikan di

dalam dataset, mulai dari usaha kecil hingga yang memiliki pendapatan jauh lebih besar.

3.2 Distribusi Kelas dan Label

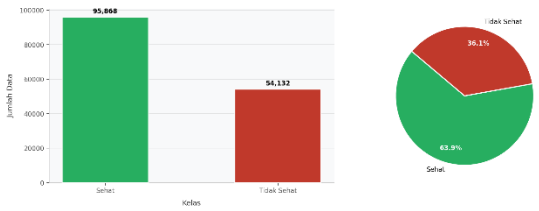
Gambar 1. Distribusi Kelas Dataset UMKM



Gambar 1. Distribusi kelas

Dataset awal terdiri dari empat kategori kondisi keuangan, yaitu Elite, Growth, Struggling, dan Critical. Hasil distribusi menunjukkan bahwa kelas Growth merupakan kelas yang paling dominan, diikuti oleh Struggling. Sementara itu, kelas Elite dan Critical memiliki jumlah yang lebih sedikit. Untuk mempermudah proses klasifikasi, dilakukan penggabungan kelas menjadi dua kategori, yaitu Sehat (Elite dan Growth) serta Tidak Sehat (Struggling dan Critical). Hasil transformasi ini menghasilkan 95.868 data Sehat (63,9%) dan 54.132 data Tidak Sehat (36,1%) sebagaimana ditampilkan pada Gambar 2.

Gambar 2. Distribusi Label Sehat / Tidak Sehat



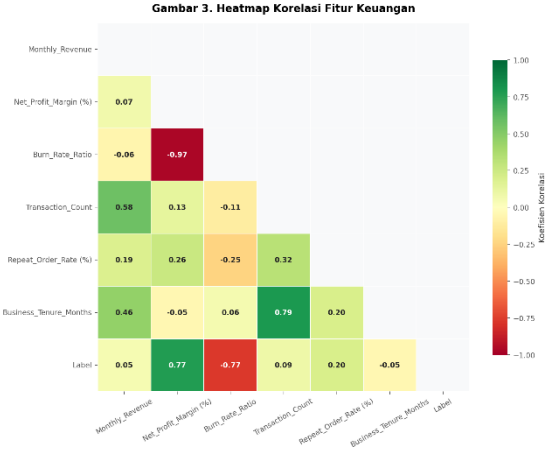
Gambar 2. Distribusi label

Terdapat ketidakseimbangan kelas dengan rasio 63,9% : 36,1%. Kondisi ini diatasi dengan penggunaan *stratified sampling* pada proses pembagian data serta penggunaan *metrik weighted average* pada evaluasi model.

3.3 Analisis Korelasi dan Distribusi Fitur

Analisis hubungan antar fitur dilakukan menggunakan korelasi Pearson dan divisualisasikan dalam bentuk heatmap.

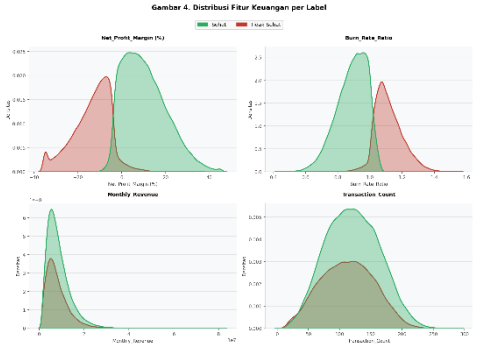
Gambar 3. Heatmap Korelasi Fitur Keuangan



Gambar 3. Korelasi heatmap

Temuan kajian menjabarkan bahwasanya *Net_Profit_Margin* serta *Burn_Rate_Ratio* memiliki hubungan negatif yang cukup kuat. Hal ini secara logis dapat dipahami, karena semakin tinggi margin keuntungan, maka tingkat pemborosan modal cenderung semakin rendah. Selain itu, *Net_Profit_Margin* juga memiliki hubungan paling kuat terhadap label kelas, sehingga dapat dianggap sebagai indikator utama dalam menentukan kondisi keuangan UMKM. Sementara itu, fitur lain seperti *Monthly_Revenue*, *Transaction*_Count, dan *Business_Tenure_Months* memiliki korelasi yang lebih lemah terhadap label, tetapi tetap memberikan kontribusi informasi dalam model.

Gambar 4. Distribusi Fitur Keuangan per Label



Gambar 4. Distribusi fitur

Dari hasil distribusi fitur menggunakan KDE, terlihat adanya perbedaan pola yang cukup jelas antara kelas Sehat dan Tidak Sehat, terutama pada fitur *Net_Profit_Margin* dan *Burn_Rate_Ratio*. UMKM yang tergolong Sehat cenderung memiliki margin keuntungan yang lebih tinggi serta rasio pembakaran modal yang lebih rendah daripada UMKM Tidak Sehat. Perihal ini mengemukakan bahwasanya kedua fitur tersebut mempunyai kemampuan yang kuat pada membedakan kelas.

