

**PREDIKSI PENYAKIT DIABETES MENGGUNAKAN
ENSAMBLE LEARNING DAN METODE RESAMPLING
PADA DATATIDAK SEIMBANG
SKRIPSI**

Diajukan untuk memenuhi salah satu syarat mencapai derajat Sarjana
Program Studi Informatika



disusun oleh

NURWAHYUNINGSIH

22.11.4897

Kepada

**FAKULTAS ILMU KOMPUTER
UNIVERSITAS AMIKOM YOGYAKARTA
YOGYAKARTA
2026**

**PREDIKSI PENYAKIT DIABETES MENGGUNAKAN
ENSAMBLE LEARNING DAN METODE RESAMPLING
PADA DATA TIDAK SEIMBANG**

SKRIPSI

untuk memenuhi salah satu syarat mencapai derajat Sarjana
Program Studi Informatika



disusun oleh
NURWAHYUNINGSIH
22.11.4897

Kepada

**FAKULTAS ILMU KOMPUTER
UNIVERSITAS AMIKOM YOGYAKARTA
YOGYAKARTA**

2026

HALAMAN PERSETUJUAN

SKRIPSI

**PREDIKSI PENYAKIT DIABETES MENGGUNAKAN ENSAMBLE
LEARNING DAN METODE RESAMPLING PADA DATA TIDAK
SEIMBANG**

yang disusun dan diajukan oleh

Nurwahyuningsih

22.11.4897

telah disetujui oleh Dosen Pembimbing Skripsi
pada tanggal 02 Maret 2026

Dosen Pembimbing,



I Made Artha Agastya, S.T., M.Eng., Ph.D.,

NIK. 190302352

HALAMAN PENGESAHAN

SKRIPSI

**PREDIKSI PENYAKIT DIABETES MENGGUNAKAN ENSAMBLE
LEARNING DAN METODE RESAMPLING PADA DATA TIDAK
SEIMBANG**

yang disusun dan diajukan oleh

Nurwahyuningsih

22.11.4897

Telah dipertahankan di depan Dewan Penguji
pada tanggal 02 Maret 2026

Susunan Dewan Penguji

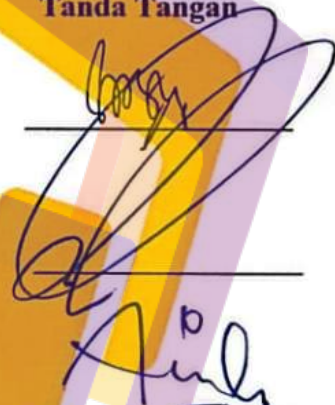
Nama Penguji

Tanda Tangan

Dr. Emigawaty, M.Kom
NIK. 190302226

Rizqi Sukma Kharisma, S.Kom., M.Kom.
NIK. 190302215

I Made Artha Agastya, S.T., M.Eng., Ph.D.
NIK. 190302352



Skripsi ini telah diterima sebagai salah satu persyaratan
untuk memperoleh gelar Sarjana Komputer
Tanggal 02 Maret 2026

DEKAN FAKULTAS ILMU KOMPUTER



Prof. Dr. Kusrini, M.Kom.
NIK. 190302106

HALAMAN PERNYATAAN KEASLIAN SKRIPSI

Yang bertandatangan di bawah ini,

Nama mahasiswa : Nurwahyuningsih
NIM : 22.11.4897

Menyatakan bahwa Skripsi dengan judul berikut:

PREDIKSI PENYAKIT DIABETES MENGGUNAKAN ENSAMBLE LEARNING DAN METODE RESAMPLING PADA DATA TIDAK SEIMBANG

Dosen Pembimbing : I Made Artha Agastya, S.T., M.Eng., Ph.D.,

1. Karya tulis ini adalah benar-benar ASLI dan BELUM PERNAH diajukan untuk mendapatkan gelar akademik, baik di Universitas ANIIKOM Yogyakarta maupun di Perguruan Tinggi Iainnya.
2. Karya tulis ini merupakan gagasan, rumusan dan penelitian SAYA sendiri, tanpa bantuan pihak Iain kecuali arahan dari Dosen Pembimbing.
3. Dalam karya tulis ini tidak terdapat karya atau pendapat orang Iain, kecuali secara tertulis dengan jelas dicantumkan sebagai acuan dalam naskah dengan disebutkan nama pengarang dan disebutkan dalam Daftar Pustaka pada karya tulis ini.
4. Perangkat lunak yang digunakan dalam penelitian ini sepenuhnya menjadi tanggung jawab SAYA, bukan tanggung jawab Universitas AMIKOM Yogyakarta.
5. Pernyataan ini SAYA buat dengan sesungguhnya, apabila di kemudian hari terdapat penyimpangan dan ketidakbenaran dalam pernyataan ini, maka SAYA bersedia menerima SANKSI AKADEMIK dengan pencabutan gelar yang sudah diperoleh, serta sanksi Iainnya sesuai dengan norma yang berlaku di Perguruan Tinggi.

