

**KOMPARASI PERFORMA ALGORITMA MACHINE
LEARNING PADA DATASET DIABETES DENGAN
PENANGANAN IMBALANCED DATA**

SKRIPSI

Diajukan untuk memenuhi salah satu syarat mencapai derajat Sarjana
Program Studi S1 Informatika



disusun oleh

IQBAL FARIS IHSANI

22.11.4850

Kepada

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

**KOMPARASI PERFORMA ALGORITMA MACHINE
LEARNING PADA DATASET DIABETES DENGAN
PENANGANAN IMBALANCED DATA**

SKRIPSI

untuk memenuhi salah satu syarat mencapai derajat Sarjana
Program Studi S1 Informatika



disusun oleh

IQBAL FARIS IHSANI

22.11.4850

Kepada

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

HALAMAN PERSETUJUAN

SKRIPSI

**KOMPARASI PERFORMA ALGORITMA MACHINE LEARNING PADA
DATASET DIABETES DENGAN PENANGANAN IMBALANCED DATA**

yang disusun dan diajukan oleh

Iqbal Faris Ihsani

22.11.4850

telah disetujui oleh Dosen Pembimbing Skripsi
pada tanggal 20 April 2026

Dosen Pembimbing,



Afrig Aminuddin, S.Kom., M.Eng., Ph.D

NIK. 190302351

HALAMAN PENGESAHAN

SKRIPSI

**KOMPARASI PERFORMA ALGORITMA MACHINE LEARNING PADA
DATASET DIABETES DENGAN PENANGANAN IMBALANCED DATA**

yang disusun dan diajukan oleh

Iqbal Faris Ihsani

22.11.4850

Telah dipertahankan di depan Dewan Penguji
pada tanggal 20 April 2026

Susunan Dewan Penguji

Nama Penguji

Tanda Tangan

Rumini, S.Kom., M.Kom.
NIK. 190302246



Ike Verawati, S.Kom., M.Kom.
NIK. 190302237



Afrig Aminuddin, S.Kom., M.Eng., Ph.D.
NIK. 190302351



Skripsi ini telah diterima sebagai salah satu persyaratan
untuk memperoleh gelar Sarjana Komputer
Tanggal 20 April 2026

DEKAN FAKULTAS ILMU KOMPUTER



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

HALAMAN PERNYATAAN KEASLIAN SKRIPSI

Yang bertandatangan di bawah ini,

Nama mahasiswa : Iqbal Faris Ihsani
NIM : 22.11.4850

Menyatakan bahwa Skripsi dengan judul berikut:

Komparasi Performa Algoritma Machine Learning Pada Dataset Diabetes Dengan Penanganan Imbalanced Data

Dosen Pembimbing: Afrig Aminuddin, S.Kom., M.Eng., Ph.D.

1. Karya tulis ini adalah benar-benar ASLI dan BELUM PERNAH diajukan untuk mendapatkan gelar akademik, baik di Universitas AMIKOM Yogyakarta maupun di Perguruan Tinggi lainnya.
2. Karya tulis ini merupakan gagasan, rumusan dan penelitian SAYA sendiri, tanpa bantuan pihak lain kecuali arahan dari Dosen Pembimbing.
3. Dalam karya tulis ini tidak terdapat karya atau pendapat orang lain, 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 lainnya sesuai dengan norma yang berlaku di Perguruan Tinggi.

Yogyakarta, 20 April 2026

Yang Menyatakan,



Iqbal Faris Ihsani

HALAMAN PERSEMBAHAN

Karya ilmiah ini dipersembahkan kepada Allah SWT atas rahmat, kesehatan, dan kekuatan yang senantiasa diberikan sepanjang perjalanan pendidikan hingga penyusunan penelitian ini. Melalui rida dan petunjuk-Nya, setiap proses pembelajaran, tantangan, dan hambatan dapat dilalui dengan penuh ketekunan sehingga penelitian ini dapat diselesaikan dengan baik. Penyertaan-Nya menjadi sumber semangat dan keyakinan dalam setiap langkah yang dijalani.

Persembahan ini juga ditujukan kepada orang tua dan keluarga tercinta atas doa, dukungan, pengorbanan, serta kasih yang tidak pernah berhenti. Kepercayaan, motivasi, dan nilai-nilai yang diberikan menjadi landasan penting dalam membentuk semangat belajar, tanggung jawab, dan ketekunan selama menempuh pendidikan hingga penyelesaian penelitian ini.

Ucapan penghargaan turut dipersembahkan kepada dosen pembimbing, kekasih tercinta di Bandung, dan beberapa sahabat yang telah memberikan arahan, ilmu, dan bimbingan selama proses penelitian. Setiap masukan dan dukungan yang diberikan sangat berarti dalam memperkaya pemahaman serta menyempurnakan penelitian ini.

