

**KLASIFIKASI PENYAKIT *DIABETES MELLITUS* TIPE II BERBASIS  
*MACHINE LEARNING* MENGGUNAKAN LIGHTGBM**

**(Skripsi)**

**Oleh**

**SALMA IRENA FEBRIASTIA  
NPM 1915061030**



**FAKULTAS TEKNIK  
UNIVERSITAS LAMPUNG  
BANDAR LAMPUNG  
2024**

**KLASIFIKASI PENYAKIT *DIABETES MELLITUS* TIPE II BERBASIS  
*MACHINE LEARNING* MENGGUNAKAN LIGHTGBM**

**Oleh  
SALMA IRENA FEBRIASTIA**

**Skripsi**

**Sebagai Salah Satu Syarat untuk Mencapai Gelar  
SARJANA TEKNIK**

**Pada**

**Jurusan Teknik Elektro  
Fakultas Teknik Universitas Lampung**



**FAKULTAS TEKNIK  
UNIVERSITAS LAMPUNG  
BANDAR LAMPUNG  
2024**

## ABSTRAK

### **KLASIFIKASI PENYAKIT *DIABETES MELLITUS* TIPE II BERBASIS *MACHINE LEARNING* MENGGUNAKAN LIGHTGBM**

oleh

**SALMA IRENA FEBRIASTIA**

*Diabetes mellitus* merupakan sebuah kondisi metabolis serius dan kronis yang terjadi karena kenaikan kadar glukosa darah akibat tubuh tidak dapat memproduksi hormon insulin atau tidak dapat menggunakan insulin yang dihasilkan dengan efektif. Sekitar 537 juta orang di berbagai bagian dunia mengalami diabetes. Pada tahun 2019, diabetes menjadi salah satu penyebab kematian utama, mengalami peningkatan sebesar 70% sejak tahun 2000. Di Indonesia, diabetes menempati peringkat tiga sebagai penyebab kematian pada tahun 2019, mencapai persentase 6.23%. Penelitian ini bertujuan untuk melakukan klasifikasi penyakit *diabetes mellitus* tipe II menggunakan *Light Gradient-Boosting Machine* (LightGBM) dengan dua skenario, yaitu menggunakan fitur yang mengandung hasil tes laboratorium berupa FPG (skenario 1) dan tanpa fitur hasil tes laboratorium (skenario 2). Pada kedua skenario, dilakukan evaluasi menggunakan konfigurasi seluruh fitur, fitur terseleksi berdasarkan *mutual information*, dan fitur rekomendasi dari *expert*. Lalu membuat sebuah antarmuka *website* untuk memprediksi diagnosis penyakit *diabetes mellitus* tipe II. Metode yang digunakan adalah *Cross-Industry Standard Process for Data Mining* (CRISP-DM). Hasil evaluasi menunjukkan bahwa model memiliki kinerja yang baik dalam semua konfigurasi fitur pada kedua skenario. Tujuh fitur rekomendasi berdasarkan *expert* pada skenario 2 (AGE, Nocturia, Polyuria, Weight\_loss, Polydipsia, Polyphagia, BMI) dengan akurasi 94,87% digunakan sebagai bahan untuk memprediksi diagnosis pada antarmuka *website* karena relevan dengan ilmu kedokteran dan pengguna dapat melakukan *prescreening* awal *diabetes mellitus* tipe II sebelum melakukan pemeriksaan lebih lanjut ke dokter, sehingga meningkatkan efisiensi waktu dan biaya.

Kata kunci: CRISP-DM, *diabetes mellitus*, klasifikasi, LightGBM, klasifikasi, *mutual information*

## **ABSTRACT**

### **MACHINE LEARNING-BASED TYPE II DIABETES MELLITUS DISEASE CLASSIFICATION USING LIGHTGBM**

**By**

**SALMA IRENA FEBRIASTIA**

*Diabetes mellitus is a serious and chronic metabolic condition that occurs due to elevated blood glucose levels because the body cannot produce the hormone insulin or cannot use the insulin produced effectively. About 537 million people in various parts of the world have diabetes. In 2019, diabetes became one of the leading causes of death, having increased by 70% since 2000. In Indonesia, diabetes ranked third as a cause of death in 2019, reaching a percentage of 6.23%. This study aims to classify type II diabetes mellitus using Light Gradient-Boosting Machine (LightGBM) with two scenarios, namely using features containing laboratory test results in the form of FPG (scenario 1) and without laboratory test results features (scenario 2). In both scenarios, evaluation was carried out using the configuration of all features, selected features based on mutual information, and expert recommendation features. Then create a website interface to predict the diagnosis of type II diabetes mellitus. The method used is Cross-Industry Standard Process for Data Mining (CRISP-DM). The evaluation results showed that the model performed well in all feature configurations in both scenarios. Seven expert-based recommendation features in scenario 2 (AGE, Nocturia, Polyuria, Weight\_loss, Polydipsia, Polyphagia, BMI) with an accuracy of 94.87% are used as configuration for predicting diagnosis on the website interface because they are relevant to medical science and users can perform initial prescreening of type II diabetes mellitus before conducting further examinations to doctors, thereby increasing time and cost efficiency.*

**Keywords:** *Classification, CRISP-DM, diabetes mellitus, LightGBM, mutual information*



Judul Skripsi

: **KLASIFIKASI PENYAKIT *DIABETES MELLITUS* TIPE II BERBASIS *MACHINE LEARNING* MENGGUNAKAN *LIGHTGBM***

Nama Mahasiswa

: **Salma Irena Febriastia**

Nomor Pokok Mahasiswa

: **1915061030**

Program Studi

: **Teknik Informatika**

Fakultas


: **Teknik**



1. Komisi Pembimbing

Pembimbing Utama

Pembimbing Pendamping


  
**Ir. Meizano Ardhi Muhammad, S.T., M.T., IPM**  
NIP. 198105282012121001


  
**Ir. Titin Yulianti, S.T., M.Eng.**  
NIP. 198807092019032015

2. Mengetahui

Ketua Jurusan  
Teknik Elektro

Ketua Program Studi  
Teknik Informatika

  
**Herlinawati, S.T., M.T.**  
NIP. 197103141999032001

  
**Yessi Mulyani, S.T., M.T.**  
NIP. 197312262000122001



## MENGESAHKAN

### 1. Tim Penguji

Ketua : **Ir. Meizano Ardhi Muhammad, S.T., M.T., IPM** .....

Sekretaris : **Ir. Titin Yulianti, S.T., M.Eng.** .....

Penguji : **Ir. Muhamad Komarudin, S.T., M.T.** .....

### 2. Dekan Fakultas Teknik

**Dr. Eng. Ir. Helmy Fitriawan, S.T., M.Sc. }**

NIP. 197509282001121002

Tanggal Lulus Ujian Skripsi: 07 Maret 2024



## SURAT PERNYATAAN

Saya yang bertandatangan di bawah ini, menyatakan bahwa skripsi saya dengan judul “Klasifikasi Penyakit *Diabetes Mellitus* Tipe II Berbasis *Machine Learning* Menggunakan LightGBM” dibuat oleh saya sendiri. Semua hasil yang tertuang dalam skripsi ini telah mengikuti kaidah penulisan karya ilmiah Universitas Lampung. Apabila di kemudian hari terbukti bahwa skripsi ini merupakan salinan atau dibuat oleh orang lain, maka saya bersedia menerima sanksi sesuai dengan ketentuan hukum atau akademik yang berlaku.