Yogyakarta, 2 Maret 2026

Yang Menyatakan,



The image shows a red official stamp from Universitas AMIKOM Yogyakarta. The stamp contains the text 'UNIVERSITAS AMIKOM YOGYAKARTA', 'METERAI TEMPEL', and the number '00666ANX28082543'. A handwritten signature in black ink is written over the stamp. Below the signature, the name 'Nurwahyuningsih' is printed in black.

HALAMAN PERSEMBAHAN

Skripsi ini dipersembahkan dengan penuh rasa syukur dan hormat kepada semua pihak yang telah memberikan doa, dukungan, dan semangat sehingga penulis dapat menyelesaikan karya ini dengan baik. Secara khusus, persembahan ini ditujukan kepada:

1. Bapak, Ibu, dan kakak serta keluarga tercinta, yang selalu memberikan doa, cinta, dukungan, dan semangat tanpa henti selama proses belajar dan penyusunan skripsi ini.
2. Dosen Pembimbing, bapak I Made Artha Agastya, S.T., M.Eng., Ph.D., yang telah memberikan bimbingan, arahan, dan motivasi yang sangat berharga selama proses penyusunan skripsi ini.
3. Teman-teman seperjuangan, yang selalu memberikan motivasi, dukungan moral, waktu, telinga dan kebersamaan selama perjalanan akademik hingga skripsi ini selesai.
4. Semua pihak lain yang telah memberikan dukungan, bantuan, atau kontribusi, baik secara langsung maupun tidak langsung, sehingga skripsi ini dapat terselesaikan.
5. Kepada diri saya sendiri, sebagai apresiasi atas dedikasi, komitmen, dan usaha yang telah diberikan selama proses penyusunan skripsi ini.

KATA PENGANTAR

Puji syukur penulis panjatkan ke hadirat Tuhan Yang Maha Esa atas segala rahmat dan karunia-Nya sehingga skripsi ini dapat diselesaikan dengan baik. Skripsi ini disusun sebagai salah satu syarat untuk memperoleh gelar Sarjana pada Program Studi Informatika di Universitas Amikom Yogyakarta.

Selama proses penyusunan skripsi ini, penulis memperoleh berbagai bantuan, arahan, dan dukungan dari banyak pihak. Oleh karena itu, penulis menyampaikan terima kasih kepada dosen pembimbing, tim dosen penguji, serta seluruh pihak yang telah berkontribusi sehingga skripsi ini dapat terselesaikan.

Penulis menyadari bahwa skripsi ini masih memiliki keterbatasan. Oleh sebab itu, kritik dan saran yang membangun sangat diharapkan sebagai bahan perbaikan di masa mendatang. Semoga skripsi ini dapat memberikan kontribusi positif bagi pengembangan ilmu pengetahuan.