Semoga karya ini menjadi wujud rasa syukur, tanggung jawab akademik, dan dedikasi terhadap pengembangan ilmu pengetahuan, serta dapat memberikan manfaat bagi sesama dan lingkungan sekitar

KATA PENGANTAR

Puji syukur dipanjatkan ke hadirat Allah SWT atas rahmat, penyertaan, dan karunia-Nya sehingga skripsi yang berjudul "KOMPARASI PERFORMA ALGORITMA MACHINE LEARNING PADA DATASET DIABETES DENGAN PENANGANAN IMBALANCED DATA" dapat diselesaikan dengan baik. Skripsi ini disusun sebagai salah satu syarat untuk memperoleh gelar sarjana serta sebagai bentuk penerapan ilmu pengetahuan yang telah diperoleh selama masa perkuliahan. Proses penyusunan skripsi ini memberikan pengalaman berharga dalam mengembangkan kemampuan analisis, penelitian, serta pemahaman terhadap penerapan teknologi *machine learning* pada bidang medis.

Dalam proses penyusunan skripsi ini, penulis memperoleh banyak bantuan, arahan, serta dukungan dari berbagai pihak. Oleh karena itu, pada kesempatan ini penulis menyampaikan ucapan terima kasih kepada:

1. Prof. Dr. M. Suyanto, M.M., selaku Rektor yang telah memberikan kesempatan dan fasilitas sehingga proses pendidikan dapat berlangsung dengan baik.
2. Afrig Aminuddin, S.Kom., M.Eng., Ph.D., selaku dosen pembimbing yang telah meluangkan waktu, tenaga, dan pikiran dalam memberikan bimbingan, arahan, serta masukan selama proses penelitian dan penyusunan skripsi.
3. Tim dosen penguji yang telah memberikan saran dan evaluasi demi penyempurnaan skripsi ini.
4. Seluruh dosen dan staf akademik yang telah memberikan ilmu, bimbingan, serta pelayanan selama masa perkuliahan.
5. Orang tua dan keluarga tercinta yang senantiasa memberikan doa, dukungan, dan motivasi sehingga penulis dapat menyelesaikan skripsi ini.
6. Seseorang tercinta yang tidak bisa penulis sebutkan namanya telah senantiasa memberikan dukungan moral, pengertian, dan semangat yang tiada henti selama proses penyusunan penelitian ini.

7. Kepada Muh. Ilham Nadzirullah, Muhammad Arung Aditya Sahada, dan Rizki Fajar Putra Satria, yang telah menemani perjalanan saya sejak awal semester hingga titik ini. Terima kasih atas tawa, diskusi, dan solidaritas yang membuat masa perkuliahan ini menjadi jauh lebih bermakna.
8. Semua pihak yang tidak dapat penulis sebutkan satu per satu, yang telah banyak membantu baik secara langsung maupun tidak langsung dalam penyelesaian skripsi ini.

Penulis menyadari bahwa skripsi ini masih memiliki keterbatasan, baik dari segi penyajian maupun isi. Oleh karena itu, kritik dan saran yang membangun sangat diharapkan demi perbaikan dan pengembangan penelitian di masa mendatang. Semoga skripsi ini dapat memberikan manfaat bagi pengembangan ilmu pengetahuan, khususnya di bidang kecerdasan buatan dan pengolahan data medis, serta menjadi referensi bagi penelitian selanjutnya.

