P- ISSN: 2355-6684, E-ISSN: 2776-6640
https://ojs.unpkediri.ac.id/index.php/noe

Pengaruh Teknik Preprocessing terhadap Kinerja Model *Explainable Boosting Machine* (EBM) untuk Prediksi Serangan Jantung

Moch. Andri Setiawan[*1], Moh. Hasan Efendi[2], Muhammad Farizal Akbar[3], Wildan Septian Pratama[4]
Universitas Nusantara PGRI Kediri[1,2,3,4]
mochandrisetiawan2@gmail.com[1], hasanefendi1258@gmail.com[2], muhammadfarizalakbar@gmail.com[3], wseptama@gmail.com[4]
Corresponding author: Moch. Andri Setiawan

Abstrak

Serangan jantung merupakan penyakit kardiovaskular yang sering terjadi secara tiba-tiba dan menjadi salah satu penyebab kematian tertinggi. Deteksi dini terhadap risiko serangan jantung masih menjadi tantangan karena keterbatasan sistem prediksi yang akurat dan mudah dipahami. Oleh karena itu, penelitian ini penting dilakukan untuk menghasilkan model prediksi yang tidak hanya akurat, tetapi juga interpretatif. Penelitian ini bertujuan mengembangkan model prediksi risiko serangan jantung berbasis machine learning menggunakan algoritma Explainable Boosting Machine (EBM). *Proses dilakukan dengan pendekatan CRISP-DM serta optimasi pada tahap preprocessing, khususnya penanganan missing value melalui pemetaan data dan penanganan ketidakseimbangan data menggunakan metode SMOTE-ENN. Dataset yang digunakan berasal dari Kaggle, terdiri atas 158.355 baris dan 28 atribut yang mencerminkan faktor demografi, gaya hidup, lingkungan, serta kondisi klinis. Penelitian mencakup lima eksperimen berdasarkan variasi parameter SMOTE dan ENN. Hasil menunjukkan bahwa eksperimen SMOTEENN Sharp (SENS) menghasilkan akurasi tertinggi sebesar 74%, namun mengalami ketidakseimbangan klasifikasi pada kelas berisiko. Sementara itu, eksperimen SMOTEENN Aggressive (SENA) meningkatkan recall pada kelas berisiko, namun menurunkan akurasi menjadi 67%.Temuan ini menunjukkan bahwa strategi penanganan data yang optimal pada tahap preprocessing sangat berpengaruh terhadap kemampuan model dalam mengenali risiko serangan jantung secara lebih akurat dan seimbang.*

Kata Kunci : *Explainable Boosting Machine, ketidakseimbangan data, missing value, preprocessing, SMOTE-ENN*

A. PENDAHULUAN

Penyakit kardiovaskular (PKV), khususnya serangan jantung, merupakan salah satu penyebab utama kematian di Indonesia dan dunia (Azizah, 2023). Di Indonesia, prevalensi penyakit ini menunjukkan tren yang meningkat setiap tahunnya (Kementerian Kesehatan Republik Indonesia, 2022). Serangan jantung sering kali terjadi secara mendadak dan tanpa gejala awal yang jelas, sehingga banyak kasus tidak terdeteksi hingga mencapai tahap yang mengancam jiwa. Keterlambatan dalam deteksi dini, minimnya kesadaran masyarakat, serta keterbatasan akses terhadap layanan kesehatan preventif menjadi permasalahan utama yang harus segera diatasi (Azizah, 2023).

Seiring dengan berkembangnya teknologi digital dan ketersediaan data kesehatan dalam jumlah besar (*big data*), peluang untuk mengembangkan sistem prediksi risiko serangan jantung yang efisien dan terjangkau menjadi semakin terbuka. Salah satu solusi potensial adalah penerapan algoritma pembelajaran mesin yang dapat memproses data kompleks secara cepat dan akurat. Dalam konteks ini, *Explainable Boosting Machine* (EBM) muncul sebagai metode unggulan yang mampu memberikan prediksi dengan tingkat interpretabilitas tinggi (Arslan dkk., 2024: 2).

Melihat urgensi permasalahan tersebut, penelitian ini bertujuan untuk mengembangkan dan mengoptimalkan model prediksi risiko serangan jantung menggunakan algoritma *Explainable Boosting Machine* (EBM) yang dikombinasikan dengan teknik *preprocessing*, seperti penanganan *missing value* dan *SMOTE-ENN* (*Synthetic Minority Over-sampling Technique with Edited Nearest Neighbor*), untuk mengatasi ketidakseimbangan data. Fokus utama penelitian ini adalah meningkatkan performa serta kemampuan model dalam mengidentifikasi individu dengan risiko tinggi secara lebih andal dan dapat dijelaskan.

Secara teoritik, EBM merupakan bentuk dari *Generalized Additive Models* (GAM) yang diperluas dengan teknik *boosting*. Keunggulan EBM terletak pada kemampuannya membangun model

prediksi yang kuat namun tetap dapat dijelaskan secara transparan kepada pengguna non-teknis, seperti tenaga medis atau pembuat kebijakan. Sementara itu, *SMOTE-ENN* digunakan untuk menghasilkan distribusi data yang lebih seimbang antara kelas mayoritas dan minoritas, yang sangat krusial dalam konteks prediksi penyakit yang kasusnya jarang terdeteksi. Studi oleh Alahmadi dkk. (2023) menunjukkan bahwa EBM dapat mencapai performa tinggi sekaligus menjaga interpretabilitas, yang menjadi keunggulan utama dibanding model *black-box* lainnya.