Yogyakarta, 2 Maret 2026

Nurwahyuningsih

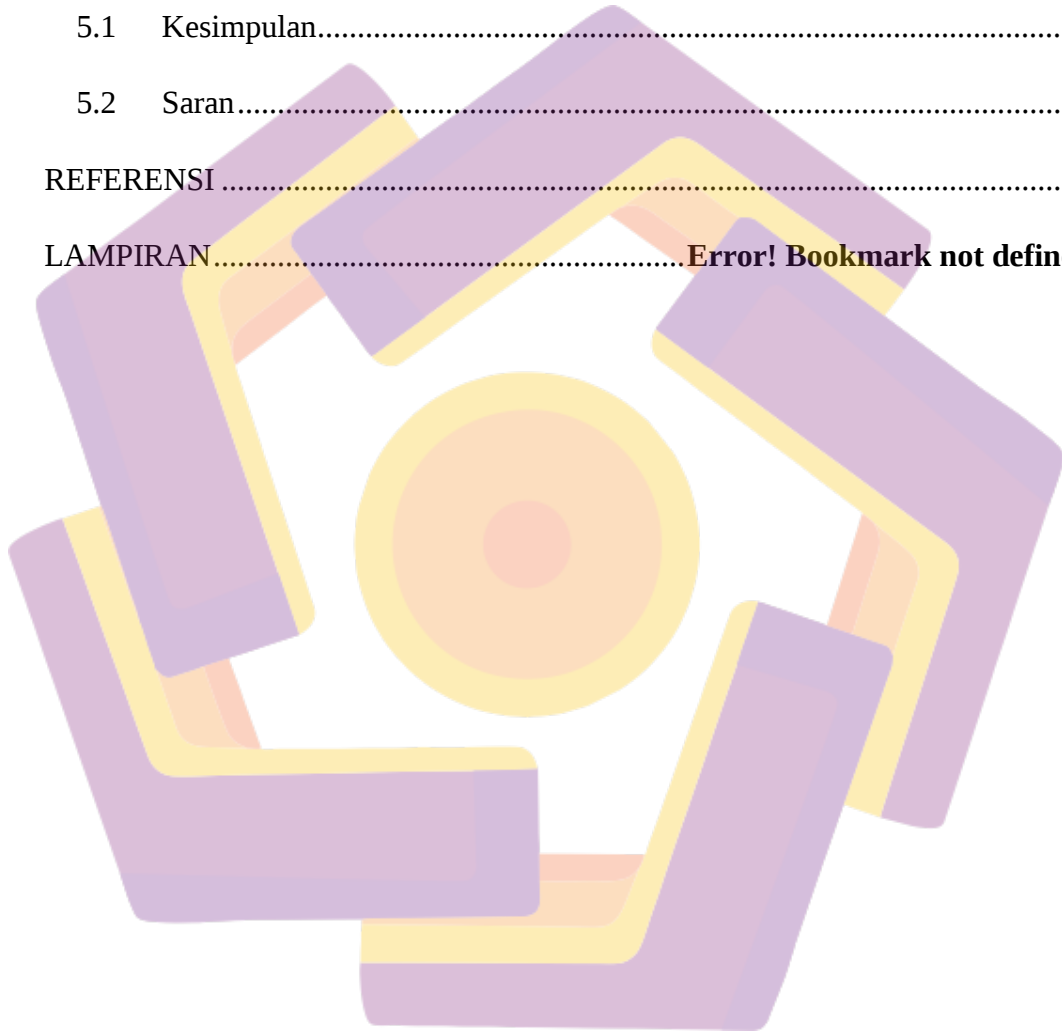
DAFTAR ISI

HALAMAN JUDUL	i
HALAMAN PERSETUJUAN.....	ii
HALAMAN PENGESAHAN	iii
HALAMAN PERNYATAAN KEASLIAN SKRIPSI.....	iv
HALAMAN PERSEMBAHAN	v
KATA PENGANTAR	vi
DAFTAR ISI.....	vii
DAFTAR TABEL.....	xi
DAFTAR GAMBAR.....	xii
DAFTAR LAMPIRAN.....	Error! Bookmark not defined.
DAFTAR LAMBANG DAN SINGKATAN	xiii
DAFTAR ISTILAH	xiv
INTISARI	xv
ABSTRACT.....	xvi
BAB I PENDAHULUAN.....	1
1. Latar Belakang	1
2. Rumusan Masalah.....	2
3. Batasan Masalah	2
4. Tujuan Penelitian	3

5. Manfaat Penelitian	3
6. Sistematika Penulisan	4
BAB II TINJAUAN PUSTAKA	5
2.1 Studi Literatur	5
2.2 Dasar Teori.....	10
2.2.1 Data Mining	10
2.2.2 Klasifikasi	10
2.2.3 Ketidakseimbangan Data (Imbalanced Data)	10
2.2.4 Metode Oversampling.....	11
2.2.4 Metode Undersampling.....	13
2.2.5 Algoritma Random Forest.....	14
2.2.6 Algoritma XGBoost.....	15
2.2.7 Evaluasi Kinerja Model	16
BAB III METODE PENELITIAN	18
3.1 Objek Penelitian	18
3.2 Alur Penelitian.....	19
3.2.1 Pengumpulan Dataset.....	20
3.2.2 Load Data.....	20
3.2.3 Exploratory Data Analysis (EDA)	20
3.2.4 Preprocessing	22
3.2.5 Pemodelan Data Asli.....	22
3.2.6 Penangan Ketidakseimbangan kelas	22