Bandar Lampung, 07 Maret 2024

Pembuat pernyataan,



Salma Irena Febriastia

NPM. 1915061030

## **MOTO**

“Your Lord hasn’t abandoned you, nor has He become hateful (of you).”  
**(Qur’an, 93:03)**

“Don’t be afraid; I am with you all the time, listening and seeing.”  
**(Qur’an, 20:46)**

"What's yours will find you."  
**-Ali ibn Abi Talib**

“God is always protecting me. Whatever is removed from my life or whatever opportunity I didn’t get, is because God is giving me something bigger. God has not failed me. God has my back and I believe that.”  
**-Thewizardliz**

“Look around. There are people who only look at you. A person receives a one-and-only love; and that person is you.”  
**-Neo Culture Technology; “Beautiful”**



## DAFTAR ISI

	Halaman
<b>DAFTAR ISI</b> .....	iv
<b>DAFTAR TABEL</b> .....	vi
<b>DAFTAR GAMBAR</b> .....	vii
<b>I. PENDAHULUAN</b> .....	1
1.1. Latar Belakang .....	1
1.2. Rumusan Masalah .....	3
1.3. Tujuan Penelitian .....	3
1.4. Manfaat Penelitian .....	4
1.5. Batasan Masalah .....	4
1.6. Sistematika Penulisan Skripsi .....	4
<b>II. TINJAUAN PUSTAKA</b> .....	6
2.1. <i>Diabetes Mellitus</i> .....	6
2.2. <i>Machine Learning</i> .....	8
2.3. <i>Light Gradient-Boosting Machine (LightGBM)</i> .....	10
2.4. <i>Python</i> .....	12
2.5. <i>Mutual Information</i> .....	13
2.6. <i>Confusion Matrix</i> .....	14
2.7. <i>Receiver Operating Characteristic Are Under the Curve (ROC AUC)</i> ..	16
2.8. <i>Flask</i> .....	17
2.9. <i>Cross-Industry Standard Process for Data Mining (CRISP-DM)</i> .....	18
2.10. <i>Google Colab</i> .....	20
2.11. <i>Visual Studio Code</i> .....	21
2.12. <i>Penelitian Terdahulu</i> .....	21
<b>III. METODOLOGI PENELITIAN</b> .....	27
3.1. Waktu dan Tempat Penelitian .....	27
3.2. Alat dan Bahan Penelitian .....	28
3.3. Tahapan Penelitian .....	29
3.4. Tahapan CRISP-DM .....	30
3.4.1. <i>Business Understanding</i> .....	32
3.4.2. <i>Data Understanding</i> .....	32

3.4.3.	<i>Data Preparation</i> .....	33
3.4.4.	<i>Modeling</i> .....	33
3.4.5.	<i>Evaluation</i> .....	33
3.4.6.	<i>Deployment</i> .....	33
<b>IV.</b>	<b>HASIL DAN PEMBAHASAN</b> .....	34
4.1.	<i>Business Understanding</i> .....	34
4.2.	<i>Data Understanding</i> .....	36
4.2.1.	<i>Dataset Exploration</i> .....	38
4.3.	<i>Data Preparation</i> .....	41
4.3.1.	<i>Data Cleaning</i> .....	41
4.3.2.	<i>Data Transformation</i> .....	41
4.3.3.	<i>Feature Selection</i> .....	43
4.4.	<i>Modeling</i> .....	45
4.5.	<i>Evaluation</i> .....	47
4.5.1.	Hasil Evaluasi Skenario 1 (Menggunakan Fitur Hasil Tes Lab).....	47
4.5.2.	Hasil Evaluasi Skenario 2 (Tanpa Fitur Hasil Tes Lab).....	56
4.5.3.	Perbandingan Hasil Evaluasi Antara Skenario 1 dan Skenario 2 ....	65
4.6.	<i>Deployment</i> .....	70
<b>V.</b>	<b>KESIMPULAN DAN SARAN</b> .....	74
5.1.	Kesimpulan.....	74
5.2.	Saran .....	75
	<b>DAFTAR PUSTAKA</b> .....	77
	<b>LAMPIRAN</b> .....	82
	Lampiran A. <i>Source Code</i> .....	A
	Lampiran B. Dokumentasi Wawancara .....	J

## DAFTAR TABEL

Tabel	Halaman
Tabel 1. Penelitian Terdahulu .....	24
Tabel 2. Jadwal Penelitian.....	27
Tabel 3. Alat Penelitian .....	28
Tabel 4. Distribusi Jumlah <i>Dataset</i> .....	29
Tabel 5. Pengelompokan Diabetes Berdasarkan Hasil Tes .....	35
Tabel 6. Atribut <i>Dataset</i> .....	36
Tabel 7. Informasi Tipe Data dan Jumlah <i>Non-null</i> .....	38
Tabel 8. Statistik Deskriptif.....	39
Tabel 9. Hasil Pengecekan <i>Missing Values</i> dan <i>Zero Count</i> .....	40
Tabel 10. Distribusi Kelas Target .....	41
Tabel 11. Distribusi Kelas Target Setelah <i>Data Cleaning</i> .....	41
Tabel 12. <i>Label Encoding</i> .....	41
Tabel 13. Informasi <i>Dataset</i> Setelah Proses <i>Data Transformation</i> .....	42
Tabel 14. Fitur yang Digunakan Pada Skenario 1 (Menggunakan Fitur Hasil Tes Lab) .....	45
Tabel 15. Fitur yang Digunakan Pada Skenario 2 (Tanpa Fitur Hasil Tes Lab) .....	45
Tabel 16. Parameter yang Digunakan pada Tahap <i>Modeling</i> .....	46
Tabel 17. <i>Classification Report</i> Skenario 1 Ketika Menggunakan Seluruh Fitur .....	49
Tabel 18. <i>Classification Report</i> Skenario 1 Ketika Menggunakan Fitur-Fitur Terseleksi Berdasarkan Mutual Information .....	52
Tabel 19. <i>Classification Report</i> Skenario 1 Ketika Menggunakan Fitur-Fitur Rekomendasi Berdasarkan Expert.....	55
Tabel 20. <i>Classification Report</i> Skenario 2 Ketika Menggunakan Seluruh Fitur .....	58
Tabel 21. <i>Classification Report</i> Skenario 2 Ketika Menggunakan Fitur-Fitur Terseleksi Berdasarkan <i>Mutual Information</i> .....	61
Tabel 22. <i>Classification Report</i> Skenario 2 Ketika Menggunakan Fitur-Fitur Rekomendasi Berdasarkan <i>Expert</i> .....	64
Tabel 23. Perbandingan Nilai Pada Skenario 1 .....	66
Tabel 24. Perbandingan Nilai Pada Skenario 2 .....	66
Tabel 25. Pengujian Fungsionalitas Antarmuka.....	72
Tabel 26. Pengujian Hasil Prediksi Antarmuka.....	73