Hasil penelitian ini diharapkan dapat memberikan kontribusi dalam bidang kesehatan preventif, khususnya dalam pengembangan sistem pendukung keputusan berbasis data untuk mendeteksi risiko serangan jantung. Selain itu, penelitian ini juga berpotensi menjadi referensi dalam pengembangan teknologi prediktif yang adaptif dan kontekstual terhadap data populasi Indonesia. Manfaat yang diharapkan mencakup peningkatan kesadaran risiko secara individual, dukungan bagi tenaga kesehatan dalam pengambilan keputusan dini, serta penguatan kebijakan kesehatan masyarakat berbasis data.

B. LANDASAN TEORI

Heart Attack (Serangan Jantung)

Serangan jantung (*myocardial infarction*) adalah kondisi medis yang terjadi ketika aliran darah ke otot jantung terhambat, umumnya akibat penyumbatan arteri koroner. Hal ini menyebabkan kerusakan jaringan jantung yang dapat berakibat fatal (Putri dkk., 2022: 144). Menurut Thygesen dkk. (2018), diagnosis serangan jantung dapat dilakukan berdasarkan gejala klinis, biomarker jantung (seperti *troponin* dan *CK-MB*), serta hasil pemeriksaan rekam jantung (*electrocardiogram*).

2. *Machine Learning*

Machine Learning adalah bagian dari ilmu komputer yang bertujuan mengembangkan sistem yang dapat belajar secara mandiri tanpa perlu diprogram secara berulang oleh manusia. Untuk dapat melakukan pembelajaran, sistem ini membutuhkan data awal sebagai dasar pemrosesan dan analisis perilaku objek. *Machine Learning* sendiri merupakan penerapan dari cabang *Artificial Intelligence* (AI) yang memanfaatkan teknik statistika guna membangun model secara otomatis berdasarkan sekumpulan data, sehingga komputer memperoleh kemampuan untuk belajar dari data tersebut (Roihan dkk., 2020: 76)

3. *Explainable Boosting Machine*

Explainable Boosting Machine (EBM) adalah algoritma *machine learning* berbasis *Generalized Additive Models (GAM)* yang dirancang untuk menghasilkan model prediktif yang akurat sekaligus mudah dipahami. EBM bekerja dengan cara menggabungkan kontribusi masing-masing fitur secara aditif melalui teknik *boosting*, sehingga pengaruh setiap fitur terhadap hasil prediksi dapat divisualisasikan dan diinterpretasikan secara terpisah. Algoritma ini cocok digunakan dalam bidang yang membutuhkan transparansi, seperti kesehatan dan keuangan, serta tahan terhadap *overfitting* karena memanfaatkan proses *binning* dan regularisasi (Arslan dkk., 2024: 3).

4. Mapping Data

Mapping data (pemetaan data) merupakan proses transformasi nilai dalam dataset agar sesuai dengan kebutuhan pemodelan, seperti mengubah data kategorikal menjadi bentuk numerik atau menstandarkan format nilai. Langkah ini penting agar data dapat dipahami dan diolah dengan baik oleh algoritma *machine learning*. Mapping juga membantu menjaga konsistensi data, meminimalkan kesalahan representasi, dan meningkatkan efektivitas proses analisis (Bunte dkk., 2018: 430).

5. Penanganan *Imbalaced Data* (*SMOTE-ENN*)

SMOTE-ENN adalah metode penanganan data tidak seimbang yang menggabungkan dua teknik, yaitu *SMOTE* (*Synthetic Minority Over-sampling Technique*) dan *ENN* (*Edited Nearest Neighbor*). *SMOTE* menambah jumlah data pada kelas minoritas dengan membuat data sintetis berdasarkan tetangga terdekat, sedangkan *ENN* menghapus data yang dianggap *noise* atau tidak konsisten berdasarkan klasifikasi tetangga terdekat. Kombinasi ini tidak hanya meningkatkan proporsi kelas minoritas, tetapi juga menjaga kualitas data, sehingga model menjadi lebih seimbang dan akurat dalam mengenali semua kelas (Pamungkas dkk., 2025: 185).

6. *Confusion Matrix*

Confusion matrix digunakan sebagai salah satu metode untuk mengevaluasi kinerja model prediksi yang telah dibangun (Saputro dkk., 2024: 68). Nilai-nilai dalam *confusion matrix* dapat dilihat pada Tabel 1.

P- ISSN: 2355-6684, E-ISSN: 2776-6640
https://ojs.unpkediri.ac.id/index.php/noe

Tabel 1. *Confusion Matrix*

		Prediksi	
		Positif	Negatif
Aktual	Positif	TP	FN
	Negatif	FP	TN

Selanjutnya, perhitungan metrik evaluasi seperti *accuracy*, *precision*, *recall*, dan *F1-score* disajikan dalam Persamaan (1), (2), (3), dan (4).

$$\text{Akurasi} = \frac{TP+TN}{TP+TN+FP+FN} \times 100\% \quad \text{.....(1)}$$

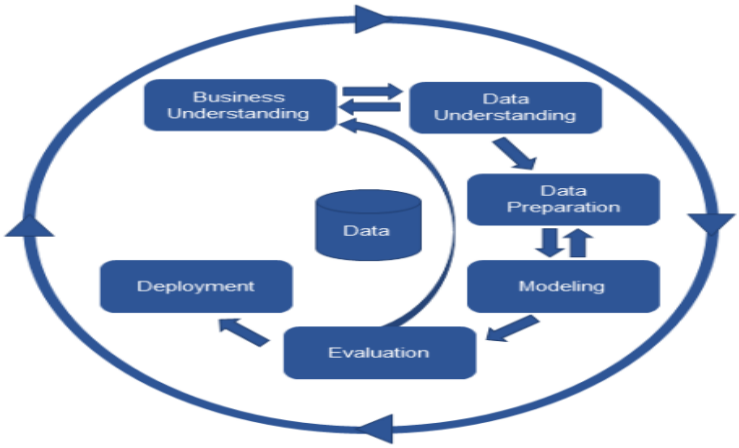
$$\text{Presisi} = \frac{TP}{TP+FP} \times 100\% \quad \text{.....(2)}$$

$$\textit{Recall} = \frac{TP}{TP+FN} \times 100\% \quad \text{.....(3)}$$

$$\textit{F1-Score} = \frac{2 \times recall \times precision}{recall + precision} \quad \text{.....(4)}$$

C. METODE PENELITIAN

Metode penelitian ini menggunakan model *CRISP-DM* (*CRoss-Industry Standard Process for Data Mining*), yaitu model yang bersifat iteratif dan interaktif untuk menemukan pengetahuan dari data (Ristyawan dkk., 2025: 31). Proses ini terdiri dari enam langkah utama dan divisualisasikan pada Gambar 1.



