TUGAS AKHIR MATA KULIAH DATA MINING

LAPORAN PENELITIAN

KLASIFIKASI PENYAKIT DIABETES BERDASARKAN DATA PASIEN MENGGUNAKAN ALGORITMA GAUSSIAN NAIVE BAYES DENGAN PENYEIMBANGAN DATA MELALUI METODE SMOTE

KELAS A

Anggota Kelompok:

Ariski Ade Raharjo	232410101015
Rendy Nayogi Pramudya	232410101016
M. Raihan Rabbani	232410101059
Rahmat Fauzul Akbar	232410101063

PROGRAM STUDI SISTEM INFORMASI

FAKULTAS ILMU KOMPUTER

UNIVERSITAS JEMBER

2025

DAFTAR ISI

BAB I
PENDAHULUAN

1.1. Latar Belakang

Diabetes merupakan salah satu penyakit kronis yang memengaruhi kemampuan tubuh dalam mengelola kadar gula darah. Penyakit ini telah menjadi perhatian serius secara global karena berdampak besar terhadap kualitas hidup individu dan angka harapan hidup masyarakat. Menurut berbagai laporan kesehatan, jumlah penderita diabetes terus meningkat setiap tahunnya, menjadikannya salah satu penyebab kematian terbesar di dunia. Dalam konteks pencegahan dan penanganan dini, deteksi serta klasifikasi kondisi diabetes sejak awal sangat penting untuk mengurangi risiko komplikasi jangka panjang.

Namun, dalam pengolahan data medis untuk klasifikasi diabetes, sering ditemui permasalahan ketidakseimbangan kelas di mana jumlah data pasien non-diabetes jauh lebih banyak dibandingkan data pasien yang terdiagnosis diabetes. Hal ini dapat memengaruhi kinerja model klasifikasi yang dikembangkan. Untuk mengatasi hal tersebut, metode Synthetic Minority Over-sampling Technique (SMOTE) digunakan untuk menyeimbangkan distribusi data antar kelas. Di sisi lain, algoritma Gaussian Naive Bayes dikenal memiliki performa yang baik dalam klasifikasi data medis, terutama pada kasus yang melibatkan asumsi distribusi probabilitas. Oleh karena itu, penerapan kombinasi SMOTE dan Gaussian Naive Bayes diharapkan dapat meningkatkan akurasi klasifikasi data diabetes dan mendukung upaya deteksi dini secara lebih efektif.

1.2. Rumusan Masalah

Berdasarkan latar belakang di atas, diperoleh rumusan masalah sebagai berikut:

1. Bagaimana penerapan algoritma Gaussian Naive Bayes dalam mengklasifikasikan penyakit diabetes berdasarkan data pasien?
2. Bagaimana metode SMOTE dapat digunakan untuk menyeimbangkan distribusi data pada dataset diabetes?
3. Bagaimana perbandingan hasil evaluasi klasifikasi sebelum dan sesudah penerapan SMOTE terhadap dataset diabetes?
4. Membandingkan tingkat akurasi antara algoritma Gaussian Naive Bayes dan Random Forest dalam mendiagnosis penyakit diabetes berdasarkan parameter evaluasi seperti akurasi, presisi, recall, dan F1-score.

1.3. Tujuan Penelitian

Berdasarkan rumusan masalah di atas, maka diperoleh tujuan penelitian sebagai berikut:

1. Mengimplementasikan algoritma Gaussian Naive Bayes untuk melakukan klasifikasi penyakit diabetes berdasarkan data pasien.
2. Menerapkan metode SMOTE (Synthetic Minority Over-sampling Technique) untuk menyeimbangkan distribusi kelas pada dataset diabetes.
3. Mengevaluasi performa klasifikasi sebelum dan sesudah penyeimbangan data menggunakan metrik evaluasi seperti akurasi, precision, recall, dan F1-score.
4. Membandingkan tingkat akurasi antara algoritma Gaussian Naive Bayes dan Random Forest dalam mendiagnosis penyakit diabetes berdasarkan parameter evaluasi seperti akurasi, presisi, recall, dan F1-score.

1.4. Batasan Masalah

Batasan dalam penelitian adalah sebagai berikut:

1. Penelitian ini hanya mengevaluasi kinerja algoritma *Gaussian Naive Bayes*, baik sebelum maupun sesudah penerapan metode penyeimbangan data menggunakan *SMOTE (Synthetic Minority Over-sampling Technique)*.
2. Meskipun dilakukan perbandingan dengan algoritma *Random Forest*, pembahasan tidak mencakup analisis mendalam terhadap algoritma lain di luar keduanya.
3. Dataset yang digunakan berasal dari platform *Kaggle* dan dibatasi pada dataset yang relevan untuk klasifikasi penyakit diabetes. Dataset ini digunakan baik untuk proses pelatihan maupun pengujian model.
4. Evaluasi performa model hanya difokuskan pada metrik-metrik dasar seperti akurasi, presisi, recall, dan F1-score, tanpa mempertimbangkan metrik lain seperti AUC-ROC atau waktu komputasi.

1.5. Manfaat

Manfaat dari penelitian adalah sebagai berikut :

1. Bagi peneliti

   Menambah pemahaman mengenai implementasi algoritma Gaussian Naive Bayes serta mengetahui tahapan detail penggunaan metode SMOTE dalam menangani data tidak seimbang pada kasus klasifikasi penyakit diabetes.

2. Bagi Akademisi

   Menambah wawasan tentang penerapan algoritma klasifikasi berbasis

probabilistik dan teknik penyeimbangan data, serta menjadi referensi dalam proses pengambilan keputusan berbasis data di bidang kesehatan dan data mining medis.

3. Bagi objek penelitian

   Memberikan gambaran tentang bagaimana model klasifikasi dapat membantu dalam proses deteksi penyakit diabetes, serta mendukung pengambilan keputusan medis yang lebih tepat dengan memanfaatkan machine learning pada data pasien.

BAB II

TINJAUAN PUSTAKA

2.1. Penelitian Terdahulu

Penelitian mengenai algoritma Naive Bayes, khususnya varian Gaussian Naive Bayes, telah banyak digunakan dalam bidang medis dan pengolahan data citra. Algoritma ini dikenal sebagai metode klasifikasi berbasis probabilistik yang sederhana namun efektif dalam menangani data numerik yang diasumsikan berdistribusi normal.