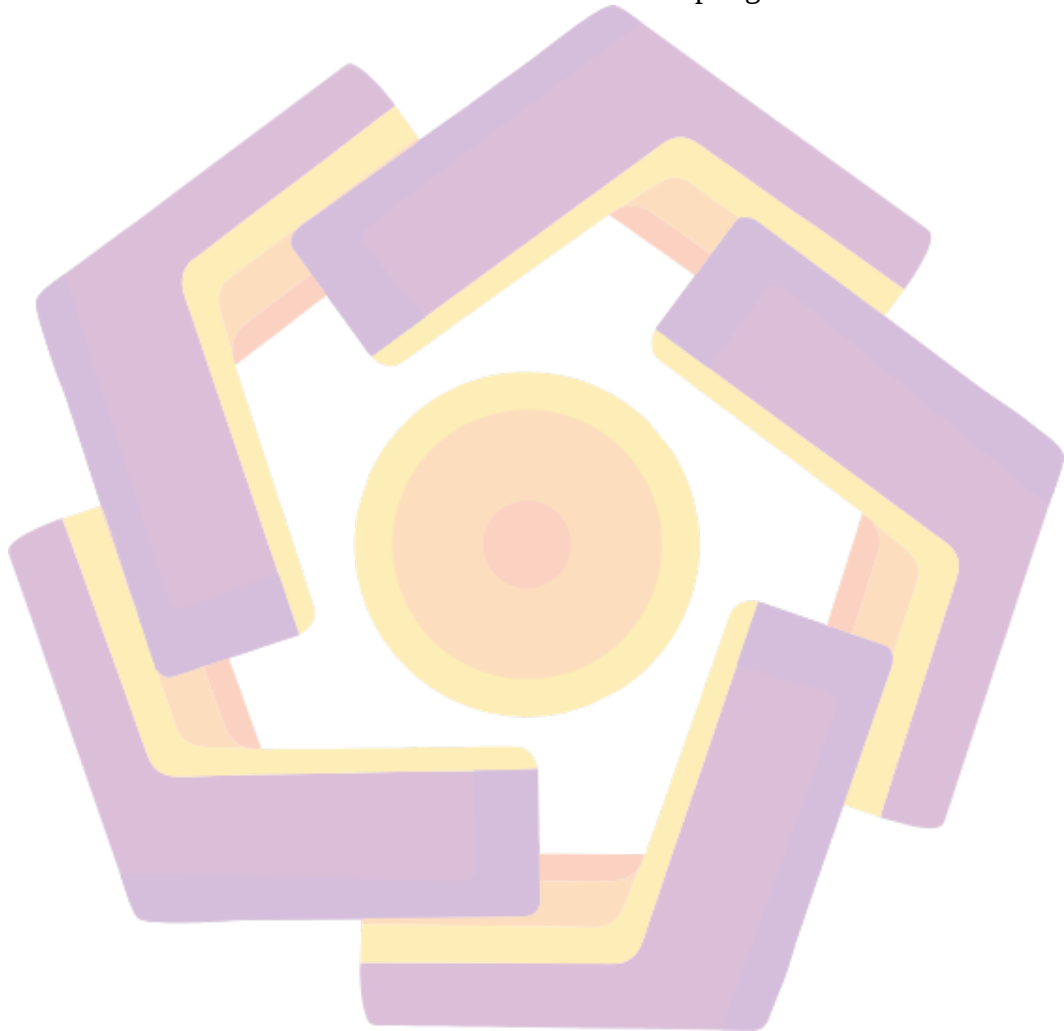
3.2.7 Pemodelan Dengan Metode Sampling.....	23
3.2.8 Evaluasi.....	23
3.2.9 Analisis Hasil	23
3.2.10 Kesimpulan	23
3.3 Alat dan Bahan.....	24
3.3.1 Data Penelitian	24
3.3.2 Alat/instrument	24
BAB IV HASIL DAN PEMBAHASAN	25
4.1 Hasil Exploratory Data Analisis (EDA).....	25
4.1.1 Analisis Distribusi Kelas.....	25
4.1.2 Analisis Fitur Numerik.....	26
4.1.3 Korelasi Fitur Numerik	31
4.1.4 Analisis Fitur Kategorial.....	32
4.1.5 Missing Values Dan Duplikasi	33
4.1.6 Identifikasi Fitur Tidak Relevan	34
4.2 Hasil Preprocessing.....	34
4.2.1 One Hot Encoding.....	34
4.2.2 Pembagian Data (Train-Test Split)	35
4.3 Pemodelan Menggunakan Data Asli (Tanpa Sampling).....	35
4.4 Penanganan Ketidakseimbangan Kelas	36
4.5 Hasil Pemodelan Setelah Sampling	36
4.6 Analisis Perbandingan Model	41

