Jurnal Infomedia : Teknik Informatika, Multimedia, dan Jaringan
Volume 10, Number 2, Desember Tahun 2025
ISSN 2527-9858 (online), 2549-1180 (print)
Artikel Penelitian

Penerapan Algoritma Binning pada Preprocessing Data untuk Meningkatkan Akurasi Klasifikasi Multi-Kelas: Studi Kasus Data SDG

Wiradika Nur Fadhillah*[1] Ronny Susetyoko[2] Isbat Uzzin Nadhori[3]
[1] Departemen Teknik Informatika dan Komputer, Politeknik Elektronika Negeri Surabaya, Jl. Raya ITS, Sukolilo, Surabaya 60111, Indonesia, dnurfadh@student.pens.ac.id
[2] Departemen Teknik Informatika dan Komputer, Politeknik Elektronika Negeri Surabaya, Jl. Raya ITS, Sukolilo, Surabaya 60111, Indonesia, rony@pens.ac.id
[3] Departemen Teknik Informatika dan Komputer, Politeknik Elektronika Negeri Surabaya, Jl. Raya ITS, Sukolilo, Surabaya 60111, Indonesia, isbat@pens.ac.id
*Corresponding Author: **wiradikanf@gmail.com**

Abstrak
Klasifikasi data memainkan peran esensial dalam analisis data, terutama untuk data *Sustainable Development Goals* (*SDGs*) yang seringkali memiliki karakteristik kompleks seperti nilai hilang dan distribusi tidak seimbang, sehingga memerlukan tahap *preprocessing* yang efektif. Penelitian ini bertujuan untuk mengevaluasi secara komprehensif efektivitas tiga teknik *binning*, yaitu *Fixed Binning*, *Random Binning*, dan *KNN Binning*, dalam meningkatkan akurasi klasifikasi multikelas pada data *SDGs*. Teknik *binning* ini diimplementasikan dan diuji menggunakan tiga algoritma klasifikasi utama, yaitu *Random Forest*, *Logistic Regression*, dan *Multilayer Perceptron* (*MLP*). Penelitian ini menggunakan dua dataset yang merepresentasikan data *SDGs*, yaitu data pembangunan berkelanjutan dan ketahanan pangan. Dataset tersebut adalah dataset UKT dengan 2.137 entri dan dataset Ketahanan pangan dengan 514 entri. *KNN Binning* dipilih karena kemampuannya mengelompokkan data berdasarkan kedekatan antar instans, adaptif terhadap distribusi data yang kompleks. Hasil penelitian secara konsisten menunjukkan bahwa *KNN Binning* memberikan peningkatan akurasi tertinggi. Secara spesifik, kombinasi *KNN Binning* dengan *Random Forest* menghasilkan akurasi 92.25% pada dataset UKT dan 73.79% pada dataset Ketahanan pangan. Lebih lanjut, kombinasi ini juga menunjukkan peningkatan pada metrik presisi, *recall*, dan F1 score. Temuan ini menggarisbawahi superioritas *KNN Binning* dalam menangani data *SDGs* yang beragam dan tidak merata, sehingga memberikan kontribusi penting bagi pengembangan teknik *preprocessing* yang lebih akurat, andal, dan dapat meningkatkan performa model klasifikasi secara keseluruhan untuk analisis data *SDGs*.
Kata Kunci: *preprocessing* data; klasifikasi multi kelas; *SDGs*; *binning*; *Random Forest*; *Logistic Regression*; *Multi Layer Perceptron*

Abstract
Data classification is essential in data analysis, especially for Sustainable Development Goals (*SDGs*) data, which often have complex characteristics such as missing values and imbalanced distributions, requiring effective preprocessing. This research aims to comprehensively evaluate the effectiveness of three binning techniques—*Fixed Binning*, *Random Binning*, and *KNN Binning*—in improving multi-class classification accuracy on *SDGs* data. These binning techniques were implemented and tested using three main classification algorithms: *Random Forest*, *Logistic Regression*, and Multilayer Perceptron (*MLP*). The study uses two datasets representing *SDGs* data: the sustainable development dataset and the food security dataset. The datasets include the UKT dataset with 2,137 entries and the Food Security dataset with 514 entries. *KNN Binning* was chosen for its ability to group data based on proximity between instances, making it adaptive to complex data distributions. The results consistently show that *KNN Binning* provides the highest accuracy improvement. Specifically, the combination of *KNN Binning* with *Random Forest* achieved an accuracy of 92.25% on the UKT dataset and 73.79% on the Food Security dataset. Furthermore, this combination also showed improvements in *precision*, *recall*, and F1 score metrics. These findings highlight the superiority of *KNN Binning* in handling diverse and uneven *SDGs* data, thus contributing to the development of more accurate and reliable preprocessing techniques that can enhance overall classification model performance for *SDGs* data analysis.
Keywords: data preprocessing; multi-class classification; *SDGs*; binning; *Random Forest*; *Logistic Regression*; Multi Layer Perceptron

PENDAHULUAN
Klasifikasi data merupakan teknik fundamental dalam *machine learning* yang bertujuan untuk mengelompokkan data ke dalam kategori kategori tertentu berdasarkan berbagai fitur relevan. Dalam konteks analisis data modern, kemampuan

untuk melakukan klasifikasi secara akurat menjadi sangat krusial. Keakuratan model klasifikasi ini sendiri sangat bergantung pada kualitas data masukan. Oleh karena itu, *preprocessing* data atau pra pemrosesan data memegang peranan sebagai tahap awal yang esensial dan seringkali menentukan dalam keseluruhan alur kerja analisis data, terutama ketika berhadapan dengan data *Sustainable Development Goals* (*SDGs*). Data *SDGs* dikenal seringkali memiliki struktur yang kompleks dan menghadapi berbagai tantangan inheren. Tantangan tersebut mencakup masalah seperti adanya nilai yang hilang (missing values), distribusi data antar kelas yang tidak seimbang, serta tingginya variasi antar fitur (Aggarwal, 2015; Sugriyono & Siregar, 2020). Berbagai tantangan ini secara langsung dapat menurunkan kemampuan generalisasi dan keandalan model klasifikasi jika tidak ditangani secara tepat melalui strategi *preprocessing* yang komprehensif. Penerapan *preprocessing* yang tepat dan sesuai dengan karakteristik data terbukti dapat secara signifikan meningkatkan akurasi serta robustisitas model klasifikasi dengan memitigasi berbagai masalah tersebut.