Nisaa et al. (2021) menerapkan algoritma Naive Bayes untuk diagnosis penyakit diabetes menggunakan dataset dari Kaggle. Penelitian ini menunjukkan bahwa dengan melalui tahapan pra-pemrosesan, pengelompokan atribut seperti usia, kelemahan tubuh, poliuria, dan lain-lain, serta evaluasi model menggunakan bahasa pemrograman Python, mereka berhasil mencapai akurasi sebesar 90%, presisi 93%, recall 89%, dan F1-Score 91%, yang menandakan performa model yang sangat baik dalam klasifikasi penyakit diabetes

Selain itu, Mahran et al. (2020) melakukan klasifikasi jenis jamur berdasarkan citra menggunakan algoritma Naive Bayes Gaussian. Dengan melakukan ekstraksi ciri statistik orde pertama (mean, skewness, variance, kurtosis, dan entropy) sebagai fitur input, mereka mencapai akurasi hingga 98.75% menggunakan metode validasi silang. Penelitian ini memperkuat keunggulan Gaussian Naive Bayes dalam menangani data numerik yang kompleks dan beragam

Penelitian-penelitian sebelumnya tersebut menunjukkan bahwa algoritma Gaussian Naive Bayes efektif dalam berbagai bidang, baik dalam pengklasifikasian data medis seperti diabetes maupun analisis citra digital seperti klasifikasi jamur. Hasil akurasi yang tinggi dalam penelitian-penelitian tersebut menjadi dasar yang kuat untuk menerapkan metode ini dalam konteks data tidak seimbang, terutama ketika dikombinasikan dengan teknik penyeimbangan seperti SMOTE.

2.2. Landasan Teori

1. Penyakit Diabetes

   Diabetes adalah penyakit kronis yang ditandai dengan tingginya kadar glukosa dalam darah akibat gangguan produksi atau fungsi hormon insulin. Penyakit ini dapat menimbulkan komplikasi serius jika tidak ditangani dengan baik, termasuk kerusakan pada jantung, ginjal, mata, dan saraf. Oleh karena itu, deteksi dini sangat penting

untuk mencegah komplikasi tersebut. Dalam beberapa penelitian, diabetes juga diklasifikasikan menjadi beberapa tipe seperti diabetes tipe 1, tipe 2, dan kondisi normal. Deteksi dini tidak hanya membantu dalam pencegahan komplikasi, tetapi juga dalam menentukan jenis penanganan yang tepat. Gejala awal yang sering muncul seperti poliuria, polidipsia, dan kelemahan sering kali diabaikan, sehingga diagnosis berbasis data menjadi salah satu alternatif penting untuk mempercepat pengambilan keputusan medis.

2. Data Mining

Data mining merupakan proses penggalian informasi penting dari kumpulan data yang besar dengan menggunakan teknik statistik, kecerdasan buatan, dan pembelajaran mesin. Dalam konteks medis, data mining sangat membantu dalam mendeteksi pola penyakit, prediksi, dan klasifikasi diagnosis berdasarkan data histori. Teknik ini memungkinkan tenaga medis maupun sistem otomatis untuk mengambil keputusan yang lebih tepat berdasarkan data gejala dan hasil pemeriksaan pasien.

Beberapa teknik umum dalam data mining meliputi klasifikasi, prediksi, estimasi, dan clustering. Selain itu, data mining juga mendukung proses pengambilan keputusan dengan membangun model diagnosis berdasarkan dataset besar yang telah dilabeli, seperti data gejala diabetes yang tersedia dari sumber terbuka seperti Kaggle. Penggunaan data mining terbukti meningkatkan efisiensi dalam menganalisis dan memproses data pasien secara cepat dan akurat.

3. Naive Bayes dan Gaussian Naive Bayes

Naive Bayes merupakan algoritma klasifikasi berbasis teorema Bayes yang mengasumsikan independensi antar fitur. Gaussian Naive Bayes adalah varian yang digunakan untuk data numerik dan mengasumsikan bahwa nilai-nilai fitur mengikuti distribusi normal (Gaussian). Model ini sederhana namun efektif dalam pengklasifikasian data medis termasuk diagnosis diabetes. Dalam penerapannya, Naive Bayes menghitung probabilitas kemunculan suatu kelas (misalnya diabetes tipe 2) berdasarkan distribusi data fitur seperti usia, tekanan darah, kadar gula, dan riwayat keluarga.

Berdasarkan hasil penelitian, algoritma ini dapat memberikan hasil klasifikasi dengan akurasi yang cukup tinggi, bahkan mencapai 90% dalam beberapa studi yang menggunakan metode evaluasi holdout (80% data pelatihan dan 20% data pengujian). Selain akurasi, metrik lain seperti presisi, recall, dan F1-score juga menjadi pertimbangan penting untuk menilai efektivitas model.

a. Teorema Bayes (dasar Naive Bayes):

$$\mathbf{P(C|X) = \frac{P(X|C).P(C)}{P(X)}}$$

Keterangan

- P(C | X) : Probabilitas posterior kelas *C* terhadap fitur *X*
- P(X | C) : Probabilitas fitur X muncul dalam kelas C
- P(C): Probabilitas awal (prior) kelas C
- P(X): Probabilitas keseluruhan fitur X

b. Evaluasi Model:

Setelah model dikembangkan, biasanya dilakukan evaluasi menggunakan beberapa metrik:

- Akurasi (Accuracy):

$$Accuracy = \frac{TP + TN}{TP+TN+FP+FN}$$

- Presisi (Precision):

$$Precision = \frac{TP}{TP+FP}$$

- Recall (Sensitivitas):

$$Recall = \frac{TP}{TP+FN}$$

- F1-Score:

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$$

Keterangan :

- TP: True Positive
- TN: True Negative
- FP: False Positive
- FN: False Negative

4. SMOTE (Synthetic Minority Over-sampling Technique)

SMOTE adalah teknik penyeimbangan data yang dibuat untuk mengatasi ketidakseimbangan kelas dalam dataset, yaitu ketika jumlah data pada satu kelas jauh lebih sedikit dibandingkan kelas lainnya. Teknik ini bekerja dengan cara menciptakan sampel sintetik dari data minoritas, yang membantu meningkatkan akurasi klasifikasi

terutama pada dataset medis yang tidak seimbang.

Dengan menerapkan SMOTE sebelum proses pelatihan model klasifikasi, performa algoritma seperti Naive Bayes dapat ditingkatkan secara signifikan. Hal ini karena SMOTE menghasilkan data baru melalui interpolasi nilai-nilai fitur dari sampel minoritas yang ada, bukan sekadar menggandakan data. Teknik ini sangat penting ketika menangani diagnosis penyakit seperti diabetes, di mana jumlah kasus positif lebih sedikit daripada kasus normal. Tanpa penyeimbangan, model akan cenderung bias terhadap kelas mayoritas, yang menyebabkan rendahnya sensitivitas terhadap kasus penyakit yang sebenarnya lebih kritis untuk dideteksi.