Yogyakarta, 20 April 2026

Penulis

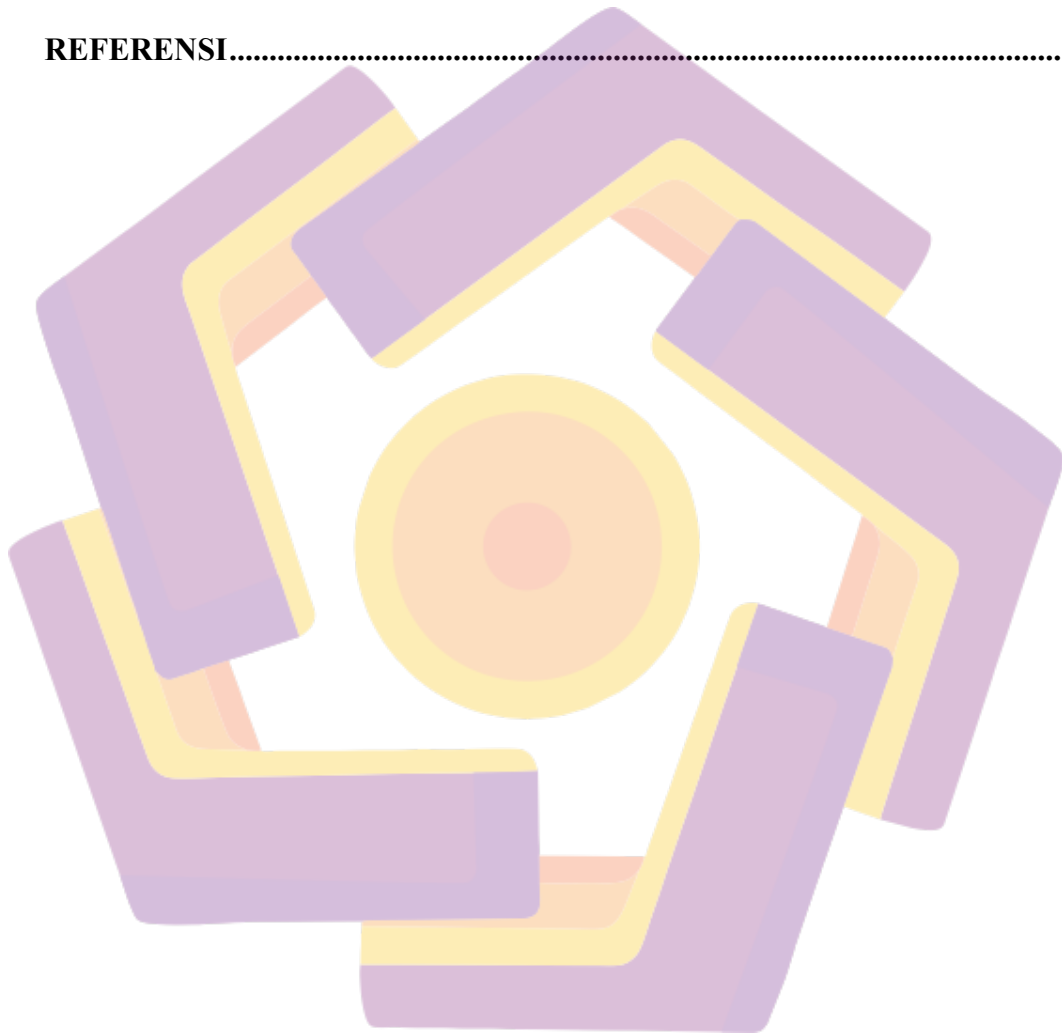
DAFTAR ISI

HALAMAN JUDUL	i
HALAMAN PERSETUJUAN	ii
HALAMAN PENGESAHAN.....	iii
HALAMAN PERNYATAAN KEASLIAN SKRIPSI.....	iv
HALAMAN PERSEMBAHAN	v
KATA PENGANTAR.....	vi
DAFTAR ISI.....	viii
DAFTAR TABLE	xii
DAFTAR GAMBAR.....	xiii
DAFTAR LAMBANG DAN SINGKATAN	xv
DAFTAR ISTILAH	xvii
INTISARI	xx
ABSTRACT.....	xxi
BAB I PENDAHULUAN.....	1
1.1 Latar Belakang	1
1.2 Rumusan Masalah	2
1.3 Batasan Masalah	2
1.4 Tujuan Penelitian	2
1.5 Manfaat Penelitian	3
1.6 Sistematika Penulisan	3

BAB II TINJAUAN PUSTAKA	5
2.1 Studi Literatur	5
2.2 Dasar Teori.....	9
2.2.1 Penyakit Diabetes Melitus	9
2.2.2 <i>Machine Learning</i> dan Klasifikasi	9
2.2.3 Ketidakseimbangan Data (<i>Imbalanced Data</i>)	10
2.2.4 Synthetic Minority Over-sampling Technique (SMOTE)	11
2.2.5 Algoritma Klasifikasi <i>Ensemble Tree</i>	13
2.2.6 <i>Confusion Matrix</i> dan Metrik Evaluasi	17
BAB III METODE PENELITIAN	20
3.1 Objek Penelitian.....	20
3.2 Alur Penelitian	21
3.2.1 Pengumpulan dataset	22
3.2.2 Exploratory Data Analysis (EDA)	22
3.2.2.1 Analisis Distribusi Kelas	22
3.2.2.2 Analisis Statistik Deskriptif.....	22
3.2.2.3 Identifikasi Nilai Tidak Logis	22
3.2.3 Preprocessing Data	23
3.2.3.1 Penanganan Missing Value	23
3.2.3.2 Pembagian Data	23
3.2.3.3 Penanganan <i>Imbalance Data</i>	23
3.2.4 Pelatihan Model.....	24
3.2.5 Hyperparameter Tuning.....	25
3.2.6 Optimasi Thershold	25
3.2.7 Evaluasi dan Perbandingan Performa	25
3.3 Alat dan Bahan.....	25
3.3.1 Alat/Instrumen.....	26
BAB IV HASIL DAN PEMBAHASAN	28

4.1	Pengumpulan Data	28
4.2	Tahap Exploratory Data Analysis (EDA)	29
4.2.1	Hasil Analisis Distribusi Kelas	29
4.2.2	Analisis Statistik Deskriptif	30
4.2.3	Identifikasi Nilai Tidak Logis	31
4.3	Tahap Preprocessing Data.....	33
4.3.1	Penanganan Missing Value	33
4.3.2	Pembagian Data	34
4.3.3	Penanganan Imbalanced Data	35
4.4	Pelatihan Model	36
4.4.1	XGBoost.....	37
4.4.2	LightGBM.....	39
4.4.3	CatBoost.....	42
4.5	Optimisasi Threshold	44
4.6	Evaluasi Model	46
4.6.1	XGBoost.....	46
4.6.2	LightGBM.....	49
4.6.3	CatBoost.....	51
4.7	Analisis Perbandingan Performa.....	54
4.7.1	Analisis Accuracy	55
4.7.2	Analisis Precision	55
4.7.3	Analisis Recall	55
4.7.4	Analisis F1-Score	56
4.7.5	Analisis ROC-AUC.....	56
4.7.6	Confusion Matrix Analysis.....	56
4.8	Perbandingan Dengan Penelitian Terdahulu.....	61
4.9	Pembahasan terhadap Hasil dan Tujuan Penelitian	62