Salah satu teknik *preprocessing* yang sering diterapkan dan terbukti efektif untuk menangani atribut data bersifat kontinu adalah *binning*. Teknik ini bekerja dengan cara mengubah data kontinu tersebut menjadi representasi kategori diskrit melalui proses pembagian ke dalam sejumlah interval atau *bin* tertentu. Pendekatan ini tidak hanya menyederhanakan struktur data sehingga lebih mudah diinterpretasi, tetapi juga membantu model klasifikasi dalam mengenali pola pola signifikan yang mungkin tersembunyi di dalam data. Dalam praktiknya, teknik *binning* dapat diimplementasikan melalui berbagai pendekatan metodologis, dimana setiap pendekatan memiliki kelebihan serta keterbatasan spesifik tergantung pada karakteristik data yang dihadapi. Di antara beragam teknik tersebut, *KNN Binning* atau *binning* berbasis *K Nearest Neighbors* menunjukkan potensi besar sebagai metode yang lebih efektif dalam upaya meningkatkan akurasi klasifikasi, khususnya pada dataset *SDGs* yang kompleks. Keunggulan ini didukung oleh temuan riset sebelumnya (Susetyoko, 2023) yang mengindikasikan bahwa *KNN Binning* memiliki kemampuan superior dalam menangani distribusi data yang tidak merata. Hal ini dimungkinkan karena proses pengelompokan data dalam *KNN Binning* didasarkan pada prinsip kedekatan atau similaritas antar instans data dalam ruang fitur, sehingga memungkinkan model untuk menangkap pola pola yang lebih rumit dan nonlinear yang sering tersembunyi dalam data sosial ekonomi dan lingkungan terkait *SDGs*.

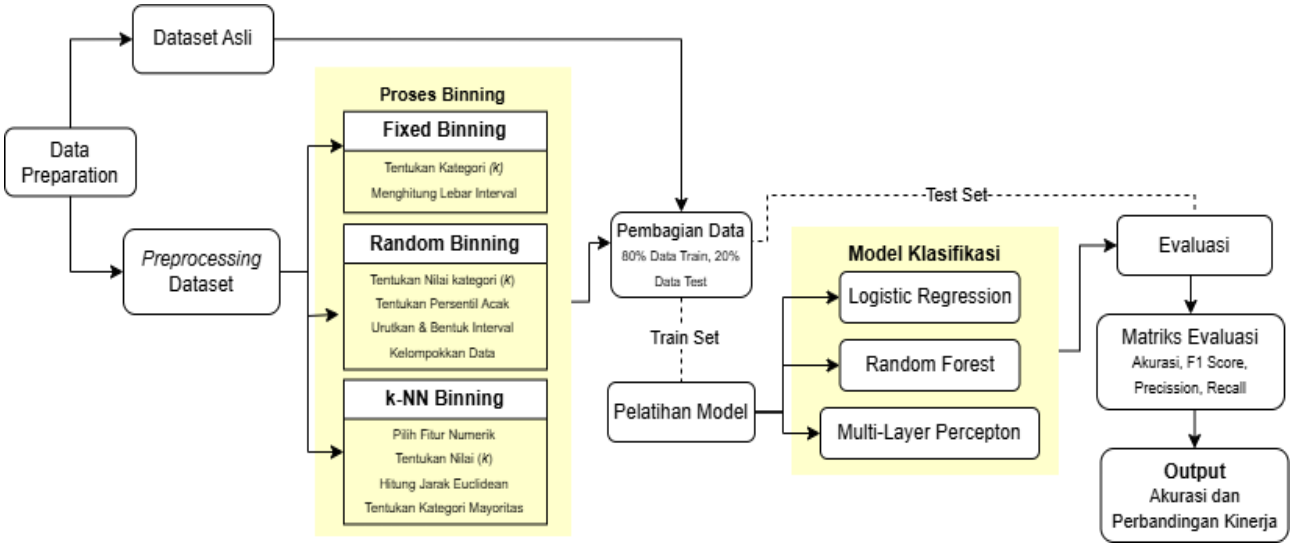
Sementara itu, metode *binning* lain seperti *Fixed Binning* dan *Random Binning* menawarkan pendekatan yang cenderung lebih sederhana namun dengan implikasi performa yang berbeda. *Fixed Binning*, yang membagi data ke dalam interval interval dengan lebar tetap, umumnya menunjukkan efektivitas yang lebih baik pada data dengan distribusi yang relatif seragam dan dapat diprediksi. Di sisi lain, *Random Binning* melakukan pembagian data berdasarkan kriteria acak, sehingga menawarkan fleksibilitas yang lebih tinggi. Meskipun demikian, sifat acaknya tersebut dapat menghasilkan variabilitas performa model yang lebih besar dan terkadang kurang stabil, terutama ketika diterapkan pada data dengan sebaran yang sangat bervariasi atau memiliki banyak pencilan. Penting untuk digarisbawahi bahwa selain pemilihan teknik *binning* yang tepat, jenis algoritma klasifikasi yang digunakan juga akan sangat memengaruhi kinerja dan keandalan model secara keseluruhan, terutama ketika berhadapan dengan keragaman dan kompleksitas data *SDGs*. Beberapa model klasifikasi yang populer dan sering digunakan dalam berbagai penelitian meliputi *Random Forest*, *Logistic Regression*, dan *Multilayer Perceptron* (*MLP*). Algoritma *Random Forest* dikenal memiliki ketahanan yang baik dalam menangani *noise* dan variasi fitur yang umum pada data *SDGs*. *Logistic Regression*, meskipun merupakan model linear, unggul dalam aspek interpretabilitas model yang memudahkan pemahaman terhadap faktor faktor yang berpengaruh. Adapun *MLP*, sebagai representasi dari jaringan saraf tiruan, memiliki kapasitas untuk menangkap pola nonlinear yang sangat kompleks yang mungkin ada dalam interaksi fitur data *SDGs*, namun seringkali memerlukan proses *tuning hyperparameter* yang lebih cermat dan sumber daya komputasi yang lebih besar (Basu & Saha, 2018; Susetyoko, 2023).