Gambar 1. Model CRISP-DM

1. Studi Literatur

Pada tahap ini, penulis mempelajari teori dari jurnal-jurnal terdahulu, buku, serta informasi yang berkaitan dengan metode klasifikasi *Explainable Boosting Machine* (EBM).

2. *Business Understanding*

Tahap ini fokus pada pemahaman tujuan dan kebutuhan proyek dari sudut pandang bisnis dan kesehatan, kemudian mengubah pemahaman tersebut menjadi definisi masalah data mining serta rencana awal proyek untuk mencapai tujuan tersebut (Ristyawan dkk., 2025: 31).

Tujuan dari penelitian ini adalah untuk menentukan pendekatan *preprocessing* yang paling optimal, guna meningkatkan performa algoritma *Explainable Boosting Machine* (EBM) dalam memprediksi risiko serangan jantung secara lebih akurat.

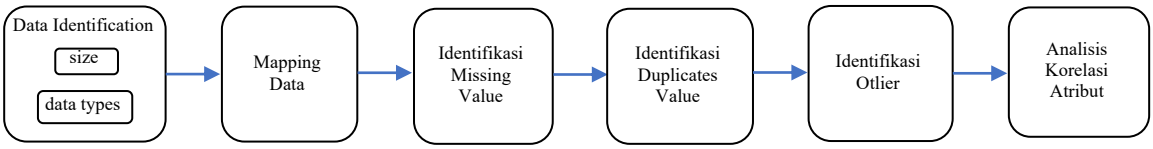
3. *Data Understanding*

Tahap ini mencakup proses pengumpulan data dan eksplorasi awal untuk memahami karakteristik data, mengidentifikasi masalah kualitas, serta menggali wawasan awal (Ristyawan dkk., 2025: 31). Kegiatan ini sejalan dengan konsep *Exploratory Data Analysis* (EDA), yang bertujuan meninjau struktur data, menemukan anomali, mengenali pola tersembunyi, serta mengamati hubungan antar variabel tanpa langsung melakukan pemodelan statistik atau analisis inferensial (Siambaton dan Husein, 2022: 42).

Dalam penelitian ini, *dataset* yang digunakan merupakan kumpulan data profil kesehatan individu di Indonesia yang berfokus pada prediksi risiko serangan jantung. *Dataset* ini mencakup berbagai atribut yang merepresentasikan faktor demografi, klinis, gaya hidup, lingkungan, serta sistem kesehatan yang relevan terhadap risiko penyakit kardiovaskular (Apollo, 2025). Data disusun berdasarkan kondisi dunia nyata di Indonesia, yang didapat dari *Kaggle* : *Heart Attack Dataset*. Penjelasan rinci mengenai setiap atribut dan label target disajikan pada Tabel 2.

Tabel 2. Atribut Pada Dataset

Atribut	Keterangan	Deskripsi
age	Atribut (*Input*)	Usia individu (25–90 tahun)
gender	Atribut (*Input*)	Jenis kelamin individu (Pria, Wanita)
region	Atribut (*Input*)	Area tempat tinggal (Perkotaan, Pedesaan)
income_level	Atribut (*Input*)	Status sosial ekonomi (Rendah, Menengah, Tinggi)
hipertensi	Atribut (*Input*)	Tekanan darah tinggi (1 = Ya, 0 = Tidak)
diabetes	Atribut (*Input*)	Diabetes yang *terdiagnosis* (1 = Ya, 0 = Tidak)
kadar_kolesterol	Atribut (*Input*)	Kadar kolesterol total (*mg/dL*)
obesitas	Atribut (*Input*)	*BMI* > 30 (1 = Ya, 0 = Tidak)
lingkar_pinggang	Atribut (*Input*)	Lingkar pinggang dalam *cm*
riwayat_keluarga	Atribut (*Input*)	Riwayat penyakit jantung dalam keluarga (1 = Ya, 0 = Tidak)
smoking_status	Atribut (*Input*)	Kebiasaan merokok (Tidak Pernah, Dulu, Sekarang)
alcohol_consumption	Atribut (*Input*)	Konsumsi alkohol (Tidak Ada, Sedang, Tinggi)
physical_activity	Atribut (*Input*)	Tingkat aktivitas fisik (Rendah, Sedang, Tinggi)
diet_habits	Atribut (*Input*)	Kualitas diet (Sehat, Tidak Sehat)
air_pollution_exposure	Atribut (*Input*)	Paparan polusi udara (Rendah, Sedang, Tinggi)
stress_level	Atribut (*Input*)	Tingkat stres (Rendah, Sedang, Tinggi)
sleep_hours	Atribut (*Input*)	Rata-rata jam tidur per malam (3–9 jam)
blood_pressure_systolic	Atribut (*Input*)	Tekanan darah *sistolik* (*mmHg*)
blood_pressure_diastolic	Atribut (*Input*)	Tekanan darah *diastolik* (*mmHg*)
fasting_blood_sugar	Atribut (*Input*)	Kadar gula darah puasa (*mg/dL*)
cholesterol_hdl	Atribut (*Input*)	Kadar kolesterol *HDL* (*mg/dL*)
cholesterol_ldl	Atribut (*Input*)	Kadar kolesterol *LDL* (*mg/dL*)
triglycerides	Atribut (*Input*)	Kadar *trigliserida* (*mg/dL*)
EKG_results	Atribut (*Input*)	Hasil *elektrokardiogram* (Normal, Abnormal)
previous_heart_disease	Atribut (*Input*)	Riwayat penyakit jantung sebelumnya (1 = Ya, 0 = Tidak)
medication_usage	Atribut (*Input*)	Konsumsi obat jantung saat ini (1 = Ya, 0 = Tidak)
participating_in_free_screening	Atribut (*Input*)	Mengikuti program *skrining* gratis (1 = Ya, 0 = Tidak)
heart_attack	Label (*Output*)	Kejadian serangan jantung (1 = Ya, 0 = Tidak)



