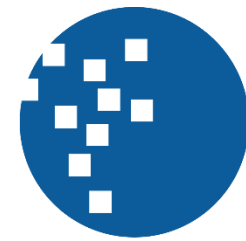


**ADVANCING EARLY-STAGE DIABETES PREDICTION:
A COMPARATIVE STUDY OF DATA OVERSAMPLING AND
CLASSIFICATION METHODS**



UMN

UNIVERSITAS
MULTIMEDIA
NUSANTARA

SKRIPSI

Michelle Melody d'Viola

00000053492

**PROGRAM STUDI SISTEM INFORMASI
FAKULTAS TEKNIK DAN INFORMATIKA
UNIVERSITAS MULTIMEDIA NUSANTARA
TANGERANG**

2025

**ADVANCING EARLY-STAGE DIABETES PREDICTION:
A COMPARATIVE STUDY OF DATA OVERSAMPLING AND
CLASSIFICATION METHODS**



Diajukan sebagai Salah Satu Syarat untuk Memperoleh
Gelar Sarjana Komputer (S.Kom.)

Michelle Melody d'Viola

00000053492

**PROGRAM STUDI SISTEM INFORMASI
FAKULTAS TEKNIK DAN INFORMATIKA
UNIVERSITAS MULTIMEDIA NUSANTARA
TANGERANG**

2025

HALAMAN PERNYATAAN TIDAK PLAGIAT

Dengan ini saya,

Nama : Michelle Melody d'Viola

Nomor Induk Mahasiswa : 00000053492

Program Studi : Sistem Informasi

Skripsi dengan judul:

“Advancing Early-Stage Diabetes Prediction: A Comparative Study of Data
Oversampling and Classification Methods”

Merupakan hasil karya saya sendiri bukan plagiat dari laporan karya tulis ilmiah yang ditulis oleh orang lain, dan semua sumber, baik yang dikutip maupun dirujuk, telah saya nyatakan dengan benar serta dicantumkan di Daftar Pustaka.

Jika di kemudian hari terbukti ditemukan kecurangan/penyimpangan, baik dalam pelaksanaan maupun dalam penulisan laporan karya tulis ilmiah, saya bersedia menerima konsekuensi dinyatakan TIDAK LULUS untuk mata kuliah yang telah saya tempuh.

Tangerang, 12 Juni 2025



(Michelle Melody d'Viola)

HALAMAN PENGESAHAN

Skripsi dengan judul

**Advancing Early-Stage Diabetes Prediction: A Comparative Study of Data
Oversampling and Classification Methods**

Oleh

Nama : Michelle Melody d'Viola
NIM : 00000053492
Program Studi : Sistem Informasi
Fakultas : Teknik dan Informatika

Telah diujikan pada hari Rabu, 11 Juni 2025

Pukul 09.00 s.d 10.00 dan dinyatakan

LULUS

Dengan susunan penguji sebagai berikut.

Ketua Sidang

Pembimbing

Johan Setiawan, S.Kom., M.M., MBA
327106402

Ir. Raymond Sunardi Setama, M.CIS
328046803

Ketua Program Studi Sistem Informasi

Ririn Ikana Desanti, S.Kom., M.Kom.
313058001

HALAMAN PERSETUJUAN PUBLIKASI KARYA ILMIAH

Yang bertanda tangan di bawah ini:

Nama : Michelle Melody d'Viola

NIM : 00000053492

Program Studi : Sistem Informasi

Jenjang : S1

Judul Karya Ilmiah : Advancing Early-Stage Diabetes Prediction:
A Comparative Study of Data Oversampling
and Classification Methods

Menyatakan dengan sesungguhnya bahwa saya bersedia* (pilih salah satu):

- ☐ Saya bersedia memberikan izin sepenuhnya kepada Universitas Multimedia Nusantara untuk mempublikasikan hasil karya ilmiah saya ke dalam repositori Knowledge Center sehingga dapat diakses oleh Sivitas Akademika UMN/Publik. Saya menyatakan bahwa karya ilmiah yang saya buat tidak mengandung data yang bersifat konfidensial.
- ☒ Saya tidak bersedia mempublikasikan hasil karya ilmiah ini ke dalam repositori Knowledge Center, dikarenakan: dalam proses pengajuan publikasi ke jurnal/konferensi nasional/internasional (dibuktikan dengan *letter of acceptance*) **.
- ☐ Lainnya, pilih salah satu:
 - ☐ Hanya dapat diakses secara internal Universitas Multimedia Nusantara
 - ☐ Embargo publikasi karya ilmiah dalam kurun waktu 3 tahun.

Tangerang, 15 Mei 2025

(Michelle Melody d'Viola)

* Pilih salah satu

** Jika tidak bisa membuktikan LoA jurnal/HKI, saya bersedia mengizinkan penuh karya ilmiah saya untuk dipublikasikan ke KC UMN dan menjadi hak institusi UMN.