Berdasarkan uraian tersebut, penelitian ini memiliki tujuan utama untuk melakukan evaluasi dan perbandingan secara mendalam mengenai efektivitas berbagai teknik *binning* yang telah disebutkan, yaitu *Fixed Binning*, *Random Binning*, dan *KNN Binning*. Fokus evaluasi adalah pada sejauh mana teknik teknik ini mampu meningkatkan akurasi model klasifikasi multikelas. Pengujian efektivitas ini akan dilakukan pada dua dataset utama yang memiliki relevansi dengan pencapaian *SDGs*, yaitu dataset Uang Kuliah Tunggal (UKT) dan dataset Ketahanan pangan. Ketiga algoritma klasifikasi yang telah dibahas sebelumnya juga akan digunakan dalam proses evaluasi ini untuk melihat interaksi antara teknik *binning* dan model klasifikasi. Diharapkan melalui rangkaian eksperimen yang sistematis ini, dapat diidentifikasi pendekatan *preprocessing* menggunakan *binning* yang paling efektif dan optimal. Luaran dari penelitian ini diharapkan tidak hanya meningkatkan akurasi tetapi juga keandalan model klasifikasi, khususnya dalam konteks analisis data *SDGs* yang penuh tantangan.

METODE PENELITIAN

Metodologi penelitian ini disusun melalui beberapa tahapan utama yang sistematis, dimulai dari persiapan data, dilanjutkan dengan pra pemrosesan data menggunakan teknik *binning*, kemudian implementasi berbagai model klasifikasi, dan diakhiri dengan evaluasi performa model. Penelitian ini memfokuskan pada investigasi penggunaan beragam teknik *binning* dengan tujuan untuk meningkatkan akurasi klasifikasi multikelas, khususnya dalam konteks data yang berkaitan dengan *Sustainable Development Goals* (*SDGs*). Proses eksperimen dalam penelitian ini melibatkan dua dataset utama sebagai studi kasus, yaitu Dataset Uang Kuliah Tunggal (UKT) dan Dataset Ketahanan pangan.

Untuk memberikan visualisasi alur kerja penelitian secara keseluruhan, Gambar 1. Diagram tersebut mengilustrasikan secara runtut tahapan proses penelitian, mulai dari tahap masukan data *SDGs*, melalui proses pra pemrosesan data yang menekankan pada penerapan teknik *binning* seperti *Fixed Binning*, *KNN Binning*, dan *Random Binning*. Setelah melalui pra pemrosesan, data kemudian dibagi menjadi data latih dan data uji. Selanjutnya, dilakukan pelatihan model dengan menggunakan tiga algoritma utama yaitu *Random Forest*, *Logistic Regression*, dan *Multilayer Perceptron*. Tahap akhir yang digambarkan adalah proses evaluasi model, yang dilakukan dengan menggunakan berbagai metrik standar untuk memungkinkan perbandingan kinerja antar model secara objektif



Gambar 1. Digram Sistem Penelitian

A. Penyiapan Data

Penelitian ini memanfaatkan dua dataset utama, yaitu dataset Uang Kuliah Tunggal (UKT) dan dataset Ketahanan pangan. Dataset UKT berasal dari Politeknik Elektronika Negeri Surabaya yang secara spesifik menghimpun data sosial ekonomi mahasiswa dan terdiri dari 2.137 entri data. Sedangkan dataset ketahanan pangan mencakup data dari 514 entri yang berasal dari berbagai daerah di Indonesia. Kedua dataset tersebut masing-masing memuat serangkaian atribut yang relevan untuk menggambarkan kondisi sosial ekonomi serta aspek ketahanan pangan pada wilayah studi.

Dataset UKT memuat informasi terkait kondisi sosial ekonomi mahasiswa, yang mencakup atribut-atribut seperti pendapatan, status rumah tangga, serta kepemilikan aset. Informasi mendetail mengenai daftar atribut, tipe data, serta deskripsi untuk setiap atribut dalam dataset UKT dapat ditemukan pada Tabel 1.

Table 1. Daftar Atribut, Tipe, dan Deskripsi Dataset UKT

No	Daftar Atribut	Tipe	Deskripsi
1	Salary	Numerik	Pendapatan per bulan
2	Status_rumah	Kategorikal	Status kepemilikan rumah
3	Rumah	Kategorikal	Kepemilikan rumah (0: Tidak, 1: Memiliki)
4	Motor	Kategorikal	Jumlah motor yang dimiliki
5	Mobil	Kategorikal	Jumlah mobil yang dimiliki
6	Listrik	Kategorikal	Kapasitas listrik rumah tangga
7	Tanah	Kategorikal	Kepemilikan tanah
8	Anak	Kategorikal	Jumlah anak yang dimiliki
9	Klas_ukt	Kategorikal	Klasifikasi UKT (1-8)

Di sisi lain, dataset ketahanan pangan memberikan gambaran mengenai berbagai indikator ketahanan pangan di Indonesia, dengan mencakup informasi seperti tingkat kemiskinan, pengeluaran pangan, akses terhadap listrik dan air bersih, serta kondisi kesehatan masyarakat. Untuk informasi lebih rinci tentang atribut, tipe data, dan deskripsi dari setiap atribut dalam dataset Ketahanan pangan, dapat dilihat pada Tabel 2.