Gambar 2. Tahap *Data Understanding*.

Tahapan *Data Understanding* dalam penelitian ini ditunjukkan pada Gambar 2, meliputi identifikasi data, pengecekan *missing value*, deteksi data duplikat, serta analisis distribusi label.

Dari identifikasi dataset yang digunakan terdiri dari 158.355 baris dan 28 atribut, dengan rincian 1 atribut bertipe float64, 17 atribut int64, dan 10 atribut object. Rincian lengkap tiap atribut dan label disajikan pada Gambar 3.

P- ISSN: 2355-6684, E-ISSN: 2776-6640
https://ojs.unpkediri.ac.id/index.php/noe

```
 #   Column                    Non-Null Count   Dtype
---  ------                    --------------   -----
 0   usia                      158355 non-null  int64
 1   jenis_kelamin             158355 non-null  object
 2   wilayah                   158355 non-null  object
 3   tingkat_pendapatan        158355 non-null  object
 4   hipertensi                158355 non-null  int64
 5   diabetes                  158355 non-null  int64
 6   tingkat_kolesterol        158355 non-null  int64
 7   obesitas                  158355 non-null  int64
 8   lingkar_pinggang          158355 non-null  int64
 9   riwayat_keluarga          158355 non-null  int64
 10  status_merokok            158355 non-null  object
 11  konsumsi_alkohol          63507 non-null   object
 12  aktivitas_fisik           158355 non-null  object
 13  kebiasaan_makan           158355 non-null  object
 14  paparan_polusi_udara      158355 non-null  object
 15  tingkat_stres             158355 non-null  object
 16  jam_tidur                 158355 non-null  float64
 17  tekanan_darah_sistolik    158355 non-null  int64
 18  tekanan_darah_diastolik   158355 non-null  int64
 19  puasa_gula_darah          158355 non-null  int64
 20  hdl_kolesterol            158355 non-null  int64
 21  ldl_kolesterol            158355 non-null  int64
 22  trigliserida              158355 non-null  int64
 23  hasil_EKG                 158355 non-null  object
 24  riwayat_penyakit_jantung  158355 non-null  int64
 25  penggunaan_obat           158355 non-null  int64
 26  ikut_skrining_gratis      158355 non-null  int64
 27  serangan_jantung          158355 non-null  int64
dtypes: float64(1), int64(17), object(10)
memory usage: 33.8+ MB
```

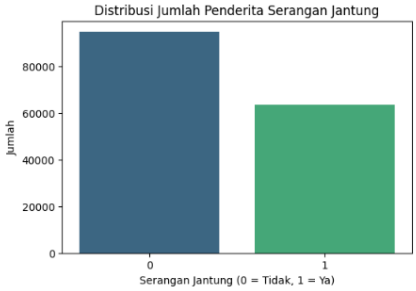
Gambar 3. Deskripsi dataset *Heart Attack*

Pada atribut *'alcohol_consumption'*, ditemukan sebanyak 94.848 data *missing value*, yang berpotensi menurunkan akurasi model prediksi. Kondisi ini ditunjukkan pada Gambar 4.

```
age                                  0
gender                               0
region                               0
income_level                         0
hypertension                         0
diabetes                             0
cholesterol_level                    0
obesity                              0
waist_circumference                  0
family_history                       0
smoking_status                       0
alcohol_consumption              94848
physical_activity                    0
dietary_habits                       0
air_pollution_exposure               0
stress_level                         0
sleep_hours                          0
blood_pressure_systolic              0
blood_pressure_diastolic             0
fasting_blood_sugar                  0
cholesterol_hdl                      0
cholesterol_ldl                      0
triglycerides                        0
EKG_results                          0
previous_heart_disease               0
medication_usage                     0
participated_in_free_screening       0
heart_attack                         0
dtype: int64
```

Gambar 4. Identifikasi *Missing Value*

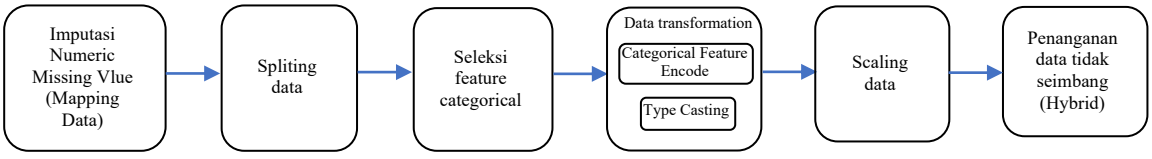
Selain itu, dataset menunjukkan ketidakseimbangan kelas (*imbalanced dataset*), sebagaimana terlihat pada Gambar 5, dengan distribusi 94.854 data untuk *class* 0 (tidak berisiko serangan jantung) dan 63.501 data untuk *class* 1 (berisiko serangan jantung). Ketimpangan ini dapat menyebabkan model cenderung lebih akurat dalam mengklasifikasikan *class* mayoritas. Sementara itu, tidak ditemukan data ganda dalam dataset yang digunakan.