5. Random Forest

Random Forest merupakan salah satu algoritma klasifikasi yang termasuk dalam kategori metode ensemble learning, yaitu teknik yang menggabungkan beberapa model pembelajaran untuk menghasilkan prediksi yang lebih akurat dan stabil. Algoritma ini bekerja dengan membentuk sejumlah pohon keputusan (decision trees) pada saat pelatihan, dan hasil klasifikasi ditentukan berdasarkan mayoritas voting dari semua pohon yang terbentuk.

Keunggulan utama dari Random Forest adalah kemampuannya untuk menangani jumlah fitur yang besar, menghindari overfitting, serta tangguh terhadap data yang tidak seimbang apabila dikombinasikan dengan metode seperti SMOTE. Random Forest juga tidak terlalu sensitif terhadap outlier dan noise karena menggunakan rata-rata atau voting dari banyak model. Secara umum, langkah kerja Random Forest meliputi:

a. Mengambil sampel data secara acak (bootstrapping) dari dataset pelatihan.
b. Membangun decision tree untuk masing-masing subset data.
c. Melakukan prediksi pada data uji berdasarkan hasil voting dari semua pohon yang telah dibentuk.

Kelebihan lainnya adalah Random Forest tidak membutuhkan asumsi distribusi data tertentu seperti Gaussian Naive Bayes, sehingga lebih fleksibel dalam mengolah berbagai jenis data, baik numerik maupun kategorikal. Dalam penelitian ini, Random Forest digunakan sebagai pembanding terhadap algoritma Gaussian Naive Bayes untuk melihat performa klasifikasi terhadap data diabetes sebelum dan sesudah dilakukan penyeimbangan menggunakan metode SMOTE.

BAB III
METODE PENELITIAN

3.1 Jenis Penelitian

Jenis penelitian ini menggunakan metode kuantitatif. Metode ini merupakan pendekatan yang menggunakan data numerik (angka). Metode kuantitatif mencakup proses pengumpulan data dari berbagai sumber. Tujuan metode kuantitatif adalah memberikan pendekatan penelitian yang terstruktur dan dapat diukur, yang difokuskan pada pengumpulan serta analisis data berupa angka.

3.2 Objek Penelitian

Objek dalam penelitian ini adalah data pasien terkait kondisi kesehatan yang berhubungan dengan potensi menderita penyakit diabetes. Data tersebut mencakup berbagai variabel seperti tekanan darah, kadar kolesterol, indeks massa tubuh (BMI), kebiasaan merokok, aktivitas fisik, dan faktor risiko lainnya. Objek ini akan digunakan untuk melatih dan menguji algoritma klasifikasi, dengan tujuan memprediksi status diabetes pasien (tidak diabetes, pre-diabetes, atau diabetes) menggunakan pendekatan Gaussian Naive Bayes.

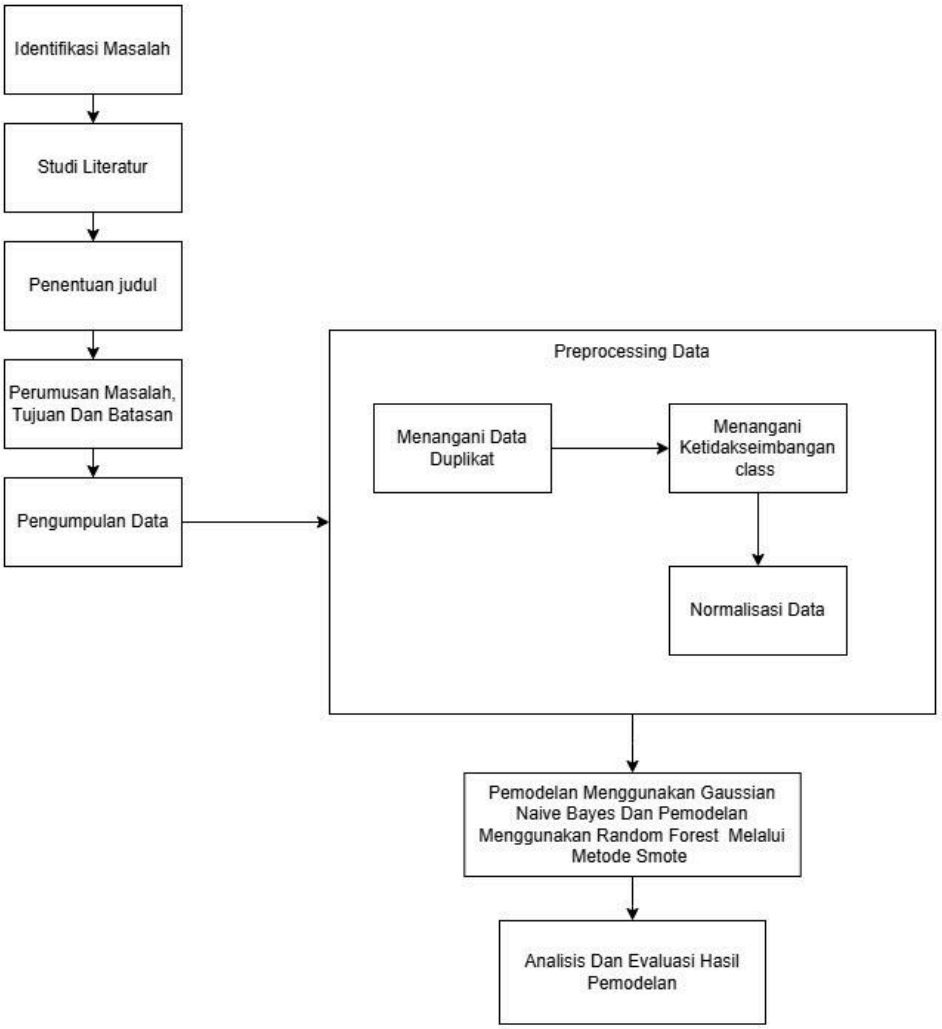
3.3 Waktu dan Tempat Penelitian

Penelitian ini dilaksanakan secara daring (online) selama periode 24 April 2025 hingga 25 Mei 2025. Seluruh proses penelitian, mulai dari pencarian dataset, pengolahan data, pelatihan model, hingga evaluasi hasil dilakukan dengan bantuan perangkat lunak *Jupyter Notebook* dan bahasa pemrograman Python.