## DAFTAR GAMBAR

Gambar	Halaman
Gambar 1. <i>Diabetes Mellitus</i> Tipe 1 [6].....	7
Gambar 2. <i>Diabetes Mellitus</i> Tipe II [6].....	7
Gambar 3. Diagram <i>Machine Learning</i> [9] .....	9
Gambar 4. <i>Level-wise Tree Growth</i> [11] .....	11
Gambar 5. <i>Leaf-wise Tree Growth</i> [11].....	11
Gambar 6. <i>Confusion Matrix</i> [8].....	14
Gambar 7. <i>Classification Report</i> [21] .....	16
Gambar 8. Kurva ROC [12] .....	16
Gambar 9. Tahapan CRISP-DM [26].....	18
Gambar 10. Tampilan Awal Google Colab .....	20
Gambar 11. Flowchart Tahapan Penelitian .....	30
Gambar 12. Tahapan CRISP-DM.....	31
Gambar 13. Ghanaian Diabetes Dataset dari kaggle.com [39] .....	32
Gambar 14. Tampilan <i>Dataset</i> Sebelum <i>Data Transformation</i> .....	42
Gambar 15. Tampilan <i>Dataset</i> Setelah <i>Data Transformation</i> .....	43
Gambar 16. Tampilan <i>Dataset</i> Pada Skenario 1 Setelah <i>Feature Selection</i> Menggunakan <i>Mutual Information</i> ....	44
Gambar 17. Tampilan <i>Dataset</i> Pada Skenario 2 Setelah <i>Feature Selection</i> Menggunakan <i>Mutual Information</i> ....	44
Gambar 18. <i>Confusion Matrix</i> Skenario 1 Ketika Menggunakan Seluruh Fitur .....	48
Gambar 19. ROC AUC Skenario 1 Ketika Menggunakan Seluruh Fitur .....	50
Gambar 20. <i>Confusion Matrix</i> Skenario 1 Ketika Menggunakan Fitur-Fitur Terseleksi Berdasarkan <i>Mutual Information</i> .....	51
Gambar 21. ROC AUC Menggunakan Fitur-Fitur Terseleksi Berdasarkan <i>Mutual Information</i> .....	53
Gambar 22. <i>Confusion Matrix</i> Skenario 1 Ketika Menggunakan Fitur-Fitur Rekomendasi Berdasarkan <i>Expert</i> .....	54
Gambar 23. ROC AUC Skenario 1 Ketika Menggunakan Fitur-Fitur Rekomendasi Berdasarkan <i>Expert</i> .....	56
Gambar 24. <i>Confusion Matrix</i> Skenario 2 Ketika Menggunakan Seluruh Fitur .....	57
Gambar 25. ROC AUC Skenario 2 Ketika Menggunakan Seluruh Fitur .....	59
Gambar 26. <i>Confusion Matrix</i> Menggunakan Fitur-Fitur Terseleksi Berdasarkan <i>Mutual Information</i> .....	60

Gambar 27. ROC AUC Skenario 2 Ketika Menggunakan Fitur-Fitur Terseleksi Berdasarkan <i>Mutual Information</i> .....	62
Gambar 28. <i>Confusion Matrix</i> Skenario 2 Ketika Menggunakan Fitur-Fitur Rekomendasi Berdasarkan <i>Expert</i> .....	63
Gambar 29. ROC AUC Skenario 2 Ketika Menggunakan Fitur-Fitur Rekomendasi Berdasarkan <i>Expert</i> .....	65
Gambar 30. Visualisasi Perbandingan Nilai Hasil Evaluasi Kelas 0 Antara Skenario 1 dan 2 .....	67
Gambar 31. Visualisasi Perbandingan Nilai Hasil Evaluasi Kelas 0 Antara Skenario 1 dan 2 .....	68
Gambar 32. Visualisasi Perbandingan Nilai Hasil Evaluasi <i>Accuracy</i> dan <i>Execution Time</i> Antara Skenario 1 dan 2 .....	68
Gambar 33. Visualiasi Perbandingan Nilai Skenario 1 dan 2 .....	69
Gambar 34. Antarmuka <i>Website</i> .....	71
Gambar 35. Contoh <i>Input</i> Pengguna .....	72
Gambar 36. Contoh Tampilan Hasil Prediksi .....	73

## **I. PENDAHULUAN**

### **1.1. Latar Belakang**

*Diabetes mellitus* atau yang lebih sering disebut diabetes merupakan sebuah kondisi metabolis serius dan kronis yang terjadi saat adanya kenaikan kadar glukosa darah akibat tubuh tidak dapat memproduksi hormon insulin atau ketidakmampuan tubuh menggunakan yang telah diproduksi secara efektif. Kenaikan kadar glukosa tersebut menyebabkan glukosa yang berlebihan di dalam tubuh penderita (hiperglikemia).

Diabetes tergolong sebagai penyakit yang berbahaya karena dapat menyebabkan komplikasi kesehatan. Komplikasi jangka panjang akibat penyakit diabetes berkembang secara bertahap. Semakin lama seseorang mengidap diabetes, semakin bertambah pula resiko komplikasi kesehatan yang terjadi. Komplikasi tersebut dapat melumpuhkan atau bahkan mengancam jiwa seseorang. Penyakit kardiovaskular, neuropati, nefropati, stroke, amputasi tungkai bawah, serta penyakit mata yang dapat menyebabkan kebutaan merupakan contoh komplikasi yang mungkin terjadi akibat diabetes.

Penderita penyakit diabetes terus bertambah setiap tahunnya. Sekitar 537 juta orang di berbagai belahan dunia menderita diabetes [1]. Pada tahun 2019, diabetes termasuk ke dalam sepuluh besar penyebab utama kematian, dengan peningkatan sebesar 70% sejak tahun 2000. Diabetes menjadi penyebab langsung dari 1,5 juta kematian di dunia. Secara tidak langsung, diabetes juga melatarbelakangi 460.000 kematian akibat penyakit ginjal yang merupakan komplikasi dari penyakit diabetes [2].



Indonesia menempati peringkat kedua sebagai penderita diabetes terbanyak di daerah Pasifik Barat, dengan jumlah penderita mencapai 7.3 juta pada tahun 2011 dan 19.5 juta penderita pada tahun 2021. Di sisi lain, pada peringkat global, Indonesia menempati peringkat lima sebagai negara dengan penderita diabetes terbanyak. Berdasarkan data dari Federasi Diabetes Internasional, diprediksi penderita diabetes di Indonesia akan mengalami pertambahan menjadi 28.6 juta jiwa pada tahun 2045 mendatang [1]. Diabetes juga menempati peringkat tiga sebagai penyakit yang menyebabkan kematian di Indonesia pada tahun 2019 dengan persentase 6.23% total kematian dengan tingkat kenaikan sebesar 49.9% dari tahun 2009 [3].

*Prescreening* awal memainkan peran penting dalam mendeteksi gejala diabetes sebelum pemeriksaan lebih lanjut. Ini memberikan solusi efisien bagi individu yang memiliki keterbatasan waktu dan memungkinkan individu untuk mengidentifikasi potensi masalah kesehatan sendiri tanpa antrian panjang. Apabila diabetes telah dideteksi sejak dini, kemudian penderita melakukan perawatan diabetes yang efektif dengan pengobatan secara rutin, diabetes dapat dikontrol dan konsekuensi komplikasi kesehatan yang serius akibat diabetes dapat dicegah atau bahkan dihindari.

*Machine learning* semakin berkembang dan memiliki peran penting dalam bidang industri, termasuk dalam melakukan diagnosis medis. Klasifikasi merupakan tugas dari *machine learning* yang bertujuan untuk memprediksi label kelas yang bersifat diskrit. Kelas tersebut terdiri dari sekumpulan kemungkinan yang telah ditentukan. Terdapat berbagai algoritma untuk melakukan klasifikasi, salah satunya adalah *Light Gradient Boosting Machine* (LightGBM). LightGBM terkenal karena efisiensi pelatihan yang tinggi serta mencapai hasil yang sangat baik dalam berbagai penelitian mengenai klasifikasi.

*Machine learning* dapat dimanfaatkan untuk mempermudah proses *prescreening* awal klasifikasi penyakit *diabetes mellitus* yang dapat membantu tenaga kesehatan

dalam menentukan diagnosis terhadap individu. Oleh karena itu, penelitian ini melakukan klasifikasi menggunakan *dataset* yang memiliki fitur yang representatif sebagai indikasi diabetes. Dengan begitu, fitur pada penelitian ini dapat menjadi dasar yang kuat bagi pengguna untuk melakukan pemeriksaan lebih lanjut. Dalam implementasinya, diperlukan antarmuka untuk memudahkan pengguna, sehingga pembuatan antarmuka *website* dilakukan dengan memanfaatkan suatu *web framework*.

## 1.2. Rumusan Masalah

Berdasarkan latar belakang yang telah dijabarkan, kajian masalah yang mendasari penelitian ini adalah sebagai berikut:

1. Bagaimana cara melakukan klasifikasi penyakit *diabetes mellitus* dengan dataset dari kaggle.com menggunakan algoritma LightGBM yang kemudian disediakan melalui suatu antarmuka *website* untuk melakukan prediksi *diabetes mellitus*?
2. Bagaimana hasil evaluasi klasifikasi dengan menggunakan seluruh fitur, fitur-fitur terseleksi dari *mutual information*, dan rekomendasi fitur dari *expert*?