Table 2. Daftar Atribut, Tipe, dan Deskripsi Dataset Pangan

No	Daftar Atribut	Tipe	Deskripsi
1	Kemiskinan	Numerik	Persentase penduduk miskin (%)
2	pengeluaran_pangan	Numerik	Pengeluaran rumah tangga untuk pangan (%)
3	tanpa_listrik	Numerik	Persentase penduduk tanpa akses listrik (%)
4	tanpa_air_bersih	Numerik	Persentase penduduk tanpa akses air bersih (%)
5	lama_sekolah_perempuan	Numerik	Rata-rata lama sekolah perempuan (tahun)
6	rasio_tenaga_kesehatan	Numerik	Jumlah tenaga kesehatan per 100.000 penduduk
7	angka_harapan_hidup	Numerik	Harapan hidup saat lahir (tahun)
8	Stunting	Numerik	Persentase balita yang mengalami stunting (%)
9	klas_ikp	Kategorikal	Klasifikasi ketahanan pangan

B. *Pre-processing*

Pada tahap *preprocessing*, fokus utama adalah menggunakan teknik *binning* untuk mengubah data kontinu menjadi kategori diskrit. Teknik *binning* ini bertujuan untuk mengelompokkan nilai atribut dalam dataset menjadi beberapa kategori yang lebih mudah dipahami dan dianalisis. Proses ini sangat penting dalam klasifikasi, karena dapat meningkatkan akurasi model dengan mengurangi ketidakteraturan dan kompleksitas dalam data. Dalam penelitian ini, tiga metode *binning* yang diterapkan adalah *Fixed Binning*, *Random Binning*, dan *KNN Binning*. Setiap metode ini memiliki pendekatan yang berbeda dalam membagi data kontinu ke dalam berbagai kategori.

Metode Diskretisasi Fitur (Binning)

Dalam konteks *preprocessing* data, teknik diskretisasi fitur, atau *binning*, diterapkan untuk mentransformasi fitur-fitur dengan nilai kontinu menjadi representasi kategorikal diskrit. Transformasi ini bertujuan untuk menyederhanakan kompleksitas data, meningkatkan robustisitas model terhadap *noise*, dan memfasilitasi pengenalan pola oleh algoritma klasifikasi. Tiga metode *binning* dievaluasi dalam penelitian ini yaitu *Fixed Binning*, *Random Binning*, dan *K-Nearest Neighbors (KNN) Binning*.

Fixed Binning

Metode *Fixed Binning* digunakan untuk mengubah atribut kontinu menjadi kategori ordinal dengan membagi data berdasarkan batas persentil tetap yang telah dihitung. Misalnya, jika kita ingin membagi data menjadi empat kategori, batas-batas kategori akan ditentukan pada persentil ke-25, ke-50, dan ke-75. Teknik ini sangat berguna untuk data yang memiliki distribusi hampir merata.

Langkah pertama adalah membaca dataset dan memilih atribut kontinu yang akan diproses. Setelah itu, jumlah kategori yang diinginkan ditentukan, seperti $k = 4$ untuk membagi data menjadi empat kategori. Selanjutnya, dihitung $k - 1$ persentil untuk menentukan batas kategori yang membagi data menjadi interval yang setara.

Setelah batas persentil ditentukan, setiap nilai dalam atribut akan dikelompokkan ke dalam kategori berdasarkan posisinya relatif terhadap batas persentil yang telah dihitung. Hasil kategori baru tersebut kemudian disimpan dalam dataset, menggantikan nilai kontinu dengan kategori ordinal yang sesuai.

***Random Binning* (Diskretisasi Berbasis Persentil Acak)**

Berbeda dengan *Fixed Binning*, *Random Binning* membentuk interval secara acak berdasarkan nilai persentil yang dipilih secara acak dari data. Jumlah kategori yang digunakan ditentukan menggunakan rumus Sturges:

$$k = 1 + 3.3 \log n \quad (2)$$

di mana *n* adalah jumlah total data. Dengan cara ini, batas interval tidak lagi uniform tetapi mengikuti pola distribusi data yang lebih fleksibel. Metode *Random Binning* dalam penelitian ini diterapkan untuk mengevaluasi kemampuan model dalam menghadapi pembagian kategori yang tidak teratur dan dampaknya terhadap performa klasifikasi.

K-Nearest Neighbors (KNN) Binning (Diskretisasi Berbasis Kedekatan KNN)

KNN Binning memanfaatkan prinsip kedekatan antar data dengan menggunakan algoritma K-Nearest Neighbors. Data dikelompokkan berdasarkan mayoritas kategori dari tetangga terdekat yang dihitung menggunakan jarak Euclidean:

1. Pertama adalah pemilihan fitur numerik. Fitur numerik yang akan dikelompokkan, seperti pendapatan atau pengeluaran, ditentukan.
2. Selanjutnya, ditentukan nilai *K*. Nilai ini merepresentasikan jumlah tetangga terdekat yang akan digunakan dalam perhitungan kedekatan antar data, misalnya *K*=3.
1. Tahap berikutnya adalah perhitungan jarak. Jarak Euclidean digunakan untuk menghitung kedekatan antara titik data yang akan diprediksi dan titik data lainnya. Rumus jarak Euclidean antara dua titik $x = (x_1, x_2, \ldots, x_n)$ dan $(y_1, y_2, \ldots, y_n)$ adalah:

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \quad (3)$$

3. Kemudian dilakukan identifikasi *K* tetangga terdekat. *K* tetangga terdekat ditemukan berdasarkan perhitungan jarak Euclidean yang telah dilakukan.
4. Setelah itu, kategori untuk titik data ditentukan berdasarkan kategori mayoritas dari *K* tetangga terdekat.
5. Langkah terakhir adalah menyimpan dataset hasil binning. Dataset yang telah dikelompokkan ke dalam kategori menggunakan KNN disimpan untuk analisis lebih lanjut.