4.7 Analisis Rata-rata Kinerja Model	42
4.7.1 Analisis F1-score	42
4.7.2 Analisis Accuracy	43
BAB V PENUTUP	46
5.1 Kesimpulan.....	46
5.2 Saran.....	46
REFERENSI	48
LAMPIRAN.....	Error! Bookmark not defined.



DAFTAR TABEL

Tabel 2.1. Keaslian Penelitian	23
Tabel 4.1. Hasil Pemodelan Random Forest Setelah Sampling	56
Tabel 4.2. Hasil Pemodelan XGBoost Setelah Sampling	57



DAFTAR GAMBAR

Gambar 2.1 Imbalanced data	28
Gambar 3.1 Alur Penelitian	36
Gambar 4.1 Distribusi Kelas	44
Gambar 4.2 Boxplot Age	45
Gambar 4.3 Boxplot BMI	46
Gambar 4.4 Boxplot HbA1c	47
Gambar 4.5 Blood Glucose	47
Gambar 4.6 Korelasi Fitur Numerik	49
Gambar 4.7 Distribusi Gender	50
Gambar 4.8 Distribusi Smoking	51
Gambar 4.9 Rata-rata F1-score	60
Gambar 4.10 Rata-rata Accuracy	61

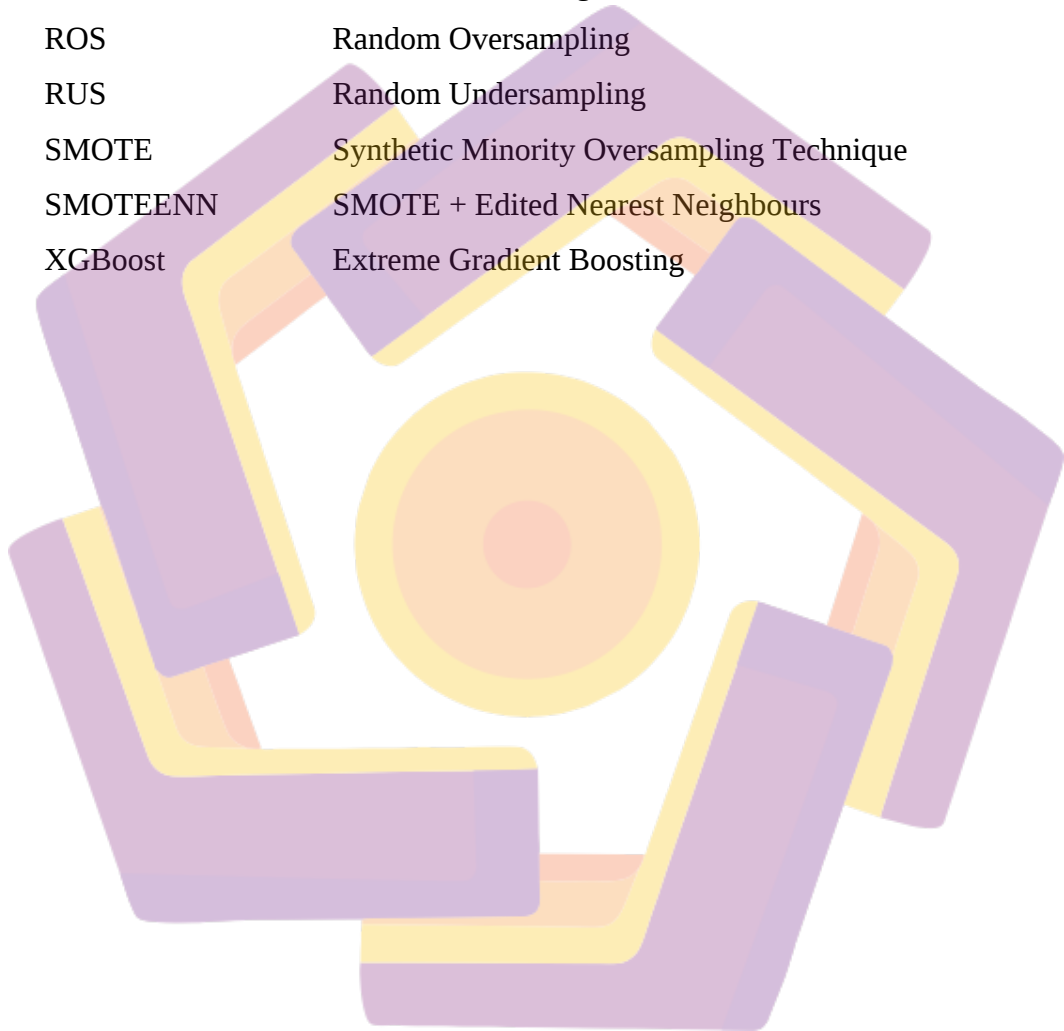
DAFTAR LAMBANG DAN SINGKATAN

TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative
Obj(t)	Fungsi objektif pada iterasi ke-t
C	Jumlah kelas dalam dataset
$\Omega(f_t)$	Fungsi regularisasi pada model
N	Jumlah data



DAFTAR ISTILAH

ADASYN	Adaptive Synthetic Sampling
ENN	Edited Nearest Neighbours
ROS	Random Oversampling
RUS	Random Undersampling
SMOTE	Synthetic Minority Oversampling Technique
SMOTEENN	SMOTE + Edited Nearest Neighbours
XGBoost	Extreme Gradient Boosting



INTISARI

Ketidakseimbangan kelas (imbalanced data) merupakan permasalahan utama dalam klasifikasi data medis, khususnya ada kasus penyakit diabetes. Kondisi ini menyebabkan model klasifikasi cenderung lebih fokus pada kelas mayoritas, sehingga kemampuan dalam mendeteksi pasien diabetes sebagai kelas minoritas menjadi kurang optimal. Apabila tidak ditangani dengan tepat, permasalahan tersebut dapat menurunkan kualitas sistem pendukung keputusan dibidang Kesehatan.

Penelitian ini bertujuan untuk menganalisis pengaruh berbagai metode sampling dalam meningkatkan kinerja model klasifikasi diabetes. Metode yang digunakan adalah Random Forest dan XGBoost dengan menggunakan beberapa pendekatan penanganan ketidakseimbangan kelas, yaitu SMOTE, ROS, SMOTENN, ADASYN, RUS, ClusterCentroids, Tomek Links, dan ENN. Evaluasi kinerja model dilakukan menggunakan metrik akurasi dan F1-score guna menilai keseimbangan performa dalam mengklasifikasikan kelas mayoritas dan minoritas.