## 1.3. Tujuan Penelitian

Adapun tujuan dari penelitian ini adalah sebagai berikut:

1. Melakukan klasifikasi penyakit *diabetes mellitus* menggunakan LightGBM dengan dua skenario, yaitu menggunakan fitur yang mengandung hasil tes laboratorium (skenario 1) dan tanpa fitur hasil tes laboratorium (skenario 2).
2. Membandingkan hasil evaluasi klasifikasi dengan menggunakan seluruh fitur, fitur-fitur terseleksi dari *mutual information*, dan rekomendasi fitur dari *expert* pada kedua skenario.
3. Membangun antarmuka *website* untuk prediksi diagnosis penyakit *diabetes mellitus*.

#### 1.4. Manfaat Penelitian

Manfaat dari penelitian ini adalah sebagai berikut:

1. Tersedia referensi untuk klasifikasi *diabetes mellitus* menggunakan LightGBM.
2. Tersedia antarmuka *website* yang dapat digunakan untuk melakukan prediksi penyakit *diabetes mellitus*.

#### 1.5. Batasan Masalah

Dalam penelitian ini, pembatasan masalah meliputi hal-hal sebagai berikut:

1. *Dataset* yang digunakan merupakan data penderita *diabetes mellitus* yang bersumber dari kaggle.com dan diakses pada bulan Januari 2024.
2. Penelitian ini tidak membahas mengenai analisis antarmuka *website* prediksi.

#### 1.6. Sistematika Penulisan Skripsi

Sistematika penulisan skripsi ini terdiri dari lima bab sebagai berikut:

### I. PENDAHULUAN

Memuat latar belakang dan dorongan dari penyusunan skripsi ini. Selain itu, juga memuat tujuan penelitian, rumusan masalah, manfaat penelitian, batasan masalah, dan sistematika penulisan skripsi.

### II. TINJAUAN PUSTAKA

Memuat dasar-dasar teori mengenai *diabetes mellitus*, *machine learning*, *Light Gradient-Boosting Machine* (LightGBM), Python, *mutual information*, *classification report* dan *confusion matrix*, ROC AUC, Flask, dan *Cross-Industry Standard Process for Data Mining* (CRISP-DM), Google Colab, dan Visual Studio Code. Selain itu juga memuat penelitian terdahulu.



### III. METODOLOGI PENELITIAN

Memuat waktu dan tempat penelitian, alat dan bahan penelitian, serta tahapan penelitian yang dilakukan menggunakan *Cross-Industry Standard Process for Data Mining* (CRISP-DM).

### IV. HASIL DAN PEMBAHASAN

Memuat hasil dan pembahasan dari penelitian yang meliputi *business understanding*, *data understanding*, *data preparation* (*data cleaning* dan *data transformation*), *modeling* menggunakan dua skenario dengan masing-masing terdapat tiga konfigurasi fitur (seluruh fitur, fitur terseleksi berdasarkan *mutual information*, dan fitur rekomendasi *expert*), *evaluation* menggunakan *confusion matrix*, *classification report*, dan skor ROC AUC, dan *deployment* menggunakan Flask.

### V. KESIMPULAN DAN SARAN

Memuat kesimpulan dari penelitian yang dilakukan serta saran-saran terhadap pengembangan penelitian selanjutnya.

### DAFTAR PUSTAKA

Memuat daftar literatur yang digunakan untuk penelitian.

## II. TINJAUAN PUSTAKA

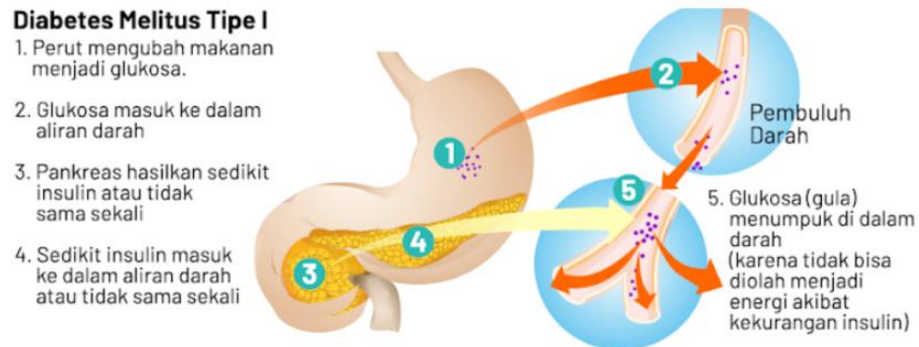
### 2.1. *Diabetes Mellitus*

*Diabetes mellitus* atau yang biasa disebut diabetes, mengacu pada kategori penyakit metabolik yang ditandai dan dibedakan dengan adanya hiperglikemia tanpa adanya pengobatan. Diabetes memiliki etiologi yang beragam, meliputi kelainan produksi insulin, kinerja insulin, maupun keduanya, serta masalah metabolisme karbohidrat, lipid, dan protein. Dampak dari penyakit kronis ini antara lain adalah retinopati, nefropati, dan neuropati. Diabetes juga meningkatkan risiko berbagai penyakit, seperti penyakit jantung, arteri perifer dan serebrovaskular, obesitas, katarak, dan disfungsi ereksi. Selain itu, penderita diabetes juga lebih rentan terhadap penyakit menular seperti TBC [4].

Seseorang yang mengidap *diabetes mellitus* dapat memiliki gejala seperti poliuria (sering buang air kecil), polidipsia (sensasi haus yang berlebihan), polifagia (rasa lapar yang berlebihan), dan penurunan berat badan tanpa penyebab yang jelas. Selain itu, gejala diabetes lainnya termasuk kelelahan dan kurangnya energi, sensasi kesemutan di tangan atau kaki, perasaan gatal, rentan terhadap infeksi bakteri atau jamur, proses penyembuhan luka yang berlangsung lebih lama dari biasanya, dan masalah penglihatan seperti penglihatan kabur. Namun, dalam beberapa kasus, individu yang menderita DM mungkin tidak menunjukkan gejala apa pun [4], [5]. Tanda-tanda klinis yang paling parah adalah ketoasidosis atau kondisi hiperosmolar non-ketotik yang dapat menyebabkan dehidrasi, ketidaksadaran, dan kematian jika tidak ada terapi yang tepat [4].

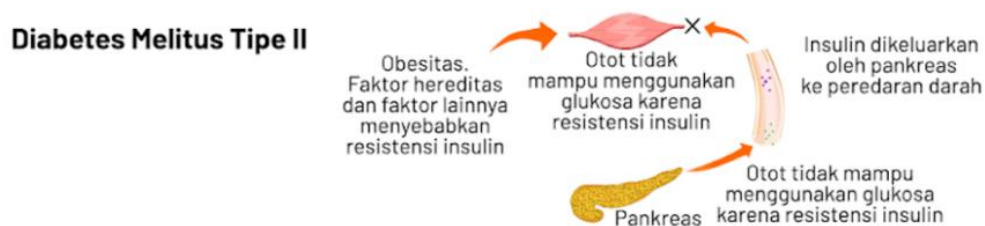
*Diabetes mellitus* dikelompokkan menjadi dua tipe sebagai berikut:

**Tipe 1** - Disebabkan oleh kenaikan kadar gula darah akibat kerusakan sel beta pankreas, sehingga sama sekali tidak ada produksi insulin. Insulin merupakan hormon yang diproduksi oleh pankreas untuk mencerna gula dalam darah. Penderita diabetes tipe ini membutuhkan asupan insulin dari luar tubuhnya [6].