C. Klasifikasi dan Evaluasi

Setelah tahap *Preprocessing* selesai, dataset yang telah diproses dengan teknik binning akan digunakan dalam algoritma klasifikasi untuk memprediksi kelas berdasarkan fitur yang ada. Klasifikasi ini bertujuan untuk memahami pola dalam data dan mengelompokkan entri ke dalam kategori-kategori yang relevan, dengan menggunakan berbagai teknik yang disesuaikan dengan karakteristik data yang berbeda. Setelah proses klasifikasi, evaluasi model dilakukan untuk mengukur seberapa baik model dapat memprediksi data yang belum pernah dilihat sebelumnya.

Pembagian Data

Tahap pertama dalam klasifikasi adalah membagi dataset menjadi dua bagian utama: data latih dan data uji. Teknik train_test_split dari pustaka Scikit-learn digunakan untuk membagi dataset menjadi 80% data latih dan 20% data uji. Pembagian ini bertujuan untuk memastikan bahwa model yang dilatih tidak hanya mempelajari pola dari satu bagian data saja, tetapi juga diuji pada data yang belum pernah dilihat sebelumnya. Dengan demikian, pembagian data yang tepat membantu model agar tidak mengalami overfitting, yaitu kondisi di mana model terlalu menyesuaikan diri dengan data latih sehingga tidak dapat menggeneralisasi dengan baik pada data baru.

Model yang Digunakan

Setelah data dibagi menjadi data latih dan data uji, langkah selanjutnya adalah penerapan berbagai model klasifikasi

untuk memprediksi kelas berdasarkan fitur yang ada dalam dataset. Pada penelitian ini, tiga model yang digunakan adalah *Random Forest*, *Logistic Regression*, dan Multilayer Perceptron (*MLP*). Setiap model ini memiliki keunggulan dan penerapan yang berbeda, tergantung pada kompleksitas dan karakteristik data.

Random Forest telah menjadi pilihan utama dalam analisis data *SDGs* karena kemampuannya menangani data kompleks dan tidak seimbang. Algoritma ini bekerja dengan membangun banyak pohon keputusan secara paralel, kemudian menggabungkan hasilnya melalui voting mayoritas (Aggarwal, 2015). Studi Hulvi & Kusrini (2024) membuktikan efektivitasnya dengan akurasi 89.7% dalam mengklasifikasikan indikator *SDGs* di Indonesia. Keunggulan utamanya terletak pada ketahanan terhadap noise dan kemampuan identifikasi fitur penting tanpa perlu normalisasi data ekstensif. Namun, interpretasi hasilnya relatif lebih kompleks dibanding metode linear.

Logistic Regression menawarkan pendekatan yang lebih sederhana namun powerful untuk analisis *SDGs*, terutama ketika interpretasi hasil menjadi prioritas. Algoritma ini memodelkan hubungan linear antara variabel prediktor dengan probabilitas keanggotaan kelas (Basu & Saha, 2018). Penelitian Susetyoko et al. (2023) menunjukkan peningkatan akurasi 15.2% ketika dikombinasikan dengan *KNN Binning*. Kelebihannya terletak pada kemudahan interpretasi koefisien yang dapat langsung mencerminkan pengaruh masing-masing indikator *SDGs*. Namun, performanya menurun ketika menghadapi hubungan non-linear yang kompleks antar variabel.

Multilayer Perceptron (*MLP*) merupakan solusi untuk menangkap pola kompleks dalam data *SDGs* yang tidak bisa diakomodasi oleh metode linear. Jaringan saraf tiruan ini mampu mempelajari representasi hierarkis dari berbagai indikator pembangunan melalui lapisan tersembunyi (Goodfellow et al., 2016). Implementasi oleh Ainayya et al. (2023) menunjukkan keunggulan *MLP* dalam mengidentifikasi interaksi tersembunyi antar variabel. Namun, algoritma ini memerlukan data dalam jumlah besar dan proses tuning yang intensif. Biaya komputasinya juga lebih tinggi dibanding dua metode sebelumnya, meskipun mampu memberikan hasil yang lebih akurat untuk pola data yang sangat kompleks.

Matriks Evaluasi

Dalam penelitian ini, evaluasi model klasifikasi dilakukan dengan menggunakan beberapa metrik statistik yang bertujuan untuk mengukur performa prediksi model secara objektif dan kuantitatif. Akurasi digunakan untuk mengukur proporsi prediksi yang benar terhadap total observasi, dihitung dengan rumus:

$$\text{Akurasi} = \frac{(\text{TP} + \text{TN})}{(\text{TP} + \text{TN} + \text{FP} + \text{FN})}. \quad (5)$$

Meskipun akurasi memberikan gambaran umum, metrik ini kurang efektif ketika distribusi kelas tidak seimbang. Oleh karena itu, *Precision* dan *Recall* digunakan untuk memberikan informasi lebih rinci. *Precision*, dihitung dengan rumus:

$$\textit{Precision} = \frac{\text{TP}}{(\text{TP} + \text{FP}),} \quad (6)$$

mengukur seberapa banyak prediksi positif yang benar, sementara *Recall*, dihitung dengan rumus:

$$\textit{Recall} = \frac{\text{TP}}{(\text{TP} + \text{FN}),} \quad (7)$$

menunjukkan seberapa banyak kasus positif yang berhasil diidentifikasi oleh model. Untuk memberikan penilaian yang lebih seimbang antara *precision* dan *recall*, digunakan *F1-Score*, yang dihitung dengan rumus:

$$\textit{F1-Score} = 2 * \frac{(\textit{Precision} * \textit{Recall})}{(\textit{Precision} + \textit{Recall}).} \quad (8)$$

F1-Score sangat berguna, terutama ketika distribusi kelas tidak merata, karena memberikan bobot yang setara pada kedua metrik tersebut. Evaluasi dengan menggunakan metrik-metrik ini memberikan wawasan yang lebih menyeluruh mengenai kelebihan dan kelemahan model dalam mendeteksi pola pada dataset yang digunakan dalam penelitian ini.