3.4 Metode Pengumpulan Data

Metode pengumpulan data yang digunakan adalah data sekunder, yaitu data yang diperoleh bukan secara langsung dari responden, melainkan dari sumber yang telah ada. Dataset yang digunakan dalam penelitian ini diambil dari situs Kaggle (https://www.kaggle.com), yang merupakan salah satu platform penyedia dataset publik untuk keperluan penelitian dan pembelajaran *machine learning*. Dataset tersebut telah melalui proses kurasi dan pembersihan awal oleh pihak penyedia.

3.5 Tahap Penelitian



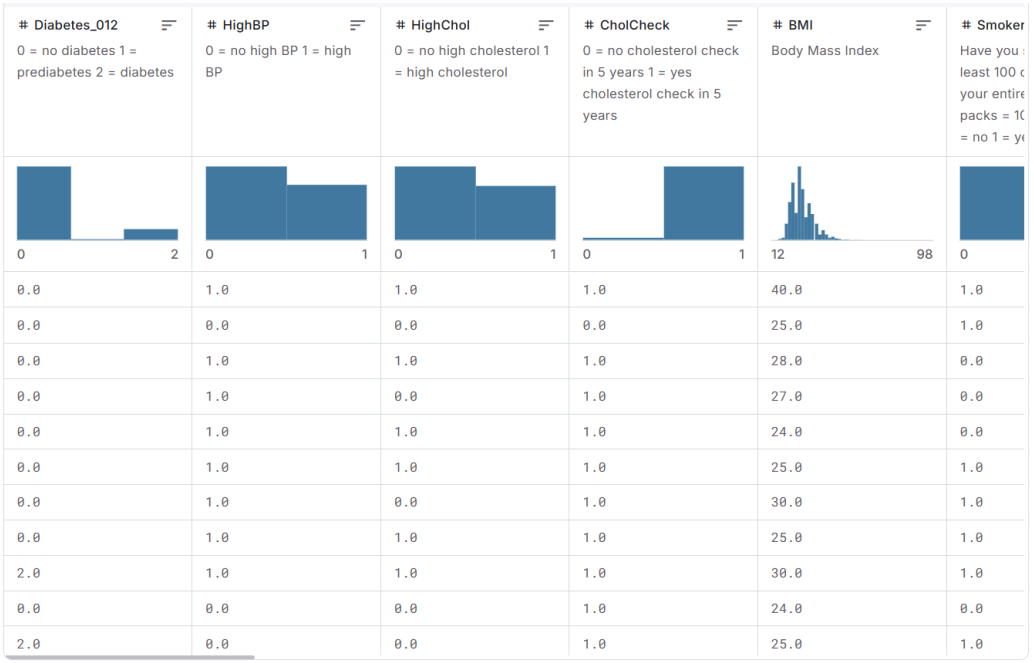
3.6 Hasil Implementasi

1. Atribut

No	Atribut	Keterangan
1.	Diabetes_012	Status diabetes responden: 0 = Tidak diabetes, 1 = Pre-diabetes, 2 = Diabetes. *(Target variabel)*
2.	HighBP	Riwayat tekanan darah tinggi: 0 = Tidak, 1 = Ya.
3.	HighChol	Kolesterol tinggi: 0 = Tidak, 1 = Ya.
4.	CholCheck	Pernah melakukan pemeriksaan kolesterol dalam 5 tahun terakhir: 0 = Tidak, 1 = Ya.
5.	BMI	Body Mass Index (Indeks Massa Tubuh) responden. Nilai numerik.
6.	Smoker	Merokok setidaknya 100 batang dalam hidupnya: 0 = Tidak, 1 = Ya.
7.	Stroke	Pernah mengalami stroke: 0 = Tidak, 1 = Ya.

8.	HeartDiseaseor Attack	Riwayat penyakit jantung koroner atau serangan jantung: 0 = Tidak, 1 = Ya.
9.	PhysActivity	Aktivitas fisik dalam 30 hari terakhir: 0 = Tidak, 1 = Ya.
10.	Fruits	Mengonsumsi buah setidaknya sekali sehari: 0 = Tidak, 1 = Ya.
11.	Veggies	Mengonsumsi sayur setidaknya sekali sehari: 0 = Tidak, 1 = Ya
12.	HvyAlcoholConsump	Konsumsi alkohol berat (lebih dari 14 (pria) atau 7 (wanita) per minggu): 0 = Tidak, 1 = Ya
13.	AnyHealthcare	Mempunyai akses ke layanan kesehatan: 0 = Tidak, 1 = Ya.
14.	NoDocbcCost	Tidak mengunjungi dokter karena alasan biaya: 0 = Tidak, 1 = Ya.
15.	GenHlth	Kesehatan umum: 1 = Sangat baik, 2 = Baik, 3 = Cukup, 4 = Buruk, 5 = Sangat buruk. *(Semakin besar = semakin buruk)*
16.	MentHlth	Jumlah hari mengalami masalah kesehatan mental dalam 30 hari terakhir (0-30)
17.	PhysHlth	Jumlah hari mengalami masalah kesehatan fisik dalam 30 hari terakhir (0-30)
18.	DiffWalk	Kesulitan berjalan atau naik tangga: 0 = Tidak, 1 = Ya.
19.	Sex	Jenis kelamin: 0 = Perempuan, 1 = Laki-laki
20.	Age	Kategori usia (numerik ordinal): 1 = 18–24, 2 = 25–29, ..., 13 = ≥80
21.	Education	Tingkat pendidikan: 1 = Tidak tamat SD, ..., 6 = Lulusan perguruan tinggi
22.	Income	Pendapatan: 1 = < $10.000, ..., 8 = ≥ $75.000.