Gambar 5. *Heart Attack data distribution*

4. *Data Preparation*

Data Preparation atau *Data Preprocessing* merupakan tahap penting dalam proses pengolahan data yang bertujuan untuk menyiapkan data mentah menjadi dataset akhir yang siap digunakan dalam pemodelan. Tahap ini dilakukan secara iteratif dan fleksibel tergantung pada kondisi dan kebutuhan data yang ditunjukkan pada Gambar 6. Pada penelitian ini, proses data preparation diawali dengan *imputasi missing value*, *spliting data*, *data transformation*, *Scaling data*, terakhir penanganan *imbalanced data* (Ristyawan dkk., 2025: 33).



Gambar 6. Tahap *Data Understanding*.

Pada tahap awal pra-pemrosesan data, dilakukan penanganan terhadap *missing value*, khususnya pada kolom *'alcohol_consumption'*. Hal ini disebabkan oleh inkonsistensi dalam pendeteksian kategori seperti *'None'*, yang seharusnya merepresentasikan individu yang tidak pernah mengonsumsi alkohol, namun tidak dikenali secara seragam sebagai suatu kategori. Ketidakkonsistenan tersebut disebabkan oleh adanya variasi penulisan seperti spasi tambahan dan perbedaan kapitalisasi. Untuk mengatasi hal ini, dilakukan pembersihan nilai pada kolom tersebut dan pengelompokan ulang kategori di luar *'Moderate'* dan *'High'* menjadi kategori *'No Alcohol'*, sehingga seluruh nilai dapat terklasifikasi dan *missing value* dapat dihilangkan.

Setelah penanganan *missing value*, dilakukan proses pemetaan ulang data kategorikal ke dalam bentuk numerik secara manual berdasarkan urutan logis. Sebagaimana terlihat pada Gambar 7.

```
mapping = {
    'income_level': {'Low': 0, 'Middle': 1, 'High': 2},
    'smoking_status': {'Never': 0, 'Past': 1, 'Current': 2},
    'alcohol_consumption': {'No Alcohol': 0, 'Moderate': 1, 'High': 2},
    'physical_activity': {'Low': 0, 'Moderate': 1, 'High': 2},
    'air_pollution_exposure': {'Low': 0, 'Moderate': 1, 'High': 2},
    'stress_level': {'Low': 0, 'Moderate': 1, 'High': 2},
    'EKG_results': {'Normal': 0, 'Abnormal': 1}
}
```

Gambar 7. Mapping Data

Pemetaan ini diperlukan karena label numerik pada data asli tidak mencerminkan hierarki atau intensitas yang sesuai. Sebagai contoh, pada kolom *'income_level'*, nilai 0 direpresentasikan sebagai *'High'*, 1 sebagai *'Low'*, dan 2 sebagai *'Moderate'*. Untuk memberikan representasi numerik yang sesuai, dilakukan pemetaan ulang sebagaimana terlihat pada Gambar 7. Hal ini bertujuan agar model dapat memahami hubungan antar kategori dengan lebih tepat dan mendukung pembelajaran yang lebih akurat (Bunte dkk., 2018: 432).

Setelah proses pemetaan, data kemudian dibagi menjadi dua bagian, yaitu 80% sebagai data latih dan 20% sebagai data uji. Pembagian data dilakukan sebelum proses penanganan *imbalanced data* untuk menghindari data *leakage* atau kebocoran data, yang dapat menyebabkan hasil evaluasi menjadi tidak valid karena informasi dari data uji dapat tercampur ke dalam data latih selama proses penyeimbangan *class* (Lemaitre, 2021).

Selanjutnya, dilakukan seleksi fitur kategorikal untuk menentukan fitur mana yang relevan dan akan digunakan dalam proses modeling. Fitur-fitur tersebut kemudian melalui proses transformasi data yang terdiri dari dua tahap, yaitu categorical *feature encoding* untuk mengubah fitur kategorikal menjadi bentuk numerik, serta *type casting* untuk menyeragamkan tipe data. Dalam penelitian ini, *type casting* digunakan untuk mengubah tipe data dari float menjadi integer pada kolom *'sleep_hours'* (Hasibuan dkk., 2022: 598).

Setelah transformasi selesai, data numerik kemudian melalui tahap *scaling* agar seluruh fitur memiliki skala yang seragam, terutama penting bagi algoritma yang sensitif terhadap perbedaan skala (Aziz dkk., 2021: 23). Setelah dilakukan *scaling* pada data numerik, maka dilanjutkan pada penanganan terhadap *imbalanced data* menggunakan *hybrid sampling*, yaitu gabungan antara *Random Over Sampling* (ROS) dan *Random Under Sampling* (RUS). Pada penelitian ini menggunakan algoritma *SMOTE-ENN*.

5. Pemodelan Metode

Setelah dilakukan *preprocessing*, data kemudian dimodelkan menggunakan metode *Explainable Boosting Machine* (EBM) dengan bantuan bahasa pemrograman *Python* melalui platform *Google Colab* untuk penulisan dan eksekusi *code*.

6. Skema Penanganan *Imbalaced Data* (*SMOTE-ENN*)

Berdasarkan eksplorasi awal dan rancangan eksperimen yang telah disusun, maka model percobaan dalam penelitian ini diuraikan seperti pada Tabel 3 berikut.