3.4 Preprocessing dan Pembagian Data

Setelah eksplorasi data, dilakukan proses preprocessing yang meliputi seleksi enam fitur keuangan, normalisasi Min-Max, dan pembagian data. Pembagian data dilakukan sebelum normalisasi menggunakan stratified sampling dengan perbandingan 80:20 untuk mencegah data leakage.

Berdasarkan tabel 3, proporsi kelas di data latih serta data uji identik dengan proporsi di dataset keseluruhan (63,9% : 36,1%), yang membuktikan bahwa *stratified sampling* berhasil menjaga distribusi kelas di kedua subset. Normalisasi Min-Max kemudian diterapkan melalui *Pipeline scikit-learn* sehingga parameter normalisasi hanya diperoleh dari data latih dan tidak ada informasi dari data uji yang bocor ke proses pelatihan.

3.5 Evaluasi Model pada Data Uji

Model Gaussian Naïve Bayes dilatih menggunakan data latih sebanyak 120.000 data melalui Pipeline yang menggabungkan proses normalisasi dan klasifikasi. Berdasarkan data latih, diperoleh probabilitas awal sebesar 0,6391 pada kelas Sehat serta 0,3609 pada kelas Tidak Sehat.

Tabel 6. Evaluasi model

Metrik	Nilai
Accuracy	94,70%
Precision (weighted)	94,76%
Recall (weighted)	94,70%
F1-Score (weighted)	94,72%
ROC-AUC	0,9859

Evaluasi dilakukan menggunakan 30.000 data uji yang benar-benar terpisah dari proses pelatihan. Temuan evaluasi mengemukakan bahwasanya model menciptakan akurasi sebesar 94,70%, dengan nilai precision 94,76%, recall 94,70%, F1-score 94,72%, serta ROC-AUC sebesar 0,9859. Secara keseluruhan, model mampu mengklasifikasikan 28.410 data dengan benar dari total 30.000 data uji.

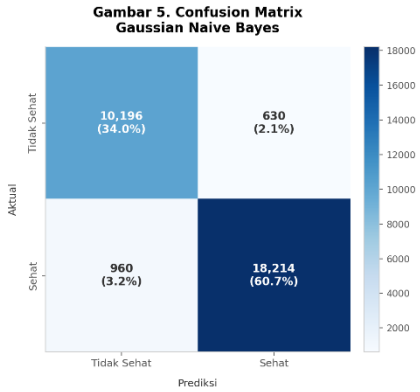
Tabel 7. Evaluasi per kelas

Kelas	Precision	Recall	F1-Score	Support

Tidak Sehat (0)	0,91	0,94	0,93	10.826
Sehat (1)	0,97	0,95	0,96	19.174
Macro Average	0,94	0,95	0,94	30.000
Weighted Average	0,95	0,95	0,95	30.000

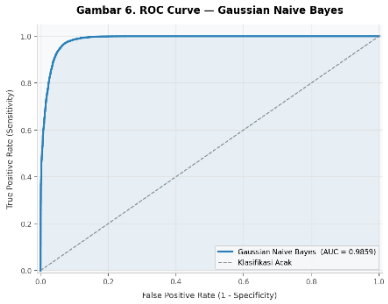
Jika dilihat per kelas, performa model lebih tinggi pada kelas Sehat dibandingkan Tidak Sehat. Namun, perbedaan tersebut relatif kecil maka menjabarkan bahwasanya model tetap lumayan seimbang pada mengenali kedua kelas.

3.6 Analisis Confusion Matrix dan ROC



Gambar 5. Confusion matrix

Temuan confusion matrix memperlihatkan bahwasanya kebanyakan data berhasil dikategorikan dengan benar. Pada kelas Tidak Sehat, sebagian besar prediksi sudah tepat, meskipun masih terdapat sebagian kecil yang salah diklasifikasikan sebagai Sehat. Hal serupa juga terjadi pada kelas Sehat. Menariknya, jumlah kesalahan berupa False Negative sedikit lebih tinggi dibandingkan False Positive. Artinya, model lebih cenderung mengklasifikasikan UMKM Sehat menjadi Tidak Sehat dibandingkan sebaliknya. Dalam konteks praktis, hal ini dapat dianggap lebih aman karena lebih baik memberikan peringatan berlebih daripada melewatkan UMKM yang sebenarnya bermasalah.



Gambar 6. Kurva ROC

Sementara itu, kurva ROC menunjukkan bahwa model memiliki kemampuan diskriminasi yang sangat baik, dengan nilai AUC sebesar 0,9859. Perihal ini menandakan bahwasanya model bisa membedakan kedua kelas dengan tingkat akurasi yang tinggi pada berbagai threshold.