2. Dataset https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset?resource=download



# Diabetes_012 	
0 = no diabetes 1 = prediabetes 2 = diabetes
	# HighBP
	0 = no high BP 1 = high BP
		# HighChol
		0 = no high cholesterol 1 = high cholesterol
			# CholCheck
			0 = no cholesterol check in 5 years 1 = yes cholesterol check in 5 years
				# BMI
				Body Mass Index
					# Smoker
					Have you least 100 your entire packs = 1 = no 1 = ye

0 2	0 1	0 1	0 1	12 98	0
0.0	1.0	1.0	1.0	40.0	1.0
0.0	0.0	0.0	0.0	25.0	1.0
0.0	1.0	1.0	1.0	28.0	0.0
0.0	1.0	0.0	1.0	27.0	0.0
0.0	1.0	1.0	1.0	24.0	0.0
0.0	1.0	1.0	1.0	25.0	1.0
0.0	1.0	0.0	1.0	30.0	1.0
0.0	1.0	1.0	1.0	25.0	1.0
2.0	1.0	1.0	1.0	30.0	1.0
0.0	0.0	0.0	1.0	24.0	0.0
2.0	0.0	0.0	1.0	25.0	1.0

*Terdapat 22 kolom dan 254 ribu baris

3. Tahapan dan Hasil Perhitungan

Link Google Colab : **A2_DIabetes Dataset | Data Mining**

IMPORT LIBRARY

```
[ ]  import pandas as pd
     import seaborn as sns
     import matplotlib.pyplot as plt
     from sklearn.model_selection import train_test_split
     from imblearn.over_sampling import SMOTE
     from sklearn.naive_bayes import GaussianNB
     from sklearn.metrics import classification_report, confusion_matrix, accuracy_score
```

EXPLANATORY DATA ANALYSIS

1. Overview Data

	Diabetes_012	HighBP	HighChol	CholCheck	BMI	Smoker	Stroke	HeartDiseaseorAttack	PhysActivity	Fruits	...	AnyHealthcare	NoDocbcCost	GenHlth	MentHlth	PhysHlth	DiffWalk	Sex	Age	Education	Income
0	0.0	1.0	1.0	1.0	40.0	1.0	0.0	0.0	0.0	0.0	...	1.0	0.0	5.0	18.0	15.0	1.0	0.0	9.0	4.0	3.0
1	0.0	0.0	0.0	0.0	25.0	1.0	0.0	0.0	1.0	0.0	...	0.0	1.0	3.0	0.0	0.0	0.0	0.0	7.0	6.0	1.0
2	0.0	1.0	1.0	1.0	28.0	0.0	0.0	0.0	0.0	1.0	...	1.0	1.0	5.0	30.0	30.0	1.0	0.0	9.0	4.0	8.0
3	0.0	1.0	0.0	1.0	27.0	0.0	0.0	0.0	1.0	1.0	...	1.0	0.0	2.0	0.0	0.0	0.0	0.0	11.0	3.0	6.0
4	0.0	1.0	1.0	1.0	24.0	0.0	0.0	0.0	1.0	1.0	...	1.0	0.0	2.0	3.0	0.0	0.0	0.0	11.0	5.0	4.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
253675	0.0	1.0	1.0	1.0	45.0	0.0	0.0	0.0	0.0	1.0	...	1.0	0.0	3.0	0.0	5.0	0.0	1.0	5.0	6.0	7.0
253676	2.0	1.0	1.0	1.0	18.0	0.0	0.0	0.0	0.0	0.0	...	1.0	0.0	4.0	0.0	0.0	1.0	0.0	11.0	2.0	4.0
253677	0.0	0.0	0.0	1.0	28.0	0.0	0.0	0.0	1.0	1.0	...	1.0	0.0	1.0	0.0	0.0	0.0	0.0	2.0	5.0	2.0
253678	0.0	1.0	0.0	1.0	23.0	0.0	0.0	0.0	0.0	1.0	...	1.0	0.0	3.0	0.0	0.0	0.0	1.0	7.0	5.0	1.0
253679	2.0	1.0	1.0	1.0	25.0	0.0	0.0	1.0	1.0	1.0	...	1.0	0.0	2.0	0.0	0.0	0.0	0.0	9.0	6.0	2.0

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 253680 entries, 0 to 253679
Data columns (total 22 columns):
 #   Column                Non-Null Count   Dtype
---  ------                --------------   -----
 0   Diabetes_012          253680 non-null  float64
 1   HighBP                253680 non-null  float64
 2   HighChol              253680 non-null  float64
 3   CholCheck             253680 non-null  float64
 4   BMI                   253680 non-null  float64
 5   Smoker                253680 non-null  float64
 6   Stroke                253680 non-null  float64
 7   HeartDiseaseorAttack  253680 non-null  float64
 8   PhysActivity          253680 non-null  float64
 9   Fruits                253680 non-null  float64
 10  Veggies               253680 non-null  float64
 11  HvyAlcoholConsump     253680 non-null  float64
 12  AnyHealthcare         253680 non-null  float64
 13  NoDocbcCost           253680 non-null  float64
 14  GenHlth               253680 non-null  float64
 15  MentHlth              253680 non-null  float64
 16  PhysHlth              253680 non-null  float64
 17  DiffWalk              253680 non-null  float64
 18  Sex                   253680 non-null  float64
 19  Age                   253680 non-null  float64
 20  Education             253680 non-null  float64
 21  Income                253680 non-null  float64
dtypes: float64(22)
memory usage: 42.6 MB
```

2. Pengecekan Data Duplikat

```
data.duplicated().sum()
```

```
np.int64(23899)
```

3. Overview Kelas dalam Dataset

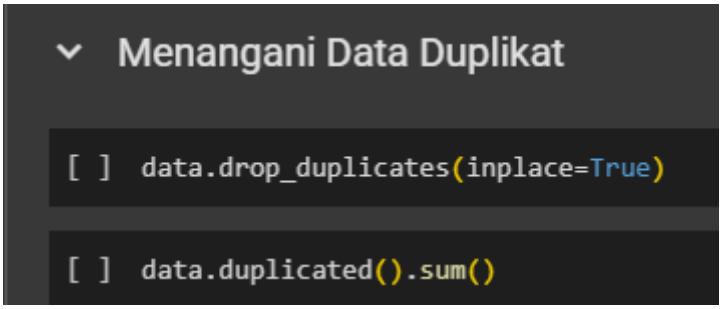
```
data.Diabetes_012.value_counts()
```

Diabetes_012	count
0.0	213703
2.0	35346
1.0	4631

dtype: int64

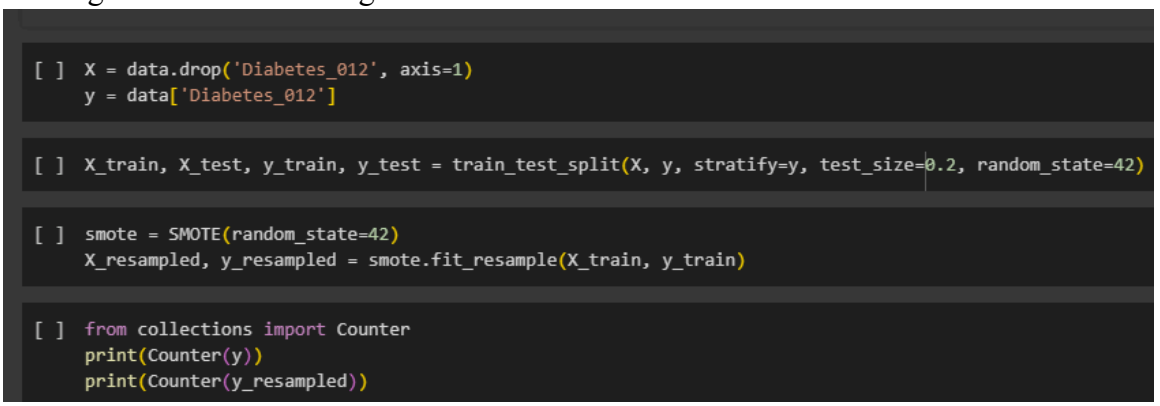
PREPROCESSING DATA

1. Menghapus data duplikat