Tabel 3. Skema Penanganan *Imbalaced Data* (*SMOTE-ENN*)

Nama Eksperimen	Parameter *SMOTE-ENN*
SMOTEENN Base (SENB)	*SMOTEENN(random_state=42)*
SMOTEENN Sharp (SENS)	*SMOTE(sampling_strategy=1.0, k_neighbors=5, random_state=42) + EditedNearestNeighbours(n_neighbors=1)*

SMOTEENN Balanced-3 (SEN3)	*SMOTEENN(enn=EditedNearestNeighbours(n_neighbors=3), random_state=42)*
SMOTEENN Conservative (SENC)	*SMOTEENN(sampling_strategy=0.7, random_state=42)*
SMOTEENN Aggressive (SENA)	*SMOTE(sampling_strategy=0.9, k_neighbors=7, random_state=42) + EditedNearestNeighbours(n_neighbors=3)*

7. *Evaluation* dan *Validation*

Setelah melakukan serangkaian percobaan menggunakan model *Explainable Boosting Machine* (EBM), langkah selanjutnya adalah melakukan evaluasi menyeluruh terhadap performa model serta meninjau kembali setiap tahapan pembentukan model guna memastikan kesesuaiannya dengan tujuan penelitian.

Evaluasi kinerja model dilakukan dengan menggunakan *confusion matrix* sebagai dasar analisis. Dari *confusion matrix* tersebut kemudian dihitung metrik evaluasi seperti *Accuracy*, *precision*, *recall*, dan *f1-score* untuk menilai sejauh mana model mampu melakukan klasifikasi dengan baik.

D. HASIL DAN PEMBAHASAN

4.1 SMOTEENN Base (SENB)

Tabel 4. Hasil *Classification report* Eksperimen SENB

Class	***precision***	***Recall***	***F1-score***	***Support***
0	0.80	0.71	0.75	18906
1	0.63	0.74	0.68	12765
Accuracy			0.72	31671

Berdasarkan Tabel 4, model memiliki *precision* 0.80 dan *recall* 0.71 pada *class* 0, yang berarti 80% prediksi *class* 0 benar, namun 29% data *class* 0 salah diklasifikasikan sebagai *class* 1. Pada *class* 1, *precision* sebesar 0.63 menunjukkan bahwa 63% prediksi *class* 1 benar, sedangkan *recall* 0.74 berarti 26% data *class* 1 salah diklasifikasikan sebagai *class* 0. *Accuracy* model secara keseluruhan adalah 0.72, menunjukkan bahwa 72% dari seluruh prediksi sesuai dengan label sebenarnya. Meskipun *recall* untuk ***class* 1** lebih tinggi daripada *class* 0, namun ***class* 0 memiliki *precision* dan *F1-score* yang lebih tinggi**, yang menunjukkan bahwa **model secara keseluruhan lebih akurat dalam mengenali *class* 0** dibanding *class* 1.

4.2 SMOTEENN Sharp (SENS)

Tabel 5. Hasil *Classification report* Eksperimen SENS

Class	***precision***	***Recall***	***F1-score***	***Support***
0	0.74	0.87	0.80	18906
1	0.74	0.53	0.62	12765
Accuracy			0.74	31671

Berdasarkan Tabel 5, model memiliki *precision* 0.74 dan *recall* 0.87 pada *class* 0, yang berarti 74% prediksi *class* 0 benar, namun 13% data *class* 0 salah diklasifikasikan sebagai *class* 1.Pada *class* 1, *precision* sebesar 0.74 menunjukkan bahwa 74% prediksi *class* 1 benar, sedangkan *recall* 0.53 berarti 47% data *class* 1 salah diklasifikasikan sebagai *class* 0. *Accuracy* model secara keseluruhan adalah 0.74, menunjukkan bahwa 74% dari seluruh prediksi sesuai dengan label sebenarnya. Meskipun nilai *precision* kedua *class* sama, nilai ***recall* yang jauh lebih tinggi pada *class* 0** serta ***F1-score* yang lebih tinggi** menunjukkan bahwa **model lebih baik dalam mengenali *class* 0** dibandingkan *class* 1.

SMOTEENN Balanced-3 (SEN3)

Tabel 6. Hasil *Classification report* Eksperimen SEN3

Class	***precision***	***Recall***	***F1-score***	***Support***
0	0.69	0.94	0.80	18906
1	0.81	0.39	0.52	12765
Accuracy			0.71	31671

Berdasarkan Tabel 6, model memiliki *precision* 0.69 dan *recall* 0.94 pada *class* 0, yang berarti 69% prediksi *class* 0 benar, namun 6% data *class* 0 salah diklasifikasikan sebagai *class* 1.Pada *class* 1, *precision* sebesar 0.81 menunjukkan bahwa 81% prediksi *class* 1 benar, sedangkan *recall* 0.39 berarti 61% data *class* 1 salah diklasifikasikan sebagai *class* 0. *Accuracy* model secara keseluruhan adalah 0.71, menunjukkan bahwa 71% dari seluruh prediksi sesuai dengan label sebenarnya. Meskipun *precision* untuk ***class* 1** lebih tinggi daripada *class* 0, namun ***class* 0 memiliki** nilai ***recall*** dan ***F1-score*** **yang jauh lebih tinggi, yang** menunjukkan bahwa **model lebih baik dalam mengenali *class* 0** dibandingkan *class* 1.

4.3 SMOTEENN Conservative (SENC)

Tabel 7. Hasil *Classification report* Eksperimen SENC

Class	***precision***	***Recall***	***F1-score***	***Support***
0	0.74	0.85	0.79	18906
1	0.71	0.57	0.63	12765
Accuracy			0.73	31671

Berdasarkan Tabel 7, model memiliki *precision* 0.74 dan *recall* 0.85 pada *class* 0, yang berarti 74% prediksi *class* 0 benar, namun 26% data *class* 0 salah diklasifikasikan sebagai *class* 1.Pada *class* 1, *precision* sebesar 0.71 menunjukkan bahwa 71% prediksi *class* 1 benar, sedangkan *recall* 0.57 berarti 43% data *class* 1 salah diklasifikasikan sebagai *class* 0. *Accuracy* model secara keseluruhan adalah 0.73, menunjukkan bahwa 73% dari seluruh prediksi sesuai dengan label sebenarnya. Model cenderung lebih baik dalam mengenali *class* 0 dibanding *class* 1, karena ***class* 0 memiliki** nilai *precision*, ***recall*** dan ***F1-score*** **yang jauh lebih tinggi** dibanding *class* 1.