4.10	Implementasi Model (Deployment)	63
BAB V PENUTUP.....		65
5.1	Kesimpulan	65
5.2	Saran	66
REFERENSI.....		68



DAFTAR TABLE

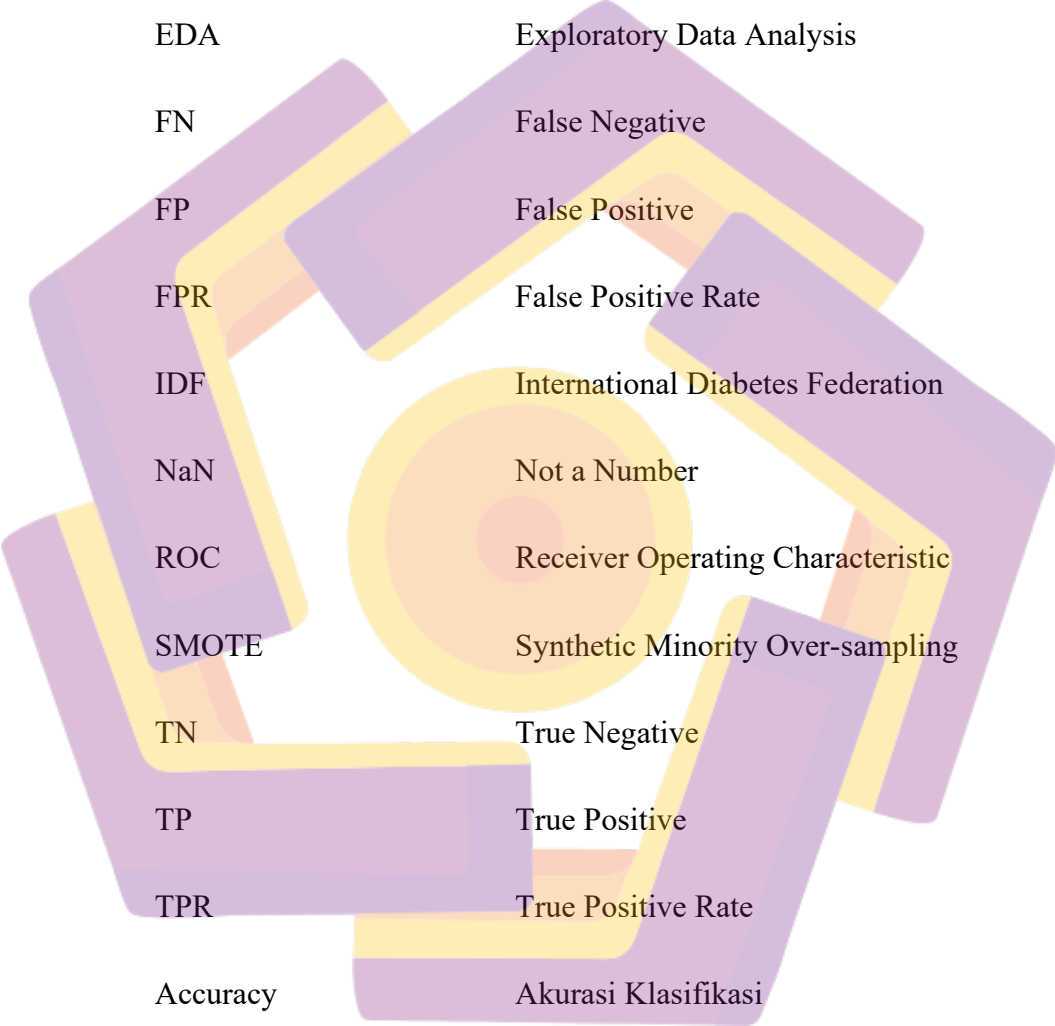
Tabel 2. 1 Keaslian Penelitian	7
Tabel 3. 1 Spesifikasi Perangkat Keras.....	26
Tabel 3. 2 Spesifikasi Perangkat Lunak.....	27
Tabel 4. 1 Hasil Pembagian data 90:10	35
Tabel 4. 2 Hasil Parameter Terbaik XGBoost	39
Tabel 4. 3 Hasil Parameter Terbaik LightGBM.....	41
Tabel 4. 4 Hasil Parameter Terbaik CatBoost.....	44
Tabel 4. 5 Hasil Optimisasi Terbaik Tanpa Smote	45
Tabel 4. 6 Hasil Optimisasi Terbaik dengan Smote.....	46
Tabel 4. 7 Hasil Evaluasi XGBoost Tanpa Smote.....	47
Tabel 4. 8 Hasil Evaluasi XGBoost dengan Smote	48
Tabel 4. 9 Hasil Evaluasi LightGBM tanpa Smote.....	49
Tabel 4. 10 Hasil Evaluasi LightGBM dengan Smote.....	50
Tabel 4. 11 Hasil Evaluasi CatBoost tanpa Smote.....	52
Tabel 4. 12 Hasil Evaluasi CatBoost dengan Smote.....	53
Tabel 4. 13 Perbandingan Performa Model tanpa Smote	54
Tabel 4. 14 Perbandingan Performa Model dengan Smote	55
Tabel 4. 15 Perbandingan Dengan Penelitian Terdahulu.....	61