KATA PENGANTAR

Puji dan syukur saya panjatkan kepada Tuhan Yang Maha Esa karena atas kasih setia dan pimpinannya saya dapat menyelesaikan tugas akhir ini. Skripsi dengan judul “ *Advancing Early-Stage Diabetes Prediction: A Comparative Study of Data Oversampling and Classification Methods*” ini disusun dengan tujuan untuk memenuhi salah satu syarat untuk meraih gelar Sarjana (S1) Jurusan Sistem Informasi, Fakultas Teknik dan Informatika, Universitas Multimedia Nusantara. Saya menyadari bahwa tanpa bimbingan serta bantuan dari berbagai pihak, saya tidak dapat menyelesaikan tugas akhir ini dengan baik. Oleh karena itu, saya ingin mengucapkan terima kasih kepada:

1. Bapak Dr. Ir. Andrey Andoko, M.Sc., selaku Rektor Universitas Multimedia Nusantara.
2. Bapak Dr. Eng. Niki Prastomo, S.T., M.Sc., selaku Dekan Fakultas Universitas Multimedia Nusantara.
3. Ibu Ririn Ikana Desanti, S.Kom., M.Kom., selaku Ketua Program Studi Universitas Multimedia Nusantara.
4. Bapak Ir. Raymond Sunardi Oetama, M.CIS, sebagai Pembimbing yang telah memberikan bimbingan, arahan, dan motivasi atas terselesainya tugas akhir ini.
5. Keluarga saya yang telah memberikan bantuan dukungan material dan moral, sehingga penulis dapat menyelesaikan tugas akhir ini.

Semoga penelitian dalam tugas akhir ini dapat memberikan manfaat yang nyata bagi masyarakat dalam mendukung deteksi dini penyakit diabetes serta menjadi sumber referensi yang berharga bagi para peneliti selanjutnya.

Tangerang, 15 Mei 2025



(Michelle Melody d'Viola)

PENGEMBANGAN PREDIKSI DIABETES TAHAP AWAL: STUDI KOMPARATIF METODE OVERSAMPLING DATA DAN KLASIFIKASI

(Michelle Melody d'Viola)

ABSTRAK

Diabetes merupakan penyakit kronis yang umum dan terus meningkat secara global, dengan sejumlah besar kasus yang belum terdiagnosis. Diperlukan model prediksi tahap awal yang optimal untuk mendukung pencegahan dini. Penelitian ini mengkaji kombinasi teknik oversampling dengan metode klasifikasi guna menentukan model yang paling optimal untuk prediksi diabetes tahap awal. Pertama, data diperluas ke dalam dua skenario menggunakan metode Synthetic Minority Over-sampling Technique (SMOTE) dan Adaptive Synthetic Sampling (ADASYN) untuk mengatasi ketidakseimbangan kelas. Random Forest (RF), Extreme Gradient Boosting (XGB), Multilayer Perceptron (MLP), serta model hibrida yang menggabungkan dua metode klasifikasi tersebut dievaluasi untuk menentukan model yang paling efektif. Kombinasi SMOTE dengan Random Forest (SMOTE + RF), ADASYN dengan Random Forest (ADASYN + RF), serta ADASYN dengan gabungan Extreme Gradient Boosting dan Multilayer Perceptron (ADASYN + XGB-MLP) menunjukkan akurasi tertinggi sebesar 99,04% dengan skor F1 sebesar 0,99. Penggabungan teknik oversampling dengan metode klasifikasi ini meningkatkan akurasi prediksi sekaligus mengatasi ketidakseimbangan kelas, sehingga menjadi alat yang berguna untuk deteksi dini diabetes.

Kata kunci: Diabetes Tahap Awal, Klasifikasi, Oversampling

ADVANCING EARLY-STAGE DIABETES PREDICTION: A COMPARATIVE STUDY OF DATA OVERSAMPLING AND CLASSIFICATION METHODS

(Michelle Melody d'Viola)

ABSTRACT (English)

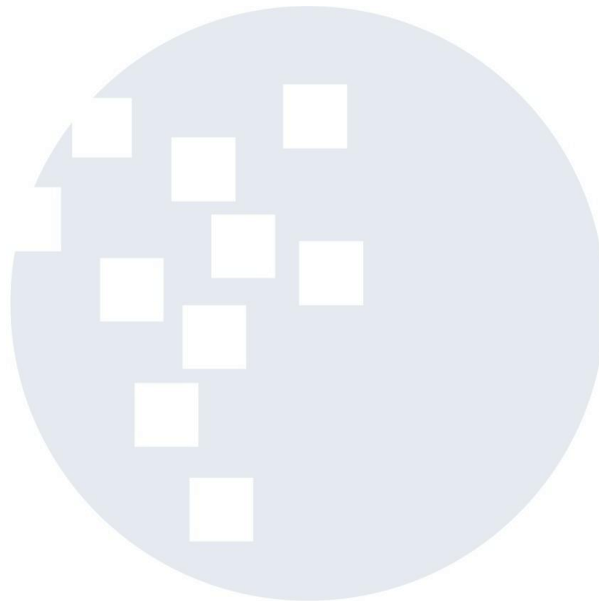
Diabetes is a common and increasing chronic disease worldwide with a large number of undiagnosed cases. An optimal early-stage prediction model is needed to support early prevention. This study has examined the combination of oversampling techniques with classifiers to determine the optimal models for early-stage diabetes prediction. First, the dataset was augmented into two scenarios using the Synthetic Minority Over-sampling Technique (SMOTE) and Adaptive Synthetic Sampling (ADASYN) methods to overcome class imbalance. Random Forest (RF), Extreme Gradient Boosting (XGB), Multilayer Perceptron (MLP), and hybrid models combining two of these classifiers were evaluated to determine the most effective model. The combinations of the Synthetic Minority Over-sampling Technique with Random Forest (SMOTE + RF), Adaptive Synthetic Sampling with Random Forest (ADASYN + RF), as well as Adaptive Synthetic Sampling with Extreme Gradient Boosting and Multilayer Perceptron (ADASYN + XGB-MLP) demonstrated the highest accuracy at 99.04% with an F1-score of 0.99. Combining oversampling techniques with these classifiers enhances the accuracy of prediction while addressing class imbalance, and thus a useful tool for early diabetes detection.