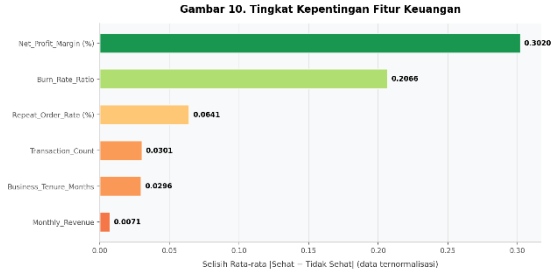
3.7 Validasi Silang (Cross Validation)

Tabel 8. Hasil Validasi silang 10-fold

Metrik	Mean	Std Dev	Min	Max
Accuracy	95,02%	0,18%	94,73%	95,38%
Precision	95,08%	0,19%	94,77%	95,45%
Recall	95,02%	0,18%	94,73%	95,38%
F1-Score	95,04%	0,18%	94,74%	95,39%
ROC-AUC	0,9869	0,0012	0,9838	0,9884

Berdasarkan Tabel 4.6, rata-rata akurasi cross validation sebesar 95,02% hanya berbeda 0,32% dari akurasi data uji sebesar 94,70%. Selisih yang sangat kecil ini mengindikasikan bahwasanya model tidak mengalami overfitting serta bisa melaksanakan generalisasi dengan baik kepada data baru. Nilai standar deviasi akurasi yang sangat kecil yaitu 0,18% menunjukkan performa model sangat konsisten di seluruh 10 fold, dengan rentang akurasi hanya 0,65% antara fold terbaik (95,38%) dan fold terburuk (94,73%).

3.8 Analisis Kepentingan Fitur



Gambar 7. Tingkat kepentingan fitur

Analisis menunjukkan bahwa fitur paling berpengaruh dalam proses klasifikasi adalah *Net_Profit_Margin*, diikuti oleh *Burn_Rate_Ratio*. Kedua fitur ini memiliki kontribusi paling besar dalam membedakan kondisi keuangan UMKM. Sementara itu, *Monthly_Revenue* memiliki kontribusi paling kecil, yang menunjukkan bahwa besarnya pendapatan tidak selalu mencerminkan kondisi kesehatan keuangan. Fitur lain seperti *Transaction_Count*, *Business_Tenure_Months*, dan *Repeat_Order_Rate* memberikan kontribusi tambahan namun tidak dominan.

3.9 Analisis dan Pembahasan

Berdasarkan hasil pengujian, metode *Gaussian Naïve Bayes* bisa mengategorikan kondisi keuangan UMKM dengan baik. Model mendapati akurasi sebesar 94,70% dengan nilai *precision*, *recall*, serta *F1-score* yang juga tinggi. Hasil ini mengemukakan bahwasanya model dapat membedakan kondisi UMKM sehat dan tidak sehat secara cukup akurat berdasarkan data laporan keuangan. Dataset yang dipakai di analisis ini yakni dataset sintetis yang didapati dari *Kaggle*. Penggunaan dataset sintetis memungkinkan penelitian dilakukan dengan jumlah data yang besar serta terstruktur, namun pola data cenderung lebih teratur dibandingkan kondisi nyata pada UMKM di lapangan.

Fitur yang paling berpengaruh dalam proses klasifikasi adalah *Net_Profit_Margin* dan *Burn_Rate_Ratio*. Hal ini menunjukkan bahwa tingkat keuntungan dan pengelolaan pengeluaran menjadi faktor utama dalam menentukan kondisi keuangan UMKM. Selain itu, hasil *cross validation* menunjukkan performa model yang stabil sehingga model mampu melakukan generalisasi dengan cukup baik. Meskipun demikian, model masih memiliki keterbatasan ketika menerima input data yang berada di luar rentang dataset pelatihan. Kondisi tersebut dapat menyebabkan error atau penurunan performa prediksi karena distribusi data baru berbeda dari pola yang telah dipelajari model. Oleh karena itu, penelitian selanjutnya disarankan menggunakan data riil UMKM dan memperluas variasi dataset agar model menjadi lebih robust.