DAFTAR GAMBAR

Gambar 2. 1 <i>Imbalance Data</i>	10
Gambar 2. 2 <i>Synthetic Minority Over-sampling Technique</i>	12
Gambar 2. 3 Boosting Tree	13
Gambar 2. 4 <i>Leaf-wise tree growth</i>	15
Gambar 2. 5 Struktur Alur CatBoost	16
Gambar 3. 1 Alur Penelitian	21
Gambar 4. 1 Dataset	28
Gambar 4. 2 Tahapan Distribusi Kelas	29
Gambar 4. 3 Hasil Distribusi Kelas	30
Gambar 4. 4 Hasil Statistik Deskriptif Dataset	31
Gambar 4. 5 Tahapan Identifikasi Nilai Tidak Logis	32
Gambar 4. 6 Hasil Penanganan Nilai Tidak Logis	32
Gambar 4. 7 Tampilan Dataset Setelah Penanganan Nilai Tidak Logis	33
Gambar 4. 8 Tahapan Pengecekan Missing Value	33
Gambar 4. 9 Tahapan Penanganan Missing Value	34
Gambar 4. 10 Hasil Penanganan Missing Value	34
Gambar 4. 11 Tahapan Pembagian data 90:10	35
Gambar 4. 12 Tahapan Penanganan Imbalance Data	35
Gambar 4. 13 Hasil Penanganan Imbalance Data	36
Gambar 4. 14 Visualisasi Hasil Distribusi Kelas Sebelum dan Sesudah SMOTE	36
Gambar 4. 15 Proses Hyperparameter Tuning XGBoost Menggunakan GridSearchCV (Tanpa SMOTE).....	38
Gambar 4. 16 Tahapan Pelatihan Model XGBoost dan Parameter Tuning	38
Gambar 4. 17 Proses Hyperparameter Tuning LightGBM Menggunakan GridSearchCV (Tanpa SMOTE).....	40
Gambar 4. 18 Tahapan Pelatihan Model LightGBM dan Parameter Tuning	41
Gambar 4. 19 Proses Hyperparameter Tuning CatBoost Menggunakan GridSearchCV (Tanpa SMOTE).....	43
Gambar 4. 20 Tahapan Pelatihan Model CatBoost dan Parameter Tuning	43

Gambar 4. 21 Optimisasi Threshold	45
Gambar 4. 22 Classification Report XGBoost tanpa Smote	47
Gambar 4. 23 Classification Report XGBoost dengan Smote	48
Gambar 4. 24 Classification Report LightGBM tanpa Smote	49
Gambar 4. 25 Classification Report LightGBM dengan Smote	50
Gambar 4. 26 Classification Report CatBoost tanpa Smote	52
Gambar 4. 27 Classification Report CatBoost dengan Smote	53
Gambar 4. 28 Confusion Matrix XGBoost tanpa Smote	57
Gambar 4. 29 Confusion Matrix XGBoost dengan Smote	58
Gambar 4. 30 Confusion Matrix LightGBM tanpa Smote	58
Gambar 4. 31 Confusion Matrix LightGBM dengan Smote	59
Gambar 4. 32 Confusion Matrix CatBoost tanpa Smote	60
Gambar 4. 33 Confusion Matrix CatBoost dengan Smote	60
Gambar 4. 34 Implementasi Model (Deploy)	64