```
np.int64(0)
```

2. Menangani Ketidakseimbangan Class




```
Counter({0.0: 190055, 2.0: 35097, 1.0: 4629})
Counter({0.0: 152043, 2.0: 152043, 1.0: 152043})
```

MODELING

1. Sebelum Penyeimbangan Class Data dengan SMOTE

Sebelum SMOTE

```
[18] X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

[19] from sklearn.preprocessing import StandardScaler
     scaler = StandardScaler()
     X_train = scaler.fit_transform(X_train)
     X_test = scaler.transform(X_test)
```

1.1. Naive Bayes Gaussian

Naive Bayes Gaussian

```
[20] model = GaussianNB()
     model.fit(X_train, y_train)
```

```
GaussianNB
GaussianNB()
```

```
[21] y_pred = model.predict(X_test)
```

```
print("Confusion Matrix:\n", confusion_matrix(y_test, y_pred))
print("\nClassification Report:\n", classification_report(y_test, y_pred))
print("Akurasi:", accuracy_score(y_test, y_pred))
```

```
Confusion Matrix:
 [[30297   100  7719]
 [  525     7   374]
 [ 2935    31  3969]]

Classification Report:
              precision    recall  f1-score   support

         0.0       0.90      0.79      0.84     38116
         1.0       0.05      0.01      0.01       906
         2.0       0.33      0.57      0.42      6935

    accuracy                           0.75     45957
   macro avg       0.43      0.46      0.42     45957
weighted avg       0.80      0.75      0.76     45957

Akurasi: 0.7457623430598168
```

1.2. Random Forest Classifier

Random Forest

```
[23] model = RandomForestClassifier(n_estimators=100, random_state=42)
     model.fit(X_train, y_train)
```

```
RandomForestClassifier(random_state=42)
```

```
[24] y_pred = model.predict(X_test)
```

```
[25] print("Confusion Matrix:\n", confusion_matrix(y_test, y_pred))
     print("\nClassification Report:\n", classification_report(y_test, y_pred))
     print("Akurasi:", accuracy_score(y_test, y_pred))
```

```
Confusion Matrix:
 [[36612    55  1449]
 [  810     0    96]
 [ 5600    11  1324]]

Classification Report:
              precision    recall  f1-score   support

         0.0       0.85      0.96      0.90     38116
         1.0       0.00      0.00      0.00       906
         2.0       0.46      0.19      0.27      6935

    accuracy                           0.83     45957
   macro avg       0.44      0.38      0.39     45957
weighted avg       0.78      0.83      0.79     45957

Akurasi: 0.8254672846356377
```

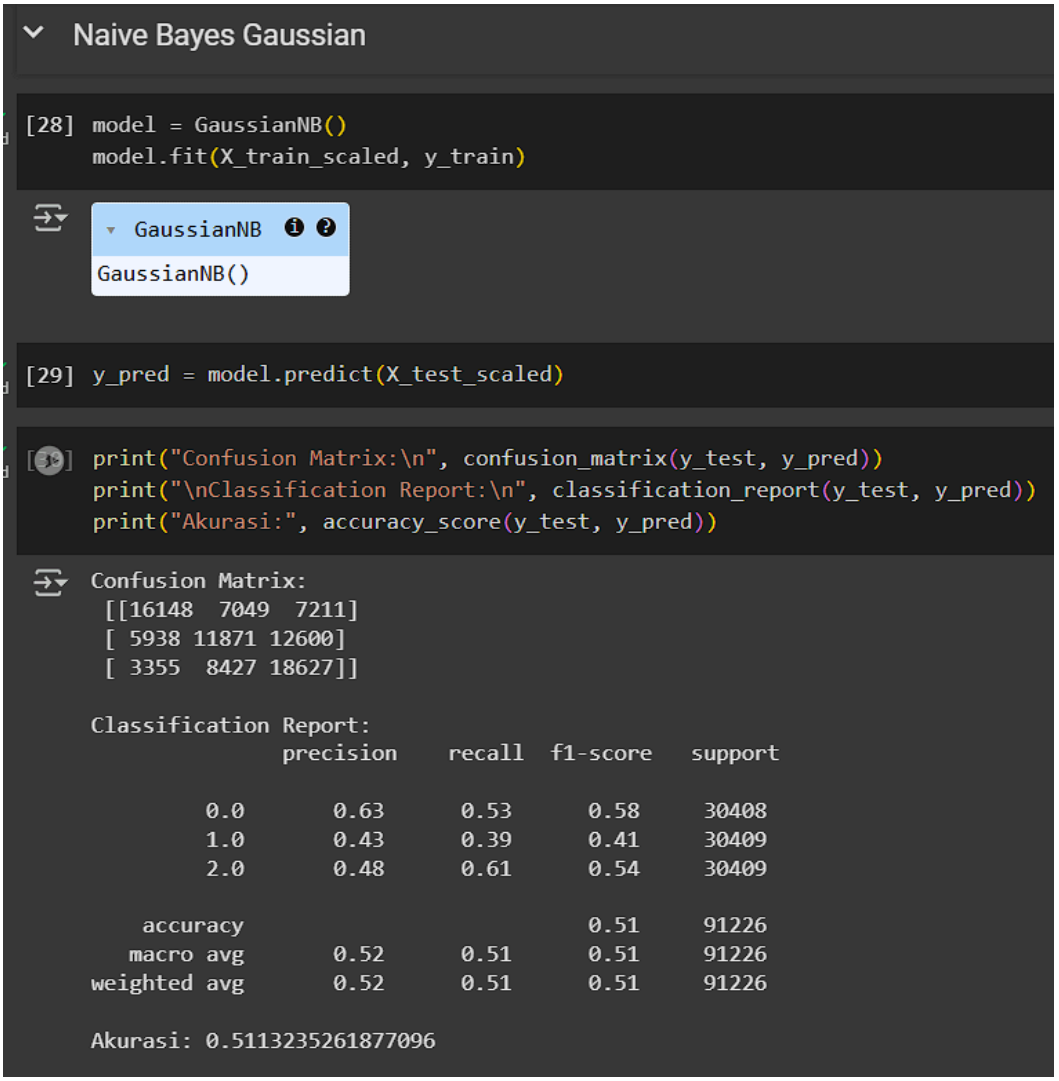
2. Setelah Penyeimbangan Class Data dengan SMOTE

Sesudah SMOTE

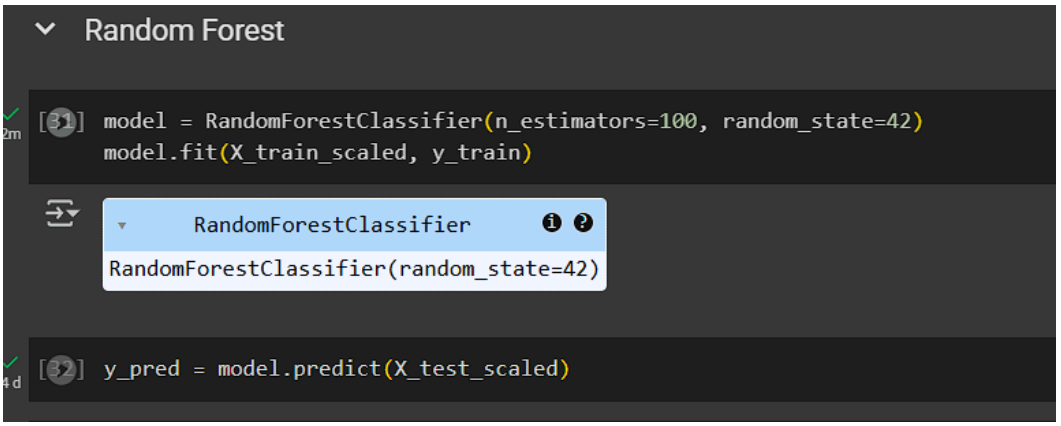
```
[26] X_train, X_test, y_train, y_test = train_test_split(
         X_resampled, y_resampled, test_size=0.2, random_state=42, stratify=y_resampled
     )
```