HASIL DAN PEMBAHASAN

Pada bagian ini, akan disajikan hasil eksperimen yang dilakukan untuk menguji efektivitas teknik binning dalam meningkatkan akurasi klasifikasi pada dua dataset yang berbeda, yaitu dataset UKT dan dataset ketahanan pangan. Evaluasi dilakukan dengan menguji berbagai kombinasi antara teknik binning dan model klasifikasi untuk memperoleh hasil yang optimal. Eksperimen ini bertujuan untuk menganalisis pengaruh teknik binning terhadap performa model klasifikasi, baik dari segi akurasi, *precision*, *recall*, dan *F1-Score*, serta untuk melihat bagaimana teknik-teknik ini bekerja pada dua jenis dataset yang memiliki karakteristik yang berbeda.

A. Hasil Evaluasi Model pada Dataset UKT

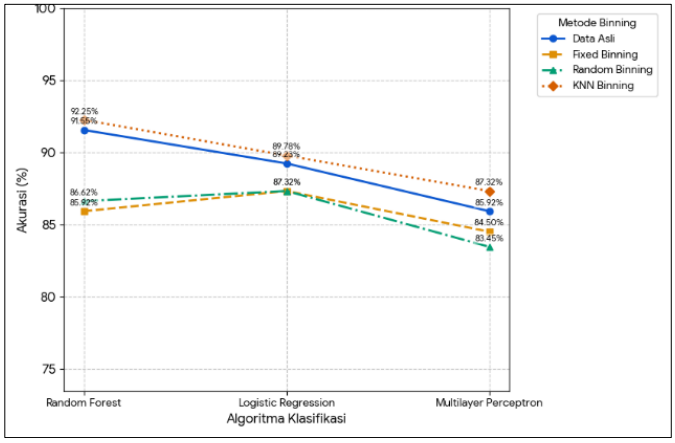
Eksperimen pertama dilakukan pada dataset UKT untuk mengukur seberapa besar pengaruh teknik binning terhadap peningkatan akurasi klasifikasi. Dataset UKT ini terdiri dari data mengenai status pembayaran biaya kuliah yang digunakan untuk mengklasifikasikan status keuangan mahasiswa. Tiga teknik binning yang diuji adalah *Fixed Binning*, *Random Binning*, dan *KNN Binning*, yang dipadukan dengan tiga model klasifikasi yang berbeda, yaitu: *Random Forest*, *Logistic Regression*, dan *Multi Layer Perceptron (MLP)*.

Hasil evaluasi menunjukkan bahwa setiap teknik binning memberikan pengaruh yang berbeda terhadap performa model. Hasil evaluasi akurasi untuk setiap kombinasi antara teknik binning dan model klasifikasi pada dataset UKT dapat dilihat pada Tabel 3 di bawah ini.

Tabel 3. Hasil Akurasi pada Dataset UKT

Metode Binning	*Random Forest*	*Logistic Regression*	Multi Layer Perceptron
Data Asli	91.55%	89.23%	85.92%
Fixed Binning	85.92%	87.32%	84.50%
Random Binning	86.62%	87.32%	83.45%
KNN Binning	92.25%	89.78%	87.32%

Dari tabel tersebut, dapat dilihat bahwa *KNN Binning* memberikan hasil terbaik pada model *Random Forest*, dengan akurasi tertinggi mencapai 92.25%. Ini menunjukkan bahwa teknik *KNN Binning* lebih efektif dalam meningkatkan akurasi klasifikasi, terutama pada model *Random Forest* yang sangat cocok untuk menangani dataset besar dan bervariasi. Berikut ini disajikan Gambar 2 yang menunjukkan perbandingan akurasi antara berbagai metode binning pada dataset UKT.



Gambar 2. Perbandingan Akurasi antara Metode Binning pada Dataset UKT

B. Hasil Evaluasi Model pada Dataset Ketahanan Pangan

Selanjutnya, eksperimen yang serupa dilakukan pada dataset ketahanan pangan untuk mengevaluasi seberapa efektif teknik binning dalam klasifikasi data ketahanan pangan. Dataset ini berfokus pada data ketahanan pangan yang berisi informasi mengenai keberagaman pangan, kecukupan gizi, dan status ketahanan pangan suatu wilayah. Seperti pada dataset UKT, tiga teknik binning yang diuji adalah *Fixed Binning*, *Random Binning*, dan *KNN Binning*, yang dipadukan dengan tiga model klasifikasi yang sama: *Random Forest*, *Logistic Regression*, dan *Multi Layer Perceptron (MLP)*.

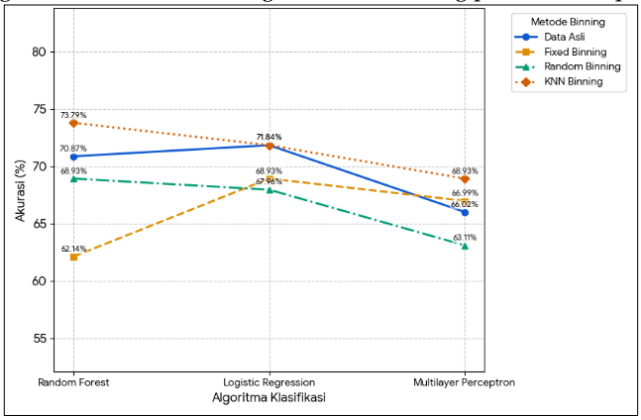
Hasil eksperimen menunjukkan bahwa *KNN Binning* kembali memberikan peningkatan performa yang signifikan, terutama pada model *Random Forest*, yang mampu menangani distribusi data yang kompleks dengan lebih baik. Tabel berikut menyajikan perbandingan akurasi antar teknik binning untuk ketiga model klasifikasi pada dataset ketahanan pangan.

Tabel 4. Hasil Akurasi pada Dataset Pangan

Metode Binning	Random Forest	Logistic Regression	Multi Layer Perceptron
Data Asli	70.87%	71.84%	66.02%
Fixed Binning	62.14%	68.93%	66.99%
Random Binning	68.93%	67.96%	63.11%
KNN Binning	73.79%	71.84%	68.93%

Pada dataset Ketahanan pangan, *KNN Binning* menunjukkan peningkatan yang cukup signifikan, dengan akurasi pada model *Random Forest* meningkat dari 70.87% menjadi 73.79%. Ini menunjukkan bahwa *KNN Binning* efektif dalam meningkatkan akurasi, terutama pada dataset yang lebih kompleks dengan distribusi data yang lebih variatif.