Gambar 1. *Diabetes Mellitus* Tipe 1 [6]

**Tipe 2** - Merupakan tipe diabetes karena adanya kenaikan gula darah akibat penurunan sekresi insulin yang rendah oleh kelenjar pankreas [6]. Laju perburukan hiperglikemia yang lambat pada tipe ini menyebabkan gejala umumnya ringan atau tidak ada sama sekali. Akibatnya, dengan tidak adanya pemeriksaan biokimia, hiperglikemia yang cukup parah dapat mengakibatkan komplikasi pada saat diagnosis [4].



Gambar 2. *Diabetes Mellitus* Tipe II [6]

## 2.2. *Machine Learning*

*Machine learning* dapat didefinisikan secara luas sebagai teknik komputasi yang memanfaatkan pengalaman sebelumnya untuk meningkatkan kinerja atau melakukan prediksi dengan tingkat akurasi yang tinggi. Pengalaman merujuk pada informasi masa lalu yang tersedia bagi pembelajar, yang umumnya berupa data elektronik yang dikumpulkan dan disediakan untuk analisis. Data tersebut bisa berupa kumpulan data pelatihan berlabel yang telah didigitalkan. Kualitas dan jumlah data sangat penting bagi keberhasilan pembelajar dalam melakukan prediksi yang akurat [7].

Jenis algoritma *machine learning* yang paling sukses merupakan algoritma yang melakukan proses pengambilan keputusan secara otomatis dengan cara melakukan generalisasi dari contoh-contoh yang sudah diketahui, atau dikenal dengan *supervised learning*. *Supervised learning* dipilih ketika penelitian berfokus untuk memprediksi hasil tertentu dari input yang diberikan, dan peneliti memiliki contoh data pasangan *input/output*. Bermodalkan contoh tersebut, model yang dibuat dapat melakukan pembelajaran dan menghasilkan prediksi yang akurat secara otomatis untuk data baru yang belum pernah dilihat sebelumnya. Klasifikasi merupakan suatu bentuk dari *supervised learning*. Klasifikasi bertujuan untuk memprediksi label kelas yang bersifat diskrit. Kelas tersebut terdiri dari sekumpulan kemungkinan yang telah ditentukan [8].

Dalam kenyataan, terdapat beberapa tugas standar dalam *machine learning* yang sudah banyak diteliti [7]:

### 1. Klasifikasi

Merupakan masalah menentukan kategori bagi setiap item ke dalam salah satu dari beberapa kelas diskrit yang telah ditentukan. Sebagai contoh, klasifikasi dokumen melibatkan penentuan kategori seperti olahraga, politik, bisnis, atau cuaca untuk setiap dokumen, sementara klasifikasi gambar melibatkan penentuan kategori kendaraan bermotor untuk setiap gambar.



## 2. Regresi

Tugas ini bertujuan untuk memprediksi nilai riil bagi setiap item. Contoh dari regresi mencakup prediksi nilai saham atau variasi variabel ekonomi.

## 3. Pemeringkatan

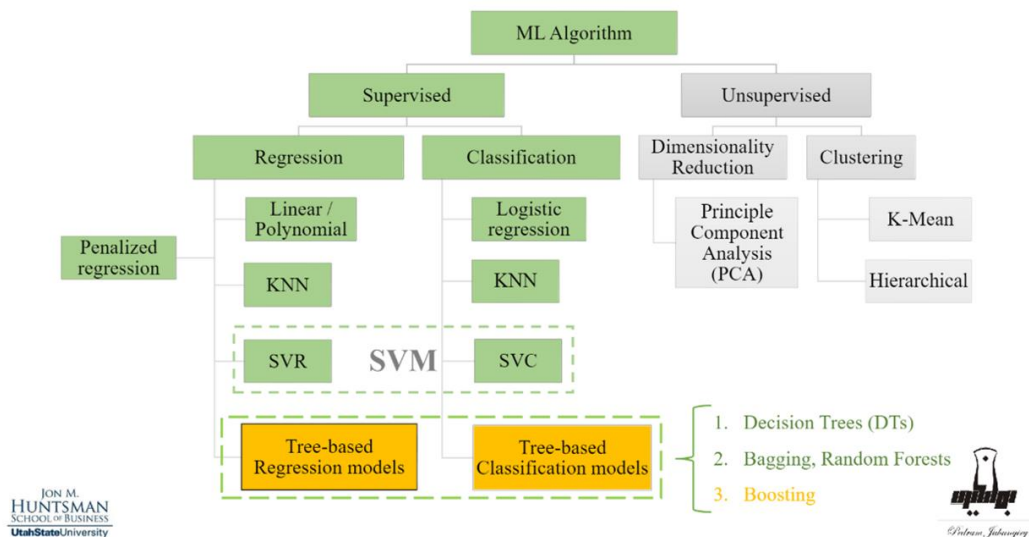
Tugas ini melibatkan pembelajaran untuk mengurutkan item berdasarkan kriteria tertentu. Contohnya adalah pencarian web yang menghasilkan halaman web yang relevan dengan kueri pencarian.

## 4. Pengelompokan (*Clustering*)

Tugas ini berfokus pada pemisahan himpunan item menjadi kelompok-kelompok yang homogen. *Clustering* sering digunakan untuk menganalisis data berskala besar.

## 5. Reduksi Dimensi

Tugas untuk mengubah representasi asli dari item ke dalam representasi yang memiliki dimensi lebih rendah, namun tetap mempertahankan beberapa sifat penting dari representasi aslinya. Contoh umum dari reduksi dimensi adalah pra-pemrosesan gambar digital dalam *computer vision*.



Gambar 3. Diagram *Machine Learning* [9]

Pada Gambar 3 terlihat bahwa klasifikasi (*classification*) termasuk ke dalam *supervised learning*. Klasifikasi dapat dilakukan dengan berbagai algoritma, salah

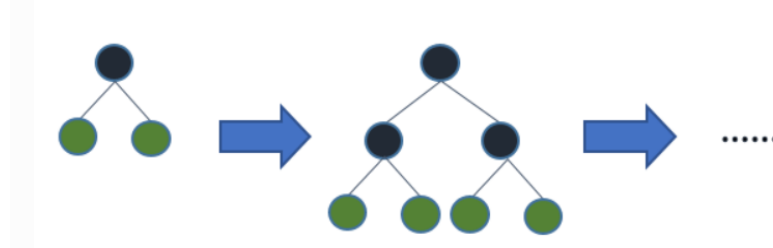
satunya dengan menggunakan model berbasis pohon (*tree-based models*). Model berbasis pohon terbagi lagi menjadi 3 cabang, yaitu *decision tree*, *bagging*, dan *boosting*. Algoritma *boosting* digunakan pada penelitian ini. *Boosting* merupakan algoritma *ensemble* berbasis pohon yang melibatkan penggunaan data pelatihan awal dan secara berulang membuat beberapa model dengan menggunakan *weak learners*. Setiap model baru berupaya memperbaiki kesalahan yang dilakukan oleh model sebelumnya. Dalam *boosting*, setiap pohon dibangun dengan memanfaatkan informasi dari pohon sebelumnya [9][10].

### 2.3. *Light Gradient-Boosting Machine (LightGBM)*

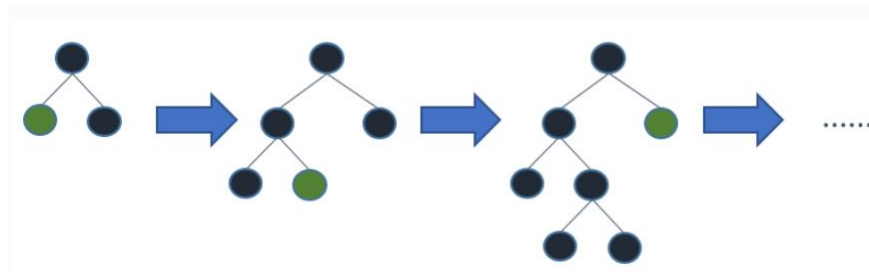
*Light Gradient-Boosting Machine (LightGBM)* merupakan suatu kerangka kerja *gradient boosting* yang memanfaatkan algoritma *ensemble* pembelajaran berbasis pohon [11]. Perilisan LightGBM dilakukan pada tanggal 17 Oktober 2016 sebagai bagian dari proyek Microsoft Distributed Machine Learning (DMTK) [12]. LightGBM dapat diaplikasikan dalam studi kasus regresi, klasifikasi biner, klasifikasi *multi-class*, *cross-entropy*, dan LambdaRank. Selain itu, LightGBM juga mendukung pembelajaran, paralel, distribusi, dan GPU [11].