```
[27] from sklearn.preprocessing import StandardScaler
     scaler = StandardScaler()
     X_train_scaled = scaler.fit_transform(X_train)
     X_test_scaled = scaler.transform(X_test)
```

2.1. Naive Bayes Classifier



2.2. Random Forest Classifier



```
[33] print("Confusion Matrix:\n", confusion_matrix(y_test, y_pred))
     print("\nClassification Report:\n", classification_report(y_test, y_pred))
     print("Akurasi:", accuracy_score(y_test, y_pred))

Confusion Matrix:
 [[28906    31  1471]
 [  621 29426   362]
 [ 3966   468 25975]]

Classification Report:
              precision    recall  f1-score   support

         0.0       0.86      0.95      0.90     30408
         1.0       0.98      0.97      0.98     30409
         2.0       0.93      0.85      0.89     30409

    accuracy                           0.92     91226
   macro avg       0.93      0.92      0.92     91226
weighted avg       0.93      0.92      0.92     91226

Akurasi: 0.924155394295486
```

3.7 Analisis Data

Penelitian ini mengimplementasikan dua algoritma klasifikasi, yaitu Gaussian Naive Bayes dan Random Forest, pada data kesehatan pasien untuk memprediksi kemungkinan status diabetes. Evaluasi dilakukan dalam dua kondisi, yaitu sebelum dan sesudah dilakukan penyeimbangan data menggunakan teknik SMOTE (Synthetic Minority Over-sampling Technique). Evaluasi dilakukan berdasarkan metrik klasifikasi seperti *precision*, *recall*, *f1-score*, dan *akurasi*.

1. Hasil analisis sebelum SMOTE

   Sebelum dilakukan penyeimbangan data, model cenderung memiliki performa yang tidak merata antar kelas. Hal ini terjadi karena distribusi data yang tidak seimbang, di mana kelas “non-diabetes” (0) jauh lebih dominan dibanding kelas “pre-diabetes” (1) dan “diabetes” (2).

Model	Akurasi	Kelas	Precision	Recall	F1-Score
Naive Bayes	0.746	0	0.90	0.79	0.84
		1	0.05	0.01	0.01
		2	0.33	0.57	0.42
Random Forest	0.825	0	0.85	0.96	0.90
		1	0.00	0.00	0.00
		2	0.46	0.19	0.27

Dari tabel di atas dapat dilihat bahwa meskipun nilai akurasi dari Random Forest lebih tinggi dibandingkan Naive Bayes, namun performa terhadap kelas minoritas (1 dan 2) sangat rendah. Model cenderung bias terhadap kelas mayoritas (0).

2. Hasil analisis sesudah SMOTE\

Setelah data diseimbangkan menggunakan SMOTE, distribusi kelas menjadi setara, sehingga model memiliki kesempatan yang sama dalam mempelajari pola dari tiap kelas. Hasil evaluasi setelah SMOTE ditunjukkan pada tabel berikut:

Model	Akurasi	Kelas	Precision	Recall	F1-Score
Naive Bayes	0.511	0	0.63	0.53	0.58
		1	0.43	0.39	0.41
		2	0.48	0.61	0.54
Random Forest	0.924	0	0.86	0.95	0.90
		1	0.98	0.97	0.98
		2	0.93	0.85	0.89

Dari hasil di atas, terlihat bahwa setelah dilakukan SMOTE untuk model Random Forest menunjukkan peningkatan signifikan pada semua kelas, dengan akurasi mencapai 92.4% dan nilai *f1-score* tinggi untuk setiap kelas, termasuk kelas minoritas.Namun, Naive Bayes mengalami hal yang sebaliknya yaitu penurunan akurasi menjadi 51.1%, namun distribusi prediksi terhadap kelas minoritas menjadi lebih merata, meskipun masih kalah jauh dari Random Forest dalam performa

keseluruhan.

Analisis yang didapatkan adalah

- Tanpa SMOTE, algoritma Random Forest tampak unggul dari segi akurasi keseluruhan, namun sangat lemah dalam mengenali kelas minoritas.
- Naive Bayes memberikan prediksi yang sedikit lebih baik untuk kelas minoritas dibanding Random Forest, tetapi akurasinya tetap rendah dan performanya tidak stabil.
- Setelah SMOTE, Random Forest menunjukkan performa terbaik dengan keseimbangan antara *precision*, *recall*, dan *f1-score* pada semua kelas.
- SMOTE sangat efektif dalam meningkatkan kemampuan model dalam mengenali kelas minoritas, terutama pada Random Forest.

3.8 Jadwal Penelitian

No.	Jenis Kegiatan	Waktu Pelaksanaan				
		April		Mei		
		1	2	3	4	5
1.	Identifikasi Masalah					
2.	Studi Literatur					
3.	Penentuan Judul					
4.	Perumusan Masalah, Tujuan, dan Batasan					
5	Pengumpulan Data					
6.	Preprocessing Data					
7.	Pemodelan Menggunakan Gaussian Nayes Bayes melalui metode Smote					
8.	Analisis dan Evaluasi Hasil Pemodelan					
9.	Kesimpulan					

BAB IV
KESIMPULAN

4.1 Kesimpulan