DAFTAR LAMBANG DAN SINGKATAN



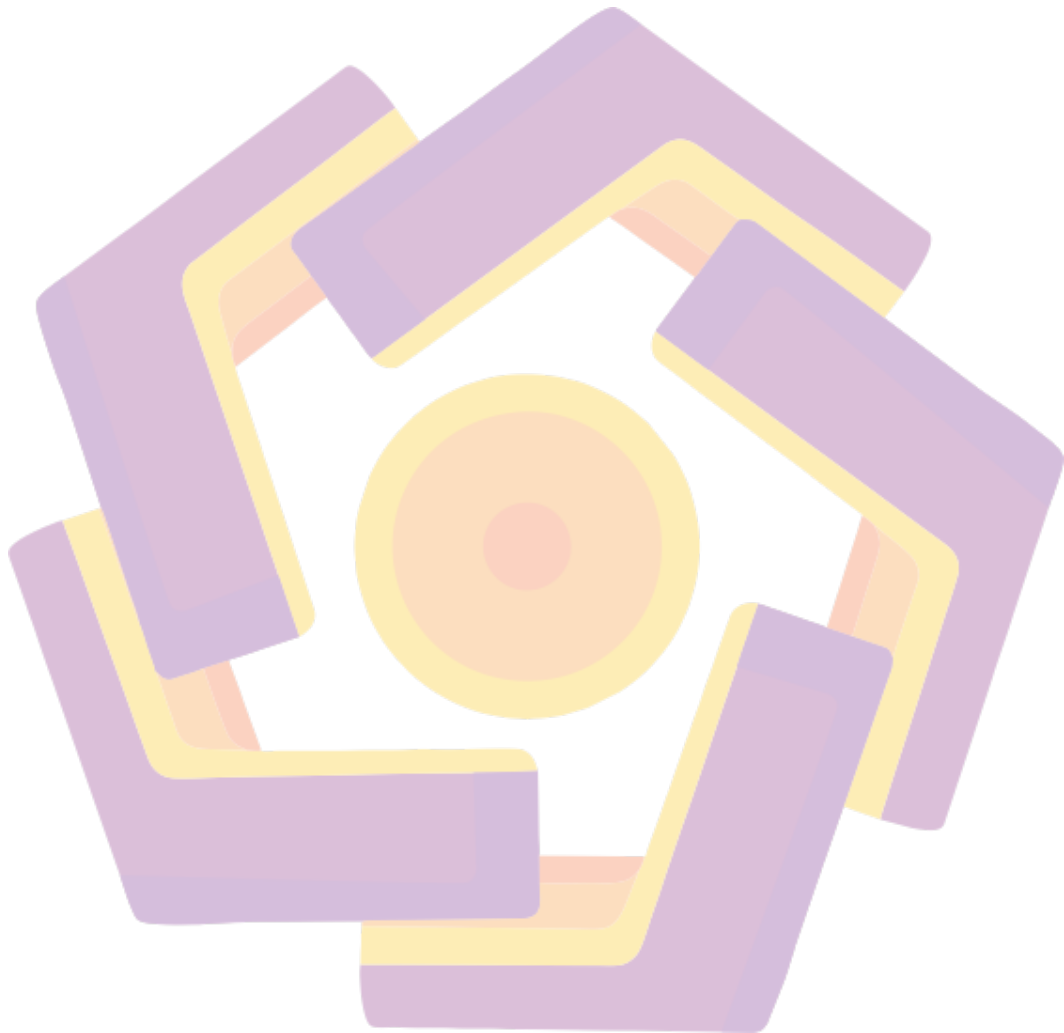
AUC	Area Under Curve
BMI	Body Mass Index
EDA	Exploratory Data Analysis
FN	False Negative
FP	False Positive
FPR	False Positive Rate
IDF	International Diabetes Federation
NaN	Not a Number
ROC	Receiver Operating Characteristic
SMOTE	Synthetic Minority Over-sampling
TN	True Negative
TP	True Positive
TPR	True Positive Rate
Accuracy	Akurasi Klasifikasi
ML	Machine Learning
GPU	Graphics Processing Unit
CPU	Central Processing Unit
CSV	Comma-Separated Values

Test Set

Data uji

Training Set

Data latih



DAFTAR ISTILAH

Accuracy

Ukuran rasio prediksi benar dari seluruh data yang diuji, yang menunjukkan tingkat ketepatan model dalam melakukan klasifikasi secara keseluruhan

CatBoost

Algoritma *gradient boosting* pada *decision trees* yang memiliki keunggulan dalam menangani fitur kategorikal secara langsung menggunakan teknik *ordered boosting*.

Confusion Matrix

Tabel yang digunakan untuk menggambarkan performa model klasifikasi berdasarkan jumlah prediksi benar dan salah pada masing-masing kelas.

Diabetes Melitus

Gangguan metabolisme kronis yang ditandai dengan tingginya kadar gula darah (hiperglikemia) seiring waktu akibat pankreas tidak memproduksi cukup insulin.

Ensemble Tree

Teknik membangun model secara bertahap dengan menggabungkan banyak model pembelajaran yang lemah (*weak learners*) menjadi satu model prediktif yang kuat.

F1-Score

Nilai rata-rata harmonis antara *Precision* dan *Recall* yang berguna untuk mengevaluasi model pada kondisi data yang tidak seimbang.

Imbalanced Data

Kondisi pada dataset di mana jumlah *instance* atau contoh data untuk masing-masing kelas tidak seimbang secara signifikan.

LightGBM

Algoritma *gradient boosting* yang menggunakan strategi konstruksi pohon secara *leaf-wise* untuk mempercepat pelatihan dan efisiensi memori.

Machine Learning

Cabang dari kecerdasan buatan (*Artificial Intelligence*) yang memungkinkan sistem komputer untuk belajar dari data dan meningkatkan kinerjanya tanpa diprogram secara eksplisit.