Peningkatan yang cukup besar ini menunjukkan bahwa teknik *KNN Binning* mampu menangani data yang lebih bervariasi, yang sering kali menjadi tantangan dalam klasifikasi data ketahanan pangan. Berikut ini disajikan Gambar 3 yang menunjukkan perbandingan akurasi antara berbagai metode binning pada dataset pangan



Gambar 3. Perbandingan Akurasi antara Metode Binning pada Dataset Pangan

C. Pembahasan Perbandingan Teknik Binning

Dari hasil eksperimen yang dilakukan, dapat disimpulkan bahwa *KNN Binning* adalah teknik binning yang paling efektif dalam meningkatkan akurasi klasifikasi pada kedua dataset, baik pada dataset UKT maupun Ketahanan pangan.

KNN Binning memiliki kemampuan untuk mengelompokkan data berdasarkan kedekatannya dengan data lain yang memiliki fitur serupa, yang memungkinkan model untuk menangkap pola-pola yang lebih kompleks. Hal ini sangat penting dalam menangani data yang memiliki distribusi tidak teratur atau sangat bervariasi, yang sering dijumpai pada dataset yang lebih besar dan kompleks.

Fixed Binning, meskipun lebih mudah diterapkan dan lebih sederhana, memberikan hasil yang lebih rendah. Hal ini disebabkan oleh ketidakfleksibelannya dalam menangani distribusi data yang tidak teratur. Pembagian data menjadi interval yang tetap dapat menyebabkan hilangnya informasi penting yang dapat digunakan dalam proses klasifikasi. Oleh karena itu, *Fixed Binning* lebih cocok digunakan pada data dengan distribusi yang relatif lebih terstruktur.

Random Binning memberikan hasil yang lebih baik dibandingkan dengan *Fixed Binning*, namun masih tertinggal dibandingkan dengan *KNN Binning*. Meskipun teknik ini lebih fleksibel karena pembagian interval dilakukan secara acak, namun proses ini dapat menghasilkan pembagian data yang tidak stabil. Pembagian yang tidak konsisten ini dapat mengurangi efektivitas model dalam mengenali pola-pola penting dalam data, terutama pada dataset yang memiliki distribusi data yang sangat bervariasi.

D. Evaluasi Metrik Klasifikasi

Untuk memberikan gambaran yang lebih komprehensif tentang kualitas klasifikasi yang dihasilkan oleh model-model yang diuji, evaluasi lebih mendalam dilakukan dengan menggunakan metrik-metrik klasifikasi tambahan seperti *precision*, *recall*, dan *F1-Score*. Metrik-metrik ini memberikan informasi lebih jauh mengenai tidak hanya akurasi model, tetapi juga bagaimana model menangani kesalahan prediksi, baik dalam hal *false positives* (prediksi positif yang salah) maupun *false negatives* (prediksi negatif yang salah). Oleh karena itu, metrik ini sangat penting untuk menilai ketepatan dan konsistensi prediksi model dalam konteks klasifikasi, terutama ketika data yang digunakan memiliki distribusi yang kompleks dan variatif.

Tabel berikut menyajikan perbandingan hasil evaluasi metrik untuk setiap teknik binning yang diterapkan pada dataset UKT. Dalam tabel ini, kami menyajikan metrik *precision*, *recall*, dan *F1-Score* untuk model *Random Forest*. Metrik ini dihitung untuk setiap teknik binning yang digunakan dalam eksperimen ini, yaitu Data Asli, *Fixed Binning*, *Random Binning*, dan *KNN Binning*.

Tabel 5. Evaluasi Metrik Klasifikasi pada Dataset UKT

Metode Binning	Model	*Precision*	*Recall*	*F1-Score*
Data Asli	*Random Forest*	0.9022	0.9155	0.9088
Fixed Binning	*Random Forest*	0.8271	0.8592	0.8429
Random Binning	*Random Forest*	0.8471	0.8662	0.8566
KNN Binning	*Random Forest*	0.9154	0.9225	0.9189

Berdasarkan tabel di atas, dapat dilihat bahwa *KNN Binning* memberikan peningkatan pada semua metrik evaluasi, terutama pada model *Random Forest*. Model ini mencapai *precision* sebesar 91.54%, *recall* sebesar 92.25%, dan *F1-Score* sebesar 91.89%. Peningkatan ini menunjukkan bahwa *KNN Binning* tidak hanya meningkatkan akurasi model, tetapi juga meningkatkan kualitas prediksi secara keseluruhan. *Precision* yang tinggi menunjukkan bahwa model ini sangat tepat dalam menghasilkan prediksi yang benar-benar positif, sementara *recall* yang tinggi menunjukkan kemampuan model dalam mengenali sebagian besar kasus positif. Secara keseluruhan, *F1-Score* yang tinggi menandakan keseimbangan yang baik antara *precision* dan *recall*, yang sangat penting dalam aplikasi dunia nyata.

Selain itu, perbandingan hasil evaluasi untuk teknik-teknik binning lainnya seperti *Fixed Binning* dan *Random Binning* juga menunjukkan hasil yang signifikan, meskipun tidak sebaik *KNN Binning*. *Fixed Binning* menunjukkan kinerja yang lebih rendah, terutama dalam *precision* dan *recall*, yang mengindikasikan bahwa teknik ini kurang efektif dalam menangani data yang lebih bervariasi. *Random Binning*, meskipun lebih fleksibel dibandingkan *Fixed Binning*, masih tidak memberikan hasil yang optimal jika dibandingkan dengan *KNN Binning*. Hal ini menunjukkan bahwa teknik binning yang lebih canggih, seperti *KNN Binning*, mampu menangani data dengan distribusi yang lebih kompleks dan memberikan hasil yang lebih stabil.