Berdasarkan hasil penelitian yang telah dilakukan, dapat disimpulkan bahwa algoritma Gaussian Naive Bayes dapat digunakan untuk melakukan klasifikasi penyakit diabetes berdasarkan data pasien. Algoritma ini bekerja berdasarkan prinsip probabilistik dan mengasumsikan bahwa setiap fitur numerik dalam data mengikuti distribusi Gaussian. Pada implementasinya, Gaussian Naive Bayes mampu memberikan hasil klasifikasi yang cukup baik, terutama dalam mengenali pola pada data kesehatan pasien.

Namun, penelitian ini juga mengungkapkan bahwa salah satu permasalahan yang muncul dalam proses klasifikasi adalah ketidakseimbangan kelas dalam dataset, di mana jumlah data pasien non-diabetes jauh lebih banyak dibandingkan pasien yang terdiagnosis diabetes. Hal ini mengakibatkan model cenderung bias terhadap kelas mayoritas, sehingga performa klasifikasi untuk kelas minoritas menjadi sangat rendah. Untuk mengatasi permasalahan tersebut, dilakukan penyeimbangan data menggunakan metode SMOTE (Synthetic Minority Over-sampling Technique). Hasil dari penerapan SMOTE menunjukkan bahwa distribusi data menjadi lebih seimbang dan memberikan kesempatan yang sama bagi model untuk mempelajari pola dari tiap kelas.

Setelah dilakukan SMOTE, algoritma Random Forest menunjukkan peningkatan performa yang signifikan, baik dari segi akurasi, presisi, recall, maupun F1-score pada seluruh kelas. Sebaliknya, performa Gaussian Naive Bayes mengalami penurunan akurasi, namun memberikan distribusi prediksi yang lebih merata terhadap kelas minoritas. Meskipun demikian, secara keseluruhan, Random Forest tetap menjadi algoritma dengan performa terbaik dalam penelitian ini.

Dengan demikian, dapat disimpulkan bahwa kombinasi algoritma klasifikasi dan teknik penyeimbangan data sangat berpengaruh terhadap hasil prediksi. Gaussian Naive Bayes tetap menjadi algoritma yang efisien dan sederhana, namun dalam konteks data yang kompleks dan tidak seimbang, algoritma Random Forest lebih unggul dalam menghasilkan klasifikasi yang akurat dan merata..

4.2 Saran

Berdasarkan hasil penelitian yang telah dilakukan, terdapat beberapa saran yang dapat diberikan untuk pengembangan dan penelitian lebih lanjut. Pertama, meskipun

algoritma Gaussian Naive Bayes mampu memberikan hasil klasifikasi yang cukup baik, performanya masih terbatas dalam mengenali kelas minoritas. Oleh karena itu, disarankan untuk mempertimbangkan penggunaan atau kombinasi dengan algoritma lain yang lebih adaptif terhadap data yang tidak seimbang, seperti XGBoost, SVM, atau metode berbasis deep learning.

Kedua, untuk meningkatkan akurasi model secara keseluruhan, perlu dilakukan penambahan atribut atau fitur relevan yang lebih beragam, seperti riwayat keluarga, kadar HbA1c, pola makan, serta aktivitas fisik secara detail. Penambahan fitur ini diharapkan dapat membantu model dalam mengenali pola yang lebih kompleks dari data pasien.

Terakhir, untuk penerapan di dunia nyata, disarankan agar model klasifikasi yang telah dibangun dapat diintegrasikan ke dalam sistem informasi kesehatan atau aplikasi berbasis web/mobile. Dengan demikian, hasil penelitian ini dapat dimanfaatkan langsung oleh tenaga medis maupun masyarakat dalam proses deteksi dini dan pengambilan keputusan terkait diagnosis diabetes.

DAFTAR PUSTAKA

Nisaa, T. A., Ningrum, S. M., & Haque, B. A. (2021). *Diagnosis of diabetes using Naïve Bayes classifier method*. C, 1, 22–29.

Mahran, A. A., Hapsari, R. K., & Nugroho, H. (2020). Penerapan Naive Bayes Gaussian pada klasifikasi jenis jamur berdasarkan ciri statistik orde pertama. *Jurnal Ilmiah NERO*, *5*(2), 91–99.

Adiningrum, N. T. R., & Harani, N. H. (2023). Analisis Perbandingan Ensemble Machine Learning dengan Teknik SMOTE untuk Prediksi Diabetes. *JEIS: Jurnal Elektro dan Informatika Swadharma*, *6*(1), 12–20. https://doi.org/10.56670/jsrd.v6i1.326

Rezki, M. K., Mazdadi, M. I., Indriani, F., Muliadi, M., Saragih, T. H., & Athavale, V. A. (2024). Application of SMOTE to address class imbalance in diabetes disease classification utilizing C5.0, Random Forest, and SVM. *Journal of Electronics, Electromedical Engineering, and Medical Informatics*, *6*(4), 343–354. https://doi.org/10.35882/jeeemi.v6i4.434

Rahmawati, S., Wibowo, A., & Masruriyah, A. F. N. (2024). Improving Diabetes Prediction Accuracy in Indonesia: A Comparative Analysis of SVM, Logistic Regression, and Naive Bayes with SMOTE and ADASYN. *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, *8*(5), 607–614. https://doi.org/10.29207/resti.v8i5.598