Precision

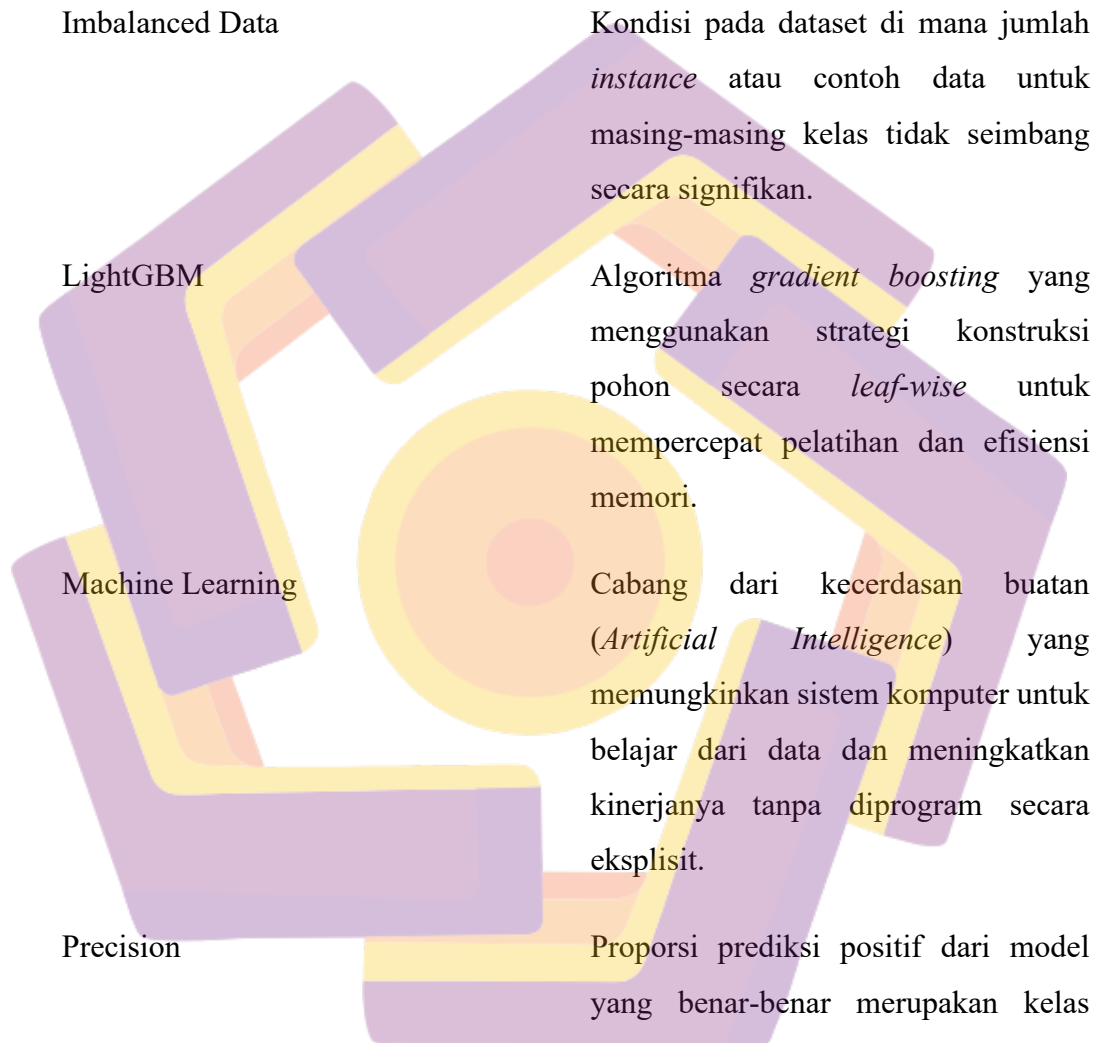
Proporsi prediksi positif dari model yang benar-benar merupakan kelas positif yang sebenarnya.