Selanjutnya, evaluasi yang serupa juga dilakukan pada dataset Ketahanan pangan, yang berfokus pada data yang lebih berkaitan dengan ketahanan pangan di berbagai daerah di Indonesia. Hasil dari evaluasi metrik klasifikasi ini memberikan wawasan tambahan tentang bagaimana teknik binning dapat membantu dalam mengklasifikasikan data yang memiliki tantangan lebih besar dalam hal variasi dan distribusi.

Tabel 6. Evaluasi Metrik Klasifikasi pada Dataset Ketahanan pangan

Metode Binning	Model	*Precision*	*Recall*	*F1-Score*
Data Asli	*Random Forest*	0.6239	0.7087	0.6603
Fixed Binning	*Random Forest*	0.5357	0.6214	0.5663
Random Binning	*Random Forest*	0.5963	0.6893	0.6303
KNN Binning	*Random Forest*	0.6854	0.7379	0.7088

Pada dataset Ketahanan pangan, teknik *KNN Binning* kembali menunjukkan hasil yang signifikan pada model *Random Forest*, dengan *precision* sebesar 68.54%, *recall* sebesar 73.79%, dan *F1-Score* sebesar 70.88%. Peningkatan yang signifikan ini menunjukkan bahwa *KNN Binning* sangat efektif dalam mengatasi data yang lebih kompleks dan variatif, sehingga memungkinkan model untuk membuat prediksi yang lebih akurat dan relevan dalam konteks ketahanan

pangan. Peningkatan pada *recall* dan *F1-Score* juga mengindikasikan bahwa model ini lebih baik dalam mengidentifikasi dan mengklasifikasikan data yang berhubungan dengan ketahanan pangan, yang sangat penting dalam konteks analisis sosial-ekonomi dan ketahanan pangan di Indonesia.

Dengan hasil yang diperoleh dari kedua dataset, dapat disimpulkan bahwa *KNN Binning* adalah teknik binning yang paling efektif untuk meningkatkan performa model klasifikasi, terutama ketika digunakan dengan model *Random Forest*. Peningkatan yang konsisten dalam semua metrik evaluasi pada kedua dataset menunjukkan bahwa teknik ini mampu menangani variasi dan distribusi data yang lebih kompleks, sehingga memberikan hasil yang lebih baik dan lebih stabil dibandingkan teknik binning lainnya.

KESIMPULAN

Berdasarkan hasil eksperimen yang dilakukan pada dua dataset yang berbeda, yaitu dataset UKT dan dataset Ketahanan pangan, dapat disimpulkan bahwa *KNN Binning* merupakan teknik binning yang paling efektif dalam meningkatkan performa model klasifikasi. Penggunaan teknik ini, terutama dengan *Random Forest*, menunjukkan peningkatan akurasi yang signifikan dibandingkan dengan teknik binning lainnya, yaitu *Fixed Binning* dan *Random Binning*.

Pada dataset UKT, kombinasi *KNN Binning* dan *Random Forest* berhasil mencapai akurasi tertinggi sebesar 92.25%, sedangkan pada dataset Ketahanan pangan, teknik yang sama meningkatkan akurasi menjadi 73.79%. Selain itu, teknik *KNN Binning* juga memberikan hasil yang lebih baik dalam hal *precision*, *recall*, dan *F1-Score*, yang menunjukkan kemampuan model untuk menangani data yang lebih kompleks dan variatif.

Dengan demikian, dapat disarankan bahwa teknik *KNN Binning* adalah pilihan terbaik untuk meningkatkan performa klasifikasi, terutama ketika digunakan dengan model *Random Forest*. Teknik ini mampu menangani data yang memiliki distribusi yang lebih tidak teratur dan kompleks, serta memberikan hasil yang optimal dalam pengklasifikasian data pada kedua jenis dataset yang digunakan.

REFERENSI

Aggarwal, C. C. (2015). Data mining: The textbook. Springer. https://doi.org/10.1007/978-3-319-14142-8

Ainayya, A., Prasetyo, D. E., & Rahmawati, R. D. (2023). Penerapan data transformation pada database sistem informasi manajemen rumah sakit. Sintak, 8(1), 45–55. https://www.unisbank.ac.id/ojs/index.php/sintak/article/view/7595

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press. http://www.deeplearningbook.org

Hulvi, A., & Kusrini. (2024). Optimasi rekomendasi Sustainable Development Goals (SDGs) di Indonesia menggunakan content-based filtering dan algoritma machine learning. Building of Informatics, Technology and Science (BITS), 6(2), 1045–1058. https://ejurnal.seminar-id.com/index.php/bits/article/view/5807

Junaedi, H., Budianto, H., Maryati, I., & Melani, Y. (2011). Data transformation pada data mining. Prosiding Konferensi Nasional Inovasi dalam Desain dan Teknologi, 7, 93–99.

Nguyen, T. T., et al. (2023). Application of AI/ML techniques in achieving SDGs: a systematic bibliometric review. Environment, Development and Sustainability. https://doi.org/10.1007/s10668-023-03935-1

Saha, Sanjit Kumar. (2025). A comparative analysis of logistic regression and random forest for individual fairness in machine learning. International Journal of Advanced Engineering Research and Science, 12(5), 33–38. https://dx.doi.org/10.22161/ijaers.125.5

Sugriyono, S., & Siregar, M. U. (2020). Prapemrosesan klasifikasi algoritme kNN menggunakan K-means dan matriks jarak untuk dataset hasil studi mahasiswa. Jurnal Teknologi dan Sistem Komputer, 8(4), 311–316. https://doi.org/10.14710/jtsiskom.2020.13874

Susetyoko, R., Purwantini, E., Iman, B. N., & Satriyanto, E. (2023). An improved accuracy of multiclass random forest classifier with continuous attribute transformation using random percentile generation. International Journal on Advanced Science, Engineering and Information Technology, 13(3), 943–953. https://doi.org/10.18517/ijaseit.13.3.18379