IV. KESIMPULAN

Kajian ini sukses menerapkan metode *Gaussian Naïve Bayes* guna klasifikasi kondisi keuangan UMKM berdasarkan laporan keuangan sederhana dan mengintegrasikannya ke dalam sistem machine learning. Hasil klasifikasi membagi kondisi keuangan UMKM ke dalam dua kategori, yakni Sehat serta Tidak Sehat, berdasarkan enam fitur utama yang digunakan di penelitian. Model yang dibangun memperoleh akurasi sebesar 94,70% dengan nilai *F1-Score* sebesar 94,72% dan ROC-AUC sebesar 0,9859, sehingga menunjukkan performa klasifikasi yang baik dan stabil. Penggunaan dataset sintetis dari *Kaggle* membantu proses pelatihan model dalam jumlah data yang besar, namun model masih memiliki keterbatasan ketika menerima input data di luar rentang dataset pelatihan. Implementasi ini diharapkan bisa membantu pelaku UMKM pada menganalisis kondisi keuangan usaha dengan lebih cepat serta sederhana sebagai dasar pengambilan keputusan bisnis. Untuk penelitian selanjutnya, disarankan menggunakan data riil UMKM, memperluas variasi dataset, serta membandingkan performa Gaussian Naïve Bayes dengan metode machine learning lainnya agar model menjadi lebih adaptif dan robust terhadap data baru.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada semua pihak yang telah memberikan dukungan, bantuan, dan masukan selama proses penelitian serta penyusunan artikel ini. Semoga hasil penelitian ini dapat memberikan manfaat bagi pengembangan ilmu pengetahuan, khususnya dalam penerapan machine learning untuk penerapan machine learning untuk klasifikasi kondisi keuangan UMKM.

REFERENSI

[1] Lisnawati, "Isu Sepekan EDITOR Polhukam," 2024. [Online]. Available: https://puslit.dpr.go.id

[2] R. Desi Safitri, "Peran Financial Technology dalam Meningkatkan Pengelolaan Keuangan UMKM," *Ilmu Ekonomi Manajemen dan Akuntansi*, vol. 5, no. 2, pp. 428–437, Oct. 2024, doi: 10.37012/ileka.v5i2.2352.

[3] A. H. Nareswari and W. Winarsih, "Pengaruh literasi keuangan, sistem informasi akuntansi, adopsi it dan green innovation performance terhadap kinerja keuangan umkm di Jawa Tengah," *Jurnal Riset Ekonomi dan Bisnis*, vol. 17, no. 1, p. 51, Apr. 2024, doi: 10.26623/jreb.v17i1.8553.

[4] H. F. Putro, R. T. Vulandari, and W. L. Y. Saptomo, "Penerapan Metode Naive Bayes Untuk Klasifikasi Pelanggan," *Jurnal Teknologi Informasi dan Komunikasi (TIKomSiN)*, vol. 8, no. 2, Oct. 2020, doi: 10.30646/tikomsin.v8i2.500.

[5] A. Haris, A. Hananto, F. nurapriani Sistem Informasi, U. Buana Perjuangan Karawang Jl Ronggo Waluyo Sinarbaya, T. Timur, and J. Barat, "PENERAPAN ALGORITMA NAIVE BAYES UNTUK KLASIFIKASI SENTIMEN ULASAN PENGGUNA APLIKASI PIDI UMKM DENGAN PEMBOBOTAN FITUR TF-IDF," 2025.

[6] M. Erkamim, S. Suswadi, M. Z. Subarkah, and E. Widarti, "Komparasi Algoritme Random Forest dan XGBoosting dalam Klasifikasi Performa UMKM," *Jurnal Sistem Informasi Bisnis*, vol. 13, no. 2, pp. 127–134, Oct. 2023, doi: 10.21456/vol13iss2pp127-134.

[7] D. A. Huriah and N. D. Nuris, "KLASIFIKASI PENERIMA BANTUAN SOSIAL UMKM MENGGUNAKAN ALGORITMA NAÏVE BAYES," 2023.

[8] S. Ayem *et al.*, "Systematic Literature Review: Implementasi SAK EMKM Pada Penyusunan Laporan Keuangan UMKM di Indonesia," *Jurnal Literasi Akuntansi*, vol. 4, no. 2, pp. 87–99, Jun. 2024, doi: 10.55587/jla.v4i2.118.

[9] Andi Diah Kuswanto, Said Imam Puro, Jodi Hariyan, Ridho Rafliansyah, Muhammad Rival Aziz, and Pebro Vaulina Rajagukguk, "Analisa Data Shopping Trends Menggunakan Algoritma Klasifikasi Dengan Metode Naive Bayes," *Repeater : Publikasi Teknik Informatika dan Jaringan*, vol. 2, no. 3, pp. 119–134, Jul. 2024, doi: 10.62951/repeater.v2i3.118.

[10] Nurul A'ayunnisa, Y. Salim, and H. Azis, "Analisis Performa Metode Gaussian Naïve Bayes untuk Klasifikasi Citra Tulisan Tangan Karakter Arab," *Indonesian Journal of Data and Science*, vol. 3, no. 3, pp. 115–121, 2022, doi: 10.56705/ijodas.v3i3.54.

[11] I. Dwita, S. Tarigan, R. Habibi, R. Nuraini, and S. Fatonah, "Evaluasi Algoritma Klasifikasi Machine Learning Kategori Nilai Akhir Tunjangan Kinerja Pegawai," 2023.