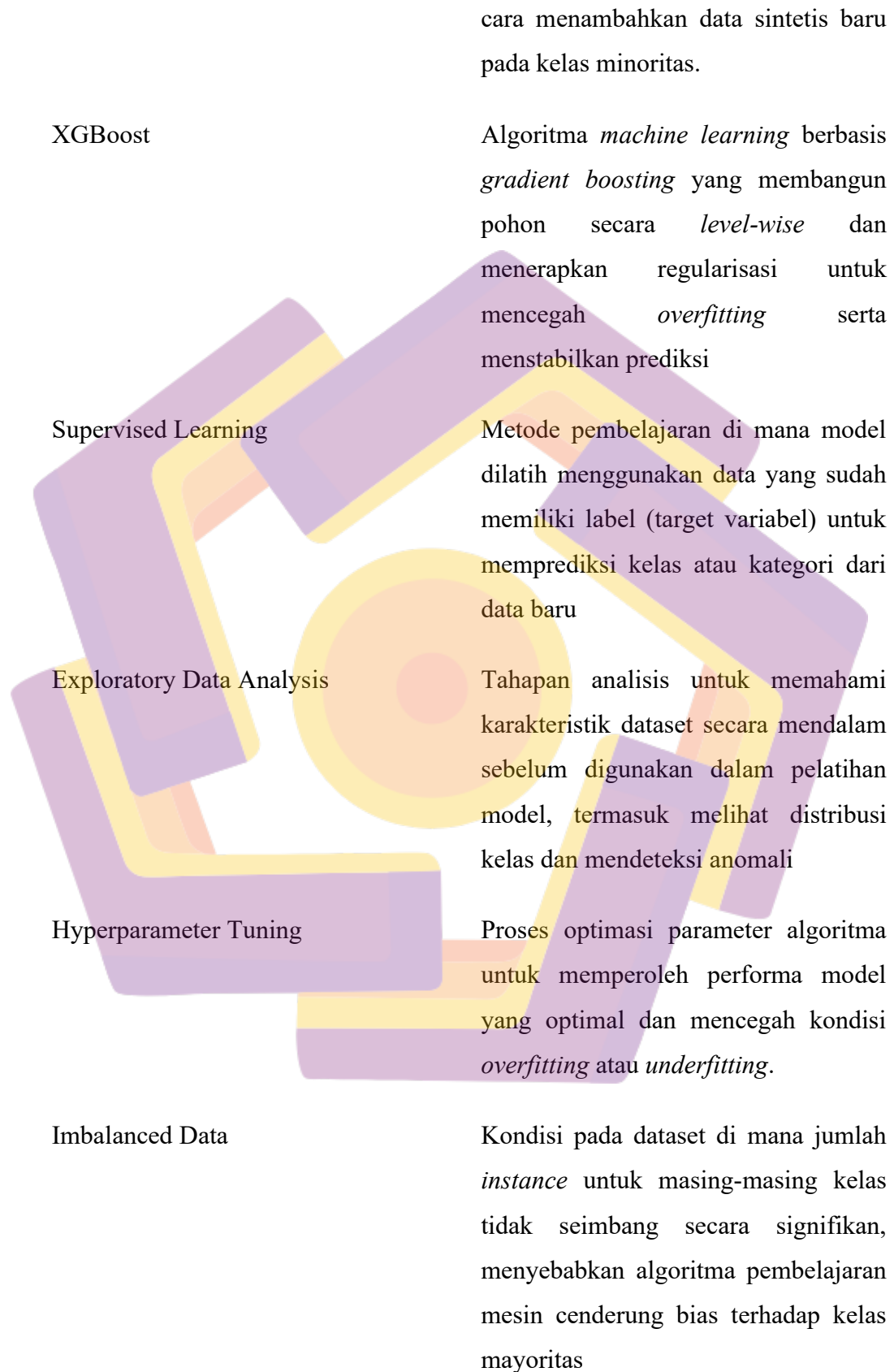
Recall

Proporsi dari kelas positif sebenarnya yang berhasil diprediksi secara tepat oleh model klasifikasi.

SMOTE

Metode *oversampling* untuk menangani data tidak seimbang dengan





INTISARI

Penyakit diabetes melitus merupakan gangguan metabolisme kronis yang memerlukan deteksi dini. Pemanfaatan machine learning pada data rekam medis sangat berpotensi untuk membantu diagnosis, namun sering terkendala oleh ketidakseimbangan kelas (imbalanced data) yang memicu bias pada prediksi model. Penelitian ini bertujuan membandingkan performa algoritma klasifikasi ensemble tree (XGBoost, LightGBM, dan CatBoost) serta menguji efektivitas Synthetic Minority Over-sampling Technique (SMOTE) dalam mengatasi masalah ketidakseimbangan data tersebut.

Eksperimen menggunakan Pima Indians Diabetes Dataset dengan menerapkan tahapan prapemrosesan seperti imputasi missing value, pembagian data latih dan uji (90:10), serta penerapan SMOTE secara eksklusif pada data latih. Model dioptimalkan melalui hyperparameter tuning dan dievaluasi menggunakan metrik accuracy, precision, recall, F1-score, serta ROC-AUC.

Hasil penelitian menunjukkan bahwa SMOTE efektif mengatasi bias prediksi terhadap kelas mayoritas (pasien sehat). Berdasarkan komparasi, algoritma XGBoost memberikan performa paling superior dan seimbang dengan capaian accuracy 85,71%, precision 75,00%, recall 88,88%, F1-Score 81,35%, dan ROC-AUC 87,85%. Secara keseluruhan, XGBoost direkomendasikan sebagai arsitektur terbaik untuk pengembangan sistem diagnosis dini risiko diabetes yang andal secara klinis

Kata kunci: Diabetes, *Machine Learning*, *Imbalanced Data*, *SMOTE*, *Ensemble Tree*

ABSTRACT

Diabetes mellitus is a chronic metabolic disorder characterized by high blood sugar levels, and its early detection is crucial to prevent fatal complications. The utilization of machine learning technology has great potential in diagnosing diabetes indications based on patients' clinical medical records. However, a major challenge in medical datasets is class imbalance (imbalanced data), which can cause classification models to be biased and tend to predict the majority class more often. This study aims to analyze and compare the performance of ensemble tree classification algorithms, namely XGBoost, LightGBM, and CatBoost, and to test the effectiveness of the Synthetic Minority Over-sampling Technique (SMOTE) method in handling the class imbalance issue

The experiment utilized the Pima Indians Diabetes Dataset by applying preprocessing stages such as missing value imputation, a 90:10 train-test data split, and the exclusive application of SMOTE on the training data. The models were optimized through hyperparameter tuning and evaluated using accuracy, precision, recall, F1-score, and ROC-AUC metrics

The results showed that SMOTE effectively overcame prediction bias towards the majority class (healthy patients). Based on the comparison, the XGBoost algorithm provided the most superior and balanced performance, achieving an accuracy of 85.71%, precision of 75.00%, recall of 88.88%, F1-Score of 81.35%, and ROC-AUC of 87.85%. Overall, XGBoost is recommended as the best architecture for developing a clinically reliable early diagnosis system for diabetes risk

Keywords: *Diabetes, Machine Learning, Imbalanced Data, SMOTE, Ensemble Tree*