LightGBM memanfaatkan algoritma berbasis histogram yang mengklasifikasikan nilai-nilai atribut kontinu menjadi bin diskrit. Hal tersebut memberikan keunggulan pada kinerja LightGBM, yaitu mengurangi konsumsi memori dan memiliki kecepatan *training* serta efisiensi yang lebih tinggi, sehingga LightGBM mampu untuk mengelola data dalam skala besar. Selain itu, LightGBM juga dapat dioptimalkan dengan baik untuk menangani fitur kategorik dan dapat menangani ketidakseimbangan kelas dengan baik [11].

Terdapat dua pendekatan utama untuk melatih pohon keputusan, yaitu *level-wise* dan *leaf-wise*. Sebagian besar algoritma pembelajaran pohon keputusan mengembangkan pohon dengan pendekatan konvensional *level-wise*, seperti yang diperlihatkan dalam ilustrasi berikut [11]:

Gambar 4. *Level-wise Tree Growth* [11]

Di sisi lain, LightGBM memperkenalkan strategi pertumbuhan *leaf-wise (best-first)* untuk melakukan optimasi pada aspek akurasi. Berbeda dengan pertumbuhan *level-wise*, pertumbuhan *leaf-wise* cenderung mencapai tingkat kerugian yang lebih rendah. Dengan begitu, LightGBM memiliki tingkat akurasi yang lebih baik [11].

Gambar 5. *Leaf-wise Tree Growth* [11]

Pada LightGBM, setiap iterasi dibangun dengan menambahkan prediksi model sebelumnya dengan kontribusi dari model baru sesuai dengan rumus berikut [13]:

$$f_m(x) = f_{m-1}(x) + T(x, \theta_m)$$

Keterangan:

$m$  : Jumlah iterasi

$f_m(x)$  : Prediksi model pada iterasi ke- $m$  untuk *input*  $x$ .

$f_{m-1}(x)$  : Prediksi model pada iterasi sebelumnya (ke- $(m-1)$ ) untuk *input*  $x$ .

$\theta$  : Parameter dari *decision tree*

$T(x, \theta_m)$  : *Decision tree* ke- $m$

Prediksi akhir merupakan hasil penjumlahan prediksi dari seluruh *decision tree* yang telah dibangun selama iterasi atau dapat dirumuskan sebagai berikut [13]:

$$f_M(x) = \sum_{m=1}^M T(x, \theta_m)$$

Keterangan:

$f_M(x)$  : *Boosting tree* yang mencakup sejumlah  $M$  *decision tree*

$m$  : Jumlah iterasi

$M$  : Jumlah *decision tree*

$f_m(x)$  : Prediksi model pada iterasi ke- $m$  untuk *input*  $x$ .

$\theta$  : Parameter dari *decision tree*

$T(x, \theta_m)$  : *Decision tree* ke- $m$

## 2.4. Python

Python merupakan bahasa pemrograman *general-purpose* yang pertama kali dirilis oleh Guido van Rossum pada tahun 1991. Popularitas dari bahasa ini mengalami pertumbuhan yang terus meningkat, menjadi pemimpin pada TIOBE index di tahun 2021 [14]. Program Python dijalankan menggunakan interpreter sehingga program dapat diuji secara langsung. Python dapat digunakan pada berbagai domain, mulai dari pengembangan *website*, *scripting*, *data mining*, hingga pengembangan *game* dan juga robotik [15].

Kelebihan dari Python antara lain adalah bersifat *open-source*, dapat digunakan pada *platform* yang beragam, relatif mudah untuk dimengerti, dan juga Python dilengkapi dengan *library* yang banyak dan beragam [15]. *Library* merupakan sekumpulan kode yang digunakan *developers* untuk mempercepat proses penulisan kode dari awal [16].



## 2.5. *Mutual Information*

*Mutual information* merupakan pendekatan yang berguna untuk data kategorikal maupun numerikal. *Mutual information* menerapkan konsep *information gain* untuk menyeleksi fitur (*feature selection*). Cara kerja dari *mutual information* adalah mengukur informasi bersama antara dua variabel dari data yang sama dan menilai pengurangan ketidakpastian pada satu variabel ketika nilai variabel lainnya diketahui. Nilai *mutual information* akan mencapai nol hanya jika dua variabel tersebut tidak memiliki hubungan satu sama lain. Semakin tinggi nilai dari *mutual information*, menunjukkan adanya hubungan yang semakin kuat antara dua variabel. Variabel yang dimaksud adalah fitur target dengan fitur-fitur lain [17], [18].

Scikit-learn, suatu pustaka *machine learning*, menyediakan implementasi dari konsep *mutual information* untuk melakukan seleksi fitur melalui fungsi `mutual_info_classif()`. Ketika melakukan seleksi fitur, dapat digunakan *class* `SelectKBest` yang berfungsi untuk menghilangkan semua fitur kecuali sejumlah  $k$  fitur dengan skor terbaik [18][19]. Perumusan dari *mutual information* adalah sebagai berikut [20]:

$$MI(U, V) = \sum_{i=1}^{|U|} \sum_{j=1}^{|V|} \frac{|U_i \cap V_j|}{N} \log \frac{N|U_i \cap V_j|}{|U_i||V_j|}$$

$MI(U, V)$  : *Mutual information* untuk  $U$  dan  $V$

$N$  : Jumlah keseluruhan objek

$|U_i|$  : Jumlah sampel pada  $U$

$|V_j|$  : Jumlah sampel pada  $V$

$|U_i \cap V_j|$  : Jumlah objek  $U$  dan  $V$  yang beririsan

Metrik bersifat simetris, yang berarti apabila mengganti  $U$  dengan  $V$  akan menghasilkan skor yang sama ( $MI(U, V) = MI(V, U)$ ) [20].

## 2.6. Confusion Matrix

*Confusion matrix* adalah representasi dari evaluasi kinerja model klasifikasi biner yang menggambarkan sejauh mana model tersebut berhasil mengklasifikasikan data dengan benar. *Confusion matrix* memiliki format berupa matriks persegi dengan ukuran  $n \times n$ , di mana  $n$  adalah jumlah kelas yang ada [8]. Gambaran dari *confusion matrix* dapat dilihat pada Gambar 6.

negative class	TN	FP
positive class	FN	TP
	predicted negative	predicted positive

Gambar 6. *Confusion Matrix* [8]

Kolom dalam *confusion matrix* mewakili hasil dari prediksi model, sedangkan baris mewakili label sebenarnya dari data uji. Untuk mengevaluasi sejauh mana model tersebut efektif, berbagai metrik seperti *precision*, *recall*, *F1 score*, dan *accuracy* dapat dihitung berdasarkan informasi yang terdapat dalam *confusion matrix* [8].