4.4 SMOTEENN Aggressive (SENA)

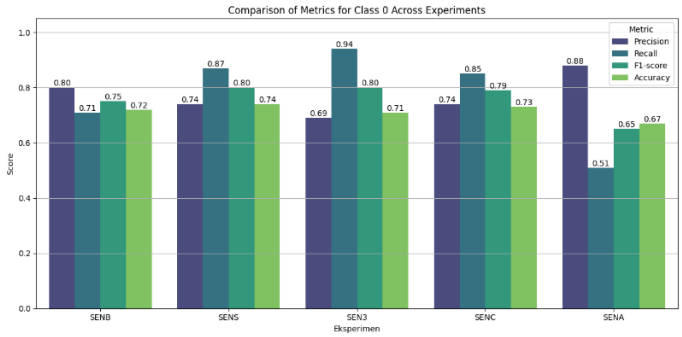
Tabel 8. Hasil *Classification report* Eksperimen SENA

Class	***precision***	***Recall***	***F1-score***	***Support***
0	0.88	0.51	0.65	18906
1	0.55	0.90	0.69	12765
Accuracy			0.67	31671

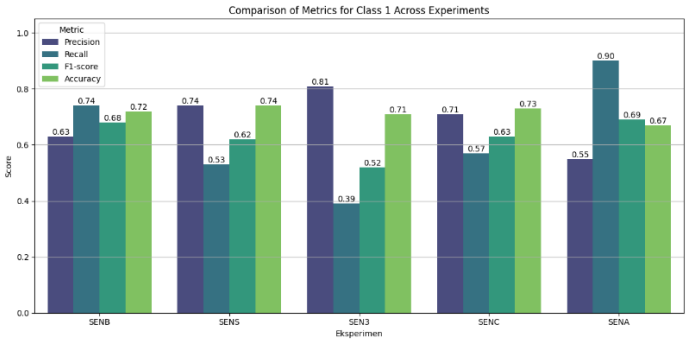
Berdasarkan Tabel 8, model memiliki *precision* 0.88 dan *recall* 0.51 pada *class* 0, yang berarti 88% prediksi *class* 0 benar, namun 49% data *class* 0 salah diklasifikasikan sebagai *class* 1.Pada *class* 1, *precision* sebesar 0.55 menunjukkan bahwa 55% prediksi *class* 1 benar, sedangkan *recall* 0.90 berarti 10% data *class* 1 salah diklasifikasikan sebagai *class* 0. *Accuracy* model secara keseluruhan adalah 0.67, menunjukkan bahwa 67% dari seluruh prediksi sesuai dengan label sebenarnya. Meskipun *precision* untuk ***class* 1** lebih rendah daripada *class* 0, namun ***class* 1 memiliki** *recall* **dan *F1-score*** **yang lebih tinggi**, yang menunjukkan bahwa **model secara keseluruhan lebih akurat dalam mengenali *class* 1** dibanding *class* 0. Namun demikian, eksperimen ini menghasilkan *Accuracy* **model terendah** dibandingkan dengan eksperimen-eksperimen sebelumnya.

Rekapitulasi kinerja pemodelan menggunakan *Explainable Boosting Machine* (EBM) setelah penerapan *preprocessing*, termasuk penanganan *missing value* dan ketidakseimbangan data dengan *SMOTE-ENN*, ditampilkan pada Gambar 8 (*class* 0) dan Gambar 9 (*class* 1). Kedua gambar tersebut menyajikan ringkasan metrik evaluasi model, yaitu *precision*, *recall*, *F1-score*, dan *Accuracy*, setelah seluruh tahapan *data preparation* diterapkan.

P- ISSN: 2355-6684, E-ISSN: 2776-6640
https://ojs.unpkediri.ac.id/index.php/noe



Gambar 8. Rekapitulasi hasil *class* 0



Gambar 9. Rekapitulasi hasil *class* 1

E. KESIMPULAN DAN SARAN

Berdasarkan hasil penelitian, dapat disimpulkan bahwa penggunaan algoritma *Explainable Boosting Machine* (EBM) dalam prediksi risiko serangan jantung memberikan performa yang baik, khususnya setelah dilakukan optimasi tahap *preprocessing* melalui penanganan *missing value* dan ketidakseimbangan data menggunakan metode SMOTE-ENN. Lima skema eksperimen yang dilakukan menunjukkan bahwa variasi parameter dalam penyeimbangan data berdampak signifikan terhadap hasil klasifikasi. Eksperimen *SMOTEENN Sharp* (SENS) menghasilkan akurasi tertinggi sebesar 74% ,akan tetapi dengan distribusi performa antar kelas yang kurang stabil. Sementara itu, Eksperimen *SMOTEENN Aggressive* (SENA) mampu meningkatkan recall pada kelas berisiko serangan jantung (*class* 1), namun menurunkan akurasi keseluruhan menjadi 67%. Temuan ini menegaskan pentingnya strategi penyeimbangan data yang tepat agar model prediksi mampu mengenali kasus berisiko serangan jantung (*class* 1) secara lebih akurat dan seimbang.