Hasil penelitian menunjukkan bahwa metode Tomek Links memberikan performa yang paling konsisten pada kedua model. Pada model XGBoost, metode ini mampu meningkatkan nilai F1-score dibanding data asli, sedangkan pada model Random Forest performanya relatif stabil dengan perbedaan yang tidak signifikan. Penelitian ini memberikan kontribusi dalam pemilihan metode sampling yang efektif untuk klasifikasi data medis. Hasil penelitian dapat dimanfaatkan oleh peneliti, praktisi data, dan institusi Kesehatan sebagai referensi dalam pengembangan sistem klasifikasi penyakit berbasis pembelajaran mesin.

Kata kunci: diabetes, imbalanced data, sampling, Random Forest, XGBoost

ABSTRACT

Class imbalance (imbalanced data) is a major problem in medical data classification, particularly in the case of diabetes. This condition causes classification models to focus more on the majority class, resulting in less than optimal ability to detect diabetes patients as a minority class. If not addressed properly, this problem can degrade the quality of decision support systems in the healthcare sector.

This study aims to analyze the effect of various sampling methods on improving the performance of diabetes classification models. The methods used were Random Forests and XGBoost, employing several approaches to addressing class imbalance: SMOTE, ROS, SMOTENN, ADASYN, RUS, ClusterCentroids, Tomek Links, and ENN. Model performance was evaluated using accuracy and F1-score metrics to assess the balance of performance in classifying the majority and minority classes.

The results showed that the Tomek Links method provided the most consistent performance across both models. In the XGBoost model, this method was able to improve the F1-score compared to the original data, while in the Random Forest model, performance was relatively stable with insignificant differences. This research contributes to the selection of effective sampling methods for medical data classification. The results can be used by researchers, data practitioners, and healthcare institutions as a reference in developing machine learning-based disease classification systems.

Keyword: diabetes, imbalanced data, sampling, Random Forest, XGBoost