**Precision.** *Precision* mengukur perbandingan antara jumlah pengamatan positif yang diprediksi dengan benar (*true positive* (TP)) terhadap jumlah total pengamatan yang diprediksi sebagai positif (*true positive* (TP) dan *false positive* (FP)). Semakin sedikit FP, maka nilai *precision*-nya akan semakin meningkat [8]. Berikut merupakan perumusan dari *precision*:

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

**Recall.** Mengukur perbandingan antara jumlah pengamatan positif yang diprediksi dengan benar (*true positive* (TP)) dengan jumlah total pengamatan pada kelas yang

sebenarnya positif (*true positive* (TP) dan *false negative* (FN)). Semakin sedikit FN, maka nilai *recall* akan semakin tinggi[8]. Berikut merupakan perumusan dari *recall*:

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

***F1-Score***. Merupakan skor yang diperoleh dari menghitung rata-rata harmonis dari *precision* dan *recall*. F1-score memberikan penekanan lebih besar pada nilai *false negative* (FN) dan *false positive* (FP), serta mengabaikan pengaruh nilai *true negative* (TN) pada skornya [8]. Berikut merupakan perumusan dari *F1-score*:

$$F1\ Score = 2 \times \frac{precision \times recall}{precision + recall}$$

***Accuracy***. Merupakan rasio prediksi yang diklasifikasikan dengan benar terhadap seluruh data yang ada (semua entri dari *confusion matrix* yang dijumlahkan). Adapun prediksi yang diklasifikasikan dengan benar artinya mencakup jumlah pengamatan positif yang diprediksi dengan benar (*true positive* (TP)) dan jumlah pengamatan negatif yang diprediksi dengan benar (*true negative* (TN)) [8]. Berikut merupakan perumusan dari *accuracy*:

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

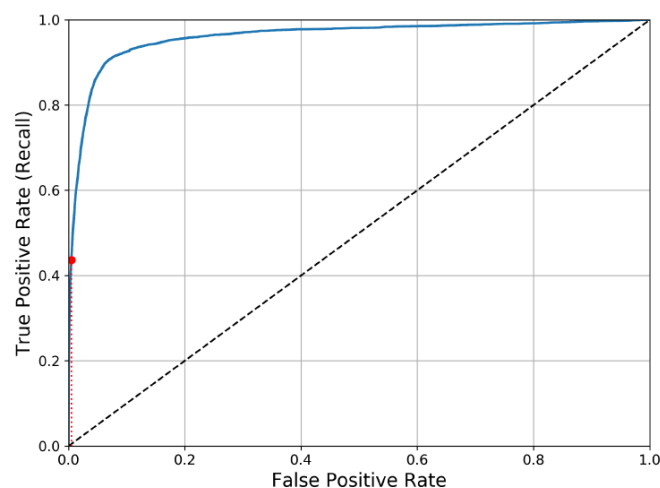
*Classification report* merupakan *built-in function* yang terdapat pada scikit-learn yang berfungsi untuk membuat laporan teks yang menampilkan metrik-metrik utama dari klasifikasi, yang memuat nilai *precision*, *recall*, *f1-score*, dan *accuracy* sebagaimana contohnya terdapat pada Gambar 2.7 [21].

	precision	recall	f1-score	support
class 0	0.50	1.00	0.67	1
class 1	0.00	0.00	0.00	1
class 2	1.00	0.67	0.80	3
accuracy			0.60	5
macro avg	0.50	0.56	0.49	5
weighted avg	0.70	0.60	0.61	5

Gambar 7. *Classification Report* [21]

## 2.7. Receiver Operating Characteristic Are Under the Curve (ROC AUC)

*Receiver Operating Characteristics Curve*, sering disingkat sebagai kurva ROC, adalah alat evaluasi yang mirip dengan kurva *precision-recall*. Alih-alih memberikan informasi tentang presisi dan recall, kurva ROC fokus pada dua metrik utama: tingkat positif palsu (*False Positive Rate* atau FPR) dan tingkat positif sejati (*True Positive Rate* atau TPR). Kurva ROC menggambarkan bagaimana model mengubah ambang batasnya, mempengaruhi FPR dan TPR. Ini membantu kita memahami sejauh mana model mampu memisahkan kelas positif dan negatif [8].



Gambar 8. Kurva ROC [12]

Area di bawah kurva ROC, yang sering disebut sebagai AUC (*Area Under the Curve*), adalah ukuran yang mengukur secara keseluruhan seberapa baik model

dapat membedakan kelas-kelas tersebut [8]. Semakin tinggi nilai AUC, semakin baik kinerja model dalam membedakan kelas positif dan negatif. Nilai AUC mencapai puncaknya pada 1 ketika semua titik positif memiliki skor yang lebih tinggi daripada semua titik negatif. Pada Gambar 2.8, garis putus-putus adalah representasi kurva ROC dari pengklasifikasi yang sepenuhnya acak; pengklasifikasi yang efektif akan berusaha untuk berada sejauh mungkin dari garis ini, yang mengarah ke sudut kiri atas. Sebuah pengklasifikasi yang sempurna akan memiliki ROC AUC setara dengan 1, sedangkan pengklasifikasi yang benar-benar acak akan memiliki ROC AUC sebesar 0,5 [12].

## 2.8. Flask

Flask merupakan sebuah *open-source light web framework* yang dibuat di Python untuk *deployment* aplikasi web [22]. Web merupakan gabungan dari berbagai halaman web, gambar, dan elemen lain yang saling terhubung membentuk sebuah dokumen yang lebih besar dan terstruktur. Dalam analogi, jika web adalah sebuah buku, maka setiap halaman adalah halaman web. Web bisa terdiri dari satu halaman atau bahkan ribuan halaman. Setiap web ditulis dalam kode, dan kode-kode tersebut menggambarkan tata letak, format, dan isi halaman. Bahasa pemrograman yang paling umum dan luas digunakan untuk membuat halaman web adalah HTML [23].

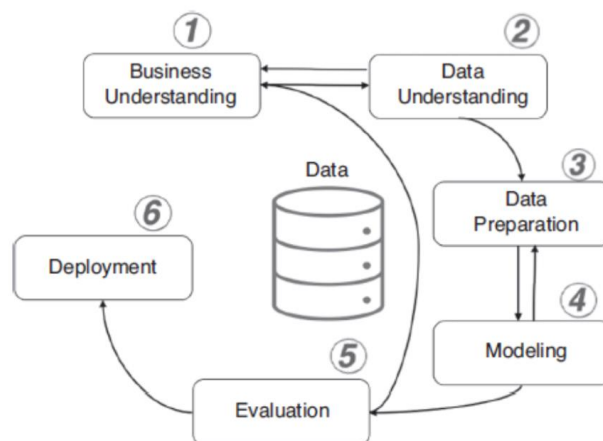
*Web framework* merupakan kumpulan sumber daya yang dibutuhkan guna mengoperasikan aplikasi web. Termasuk di dalamnya adalah modul, *library*, dan alat yang digunakan oleh pengembang untuk membangun dan mengoperasikan aplikasi dengan lancar [22].

Flask adalah sebuah kerangka kerja yang sederhana, sering disebut sebagai "*microframework*" dan mudah dipahami secara keseluruhan. Walaupun ringkas, Flask bukan berarti terbatas dalam fungsinya. Flask dibangun sebagai *framework* yang dapat diperluas dari bawah ke atas. Artinya, Flask menyediakan inti yang kuat dengan fungsi-fungsi esensial, sementara ekstensi menyediakan fungsi lainnya. Pengembang dapat memilih ekstensi apapun yang paling sesuai dengan kebutuhan

atau bahkan membuat secara khusus jika diperlukan. Hal ini berbeda dengan kerangka kerja yang lebih besar di mana banyak pilihan sudah ditentukan dan sulit untuk dimodifikasi atau diganti [24].