Disarankan agar penelitian ini dikembangkan dengan mengeksplorasi variasi parameter *SMOTEENN*, mencoba model lain seperti *Gradient Boosting* atau *K-NN*, serta menerapkan *feature selection* dan *hyperparameter tuning*. Evaluasi metrik tambahan seperti *AUC-ROC* dan penggunaan dataset yang lebih representatif juga diperlukan untuk meningkatkan kinerja dan validitas model.
dinilai produktif karena sifatnya yang fleksibel dan menjangkau berbagai lapisan masyarakat. Oleh karena itu, perhatian terhadap pola penyebaran dan jumlah UMKM di tiap wilayah sangat penting untuk perencanaan pembangunan ekonomi daerah[4].

DAFTAR PUSTAKA

Alahmadi, R., Almujibah, H., Alotaibi, S., Elshekh, A. E. A., Alsharif, M., & Bakri, M. (2023). Explainable Boosting Machine: A Contemporary Glass-Box Model to Analyze Work Zone-Related Road Traffic Crashes. *Safety*, *9*(4), 1–15. https://doi.org/10.3390/safety9040083

Apollo, R. S. (2025). *Infark miokard - Penyebab, Gejala, Diagnosis, Pengobatan, dan Pencegahan*. Apollo Hospitals.

Arslan, A. K., Yagin, F. H., Algarni, A., AL-Hashem, F., & Ardigò, L. P. (2024). Combining the Strengths of the Explainable Boosting Machine and Metabolomics Approaches for Biomarker Discovery in Acute Myocardial Infarction. *Diagnostics*, *14*(13). https://doi.org/10.3390/diagnostics14131353

Aziz, A. R., Warsito, B., & Prahutama, A. (2021). Pengaruh Transformasi Data Pada Metode Learning Vector Quantization Terhadap Akurasi Klasifikasi Diagnosis Penyakit Jantung. *Jurnal Gaussian*, *10*(1), 21–30. https://doi.org/10.14710/j.gauss.v10i1.30933

Azizah, N. (2023). *Kemenkes: Penyakit Kardiovaskular Jadi Penyebab Kematian Terbanyak di Indonesia*. Republika.Co.Id. https://news.republika.co.id/berita/s1jq78463/kemenkes-penyakit-kardiovaskular-jadi-penyebab-kematian-terbanyak-di-indonesia

Bunte, A., Li, P., & Niggemann, O. (2018). Mapping data sets to concepts using machine learning and a knowledge based approach. *ICAART 2018 - Proceedings of the 10th International Conference on Agents and Artificial Intelligence*, *2*(Icaart), 430–437. https://doi.org/10.5220/0006590204300437

Hasibuan, E., Informasi, S., Ilmu, F., Informasi, T., Gunadarma, U., Margonda, J., No, R., Cina, P., & Jawa, D. (2022). Implementasi Machine Learning untuk Prediksi Harga Mobil Bekas dengan Algoritma Regresi Linear berbasis Web. *Jurnal Ilmiah Komputasi*, *21*(4), 595–602. https://doi.org/10.32409/jikstik.21.4.3327

Kementrian Kesehatan Republik Indonesia. (2022). *Penyakit Jantung Penyebab Utama Kematian, Kemenkes Perkuat Layanan Primer*. Public. https://sehatnegeriku.kemkes.go.id/baca/rilis-media/20220929/0541166/penyakit-jantung-penyebab-utama-kematian-kemenkes-perkuat-layanan-primer/

Lemaitre, G. (2021). *8. Common pitfalls and recommended practices*. Imbalanced-Learn.Org. https://imbalanced-learn.org/stable/common_pitfalls.html

Pamungkas, B. P., Vikri, M. J., & Aristia, I. (2025). *Application of SMOTE-ENN Method in Data Balancing for Classification of Diabetes Health Indicators with C4 . 5 Algorithm*. *14*, 183–188.

Putri, R. W., Ristyawan, A., & Muzaki, M. N. (2022). Comparison Performance of K-NN and NBC Algorithm for Classification of Heart Disease. *JTECS : Jurnal Sistem Telekomunikasi Elektronika Sistem Kontrol Power Sistem Dan Komputer*, *2*(2), 143. https://doi.org/10.32503/jtecs.v2i2.2708

Ristyawan, A., Nugroho, A., & Amarya, T. K. (2025). *Optimasi Preprocessing Model Random Forest Untuk Prediksi Stroke*. *12*(1).

Roihan, A., Sunarya, P. A., & Rafika, A. S. (2020). Pemanfaatan Machine Learning dalam Berbagai Bidang: Review paper. *IJCIT (Indonesian Journal on Computer and Information Technology)*, *5*(1), 75–82. https://doi.org/10.31294/ijcit.v5i1.7951

Saputro, M. R., Mahdiyah, U., Swanjaya, D., Nusantara, U., & Kediri, P. (2024). Perbandingan Metode Adaptive Boosting dan Extreme Gradient Boosting Untuk Prediksi Hasil Pertandingan Liga Spanyol. *Jurnal Nusantara Of Engineering*, *7*(1), 67–73. https://ojs.unpkediri.ac.id/index.php/noe

Siambaton, M. Z., & Husein, A. M. (2022). Menganalisis Data Kesehatan Global : Pendekatan Analisis Data Eksplorasi Visual. *Data Sciences Indonesia (DSI)*, *1*(2), 41–49. https://doi.org/10.47709/dsi.v1i2.1315

Thygesen, K., Alpert, J. S., Jaffe, A. S., Chaitman, B. R., Bax, J. J., Morrow, D. A., White, H. D., Corbett, S., Chettibi, M., Hayrapetyan, H., Roithinger, F. X., Aliyev, F., Sujayeva, V., Claeys, M. J., Smajić, E., Kala, P., Iversen, K. K., Hefny, E. El, Marandi, T., … Parkhomenko, A. (2018). Fourth Universal Definition of Myocardial Infarction (2018). In *Circulation* (Vol. 138, Issue 20). https://doi.org/10.1161/CIR.0000000000000617