Keywords: *Classification, Early Stage Diabetes, Oversampling*

DAFTAR ISI

HALAMAN PERNYATAAN TIDAK PLAGIAT	ii
HALAMAN PENGESAHAN	iii
HALAMAN PERSETUJUAN PUBLIKASI KARYA ILMIAH	iv
KATA PENGANTAR.....	v
ABSTRAK	vi
<i>ABSTRACT (English)</i>	vii
DAFTAR ISI.....	viii
DAFTAR TABEL	x
DAFTAR GAMBAR.....	xi
DAFTAR LAMPIRAN	xii
1 Introduction	13
2 Research Method.....	15
1.1 Dataset.....	16
1.2 Data Exploration	17
1.3 Preprocessing.....	18
1.4 Methods.....	18
1.4.1 Synthetic Minority Over-sampling Technique.....	19
1.4.2 Adaptive Synthetic Sampling.....	19
1.4.3 Random Forest	19
1.4.4 Extreme Gradient Boosting.....	19
1.4.5 Multilayer Perceptron	19
1.4.6 Hybrid Model	20
1.4.7 Hyperparameter Tuning and Feature Selection	20
1.5 Evaluation	20
1.5.1 Accuracy.....	21
1.5.2 Precision	21
1.5.3 Recall.....	21
1.5.4 F1-Score	21
3 Result and Analysis	21
4 Discussion.....	27

5 Conclusion and Future Work	27
Acknowledgements.....	28
References	28
LAMPIRAN.....	32



UMN
 UNIVERSITAS
 MULTIMEDIA
 NUSANTARA

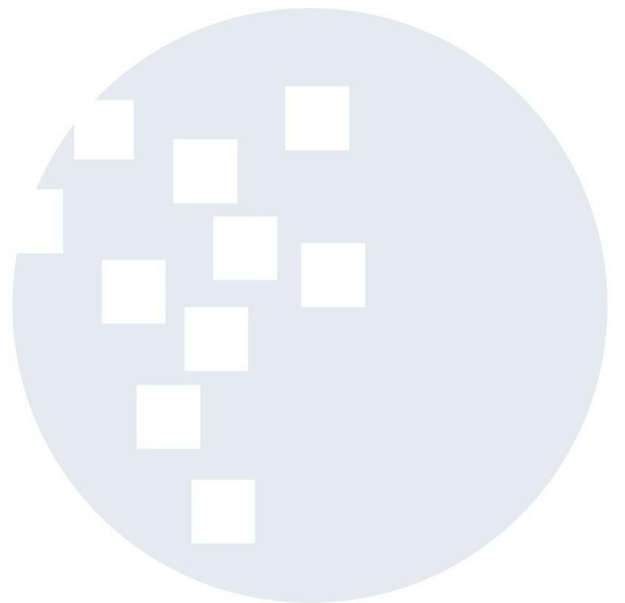
DAFTAR TABEL

Table 1. Dataset Sample - First Row (Wrapped Across 3 lines for Readability).	16
Table 2. Dataset Description	16
Table 3. Classifier Contribution in Hybrid Models.....	20
Table 4. Hyperparameter Tuning	20
Table 5. Performance Evaluation of Classifiers with Different Oversampling Techniques	24
Table 6. Hyperparameter Settings and Selected Features	25
Table 7. Optimal Models with and Without Oversampling	25
Table 8. Comparison of Previous Studies and Proposed Models on the Same Dataset.....	26



DAFTAR GAMBAR

Fig. 1. Experimental Flow	15
Fig. 2. Class Distribution	17
Fig. 3. Age Distribution	18
Fig. 4. Data Distribution for Numerical Attribute	22
Fig. 5. Data Distribution for Binary Class	23



UMN
UNIVERSITAS
MULTIMEDIA
NUSANTARA

DAFTAR LAMPIRAN

Lampiran A Turnitin Similarity Report	32
Lampiran B Form Konsultasi Bimbingan.....	36
Lampiran C LOA	37