## 2.9. Cross-Industry Standard Process for Data Mining (CRISP-DM)

Pada tahun 1999, banyak perusahaan besar, termasuk Daimler-Benz, OHRA, NCR Corp, dan SPSS, Inc. meresmikan dan menetapkan standar bahwa pendekatan teknik *data mining* adalah CRISP-DM, Cross-Industry Standard for Data Mining. Metodologi CRISP-DM menyediakan sebuah pendekatan struktural dalam merencanakan sebuah proyek data-mining. Proses yang terdapat dalam metodologi ini dirancang untuk independen dari *tools* tertentu [25]. Proses yang terkandung pada CRISP-DM terdiri atas enam fase kegiatan sebagai berikut [26]:



Gambar 9. Tahapan CRISP-DM [26]

### 1. Business Understanding

Tahap awal ini berpusat pada pemahaman tujuan serta kebutuhan proyek dari sudut pandang bisnis. Kemudian, mendefinisikan masalah *data mining* dari informasi tersebut dan merancang strategi awal untuk memenuhi tujuan.

### 2. Data Understanding

Tahap *data understanding* dimulai dengan pengumpulan data awal dan dilanjutkan dengan kegiatan untuk mendapatkan yang pemahaman lebih dalam mengenai *dataset*, seperti mengidentifikasi kualitas data, dan/atau mendeteksi



subset yang menarik untuk membentuk hipotesis tentang informasi yang tersembunyi.

### 3. *Data Preparation*

*Data preparation* meliputi seluruh aktivitas yang diperlukan untuk membuat *dataset* akhir (data yang akan dimasukkan ke tahap *modeling*) dari data mentah. Aktivitas dalam tahap ini kemungkinan dilakukan beberapa kali dan tanpa ada urutan yang baku.

### 4. *Modeling*

Selama tahap ini, berbagai teknik dipilih dan diimplementasikan. Lalu, parameternya diatur ke tingkat yang ideal. Sebelum masuk ke pemodelan algoritma, *dataset* dibagi menjadi dua bagian, yaitu *data training* yang digunakan untuk menentukan *classifier* pada data yang akan dimasukkan selanjutnya dan *data testing* yang digunakan untuk mengetahui performa sistem yang telah dibuat berdasarkan data yang telah dilatih sebelumnya [27].

### 5. *Evaluation*

Pada titik ini, dilakukan penilaian model yang telah dibuat secara menyeluruh dan meninjau metode yang digunakan untuk memastikan bahwa model tersebut memenuhi tujuan bisnis. Tahap ini juga dilakukan untuk mendapatkan keyakinan bahwa hasil pemodelan bersifat valid dan dapat diandalkan sebelum melanjutkan ke tahap selanjutnya. Apabila belum cukup baik, dapat dilakukan iterasi kembali ke tahapan sebelumnya [26][27].

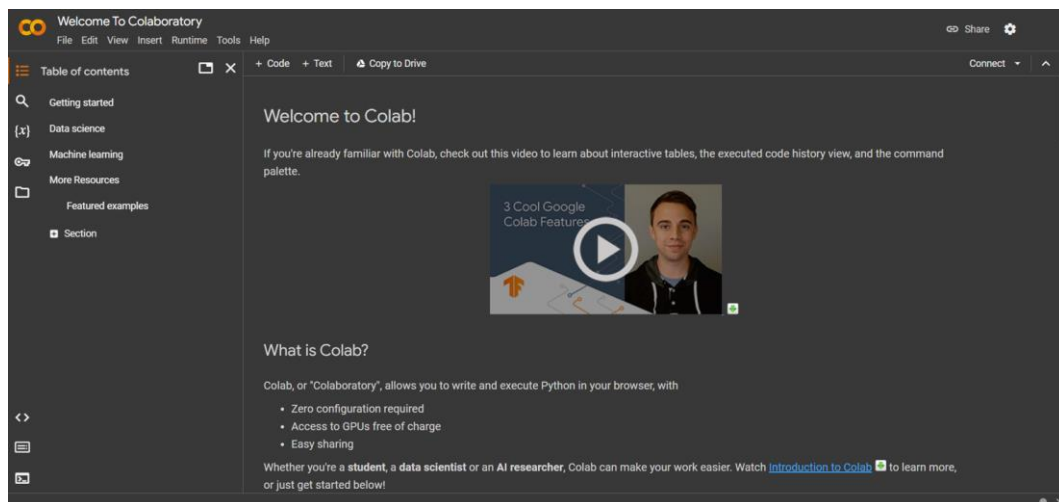
### 6. *Deployment*

Pada tahap ini, pengetahuan yang telah didapatkan dari tahap-tahap sebelumnya diatur dan disajikan sedemikian rupa agar pengguna dapat menggunakannya. Bergantung pada kebutuhan, fase *deployment* dapat sederhana maupun rumit. Pada tahap ini laporan yang mencakup hasil akhir dari proses juga dibuat.

## 2.10. Google Colab

Colaboratory, atau 'Colab', adalah produk pengembangan dari Google Research yang memungkinkan pengguna untuk menulis dan menjalankan kode Python melalui peramban web, sehingga sangat sesuai untuk *machine learning*, analisis data, dan pembelajaran. Colab bertujuan mempermudah penelitian *machine learning* dan *artificial intelligence* dengan mengatasi hambatan umum yang melibatkan kebutuhan sumber daya komputasi yang besar untuk belajar dan mencapai keberhasilan [28].

Sebagai lingkungan *notebook* Jupyter gratis berbasis *cloud* dari Google, Colab menyediakan GPU dan TPU secara gratis, fleksibel dalam konfigurasinya tanpa perlu pengaturan tambahan, dan memiliki tampilan yang *user-friendly*. *Notebook* dari Google Colab akan tersimpan secara otomatis di Google Drive. Tersedianya fitur kolaborasi antar *developer* juga menjadi keunggulan tambahan dari Google Colab [28].



Gambar 10. Tampilan Awal Google Colab

### 2.11. Visual Studio Code

Visual Studio Code (VS Code) adalah *editor* kode *open-source* yang dikembangkan oleh Microsoft, dapat diakses di berbagai platform, dan telah meraih predikat *The Most Popular Development Environment* dalam *Stack Overflow Developer Survey* tahun 2019. Selain menyediakan beragam fitur dan dapat disesuaikan sesuai pengguna, VS Code tidak hanya berfungsi untuk meng-*edit source code*, tetapi juga mendukung kolaborasi dan lingkungan *cloud-hosted environment*. Meskipun dukungan bawaannya terbatas pada Bahasa JavaScript, TypeScript, HTML, dan CSS, VS Code mendukung bahasa lain seperti Python melalui *extension*. Pengguna dapat memperluas fungsionalitas dari *editor* dengan menginstal berbagai *extension* dari Visual Studio Code Marketplace [29].

### 2.12. Penelitian Terdahulu

R. Ali, dkk., telah melakukan penelitian menggunakan teknologi pembelajaran mesin dengan Microsoft Azure untuk menciptakan sistem yang mampu melakukan prediksi terhadap diabetes berdasarkan *dataset* Pima Indian Diabetes. Sebanyak enam algoritma pembelajaran mesin digunakan (Logistic Regression, Artificial Neural Network (ANN), Decision Tree, Average Perceptron, Decision Forest, dan Support Vector Machine (SVM)) dan hasil terbaik diperoleh melalui penggunaan algoritma *decision forest*, di mana tingkat *accuracy*-nya sebesar 81,2%, *recall* sebesar 58,1%, *F1 Score* sebesar 67,9%, dan AUC sebesar 83% [30]. Kemudian, penelitian klasifikasi pada *dataset* penderita penyakit diabetes juga pernah dilakukan oleh A. M. Argina. Data yang digunakan pada penelitian adalah sebanyak 77 data dengan pembagian 90% data training dan 10% data *testing*. Hasil yang didapatkan dari penelitian ini adalah ketika besar  $K = 3$ , didapatkan akurasi tertinggi sebesar 39%, presisi tertinggi sebesar 65% (sama dengan saat  $K = 5$ ), *recall* tertinggi 36%, dan *F-measure* tertinggi sebesar 46% [31]. Kedua penelitian tersebut memiliki dua *output*, yaitu apakah seseorang menderita diabetes atau tidak [30][31].

Selanjutnya, terdapat lima penelitian yang membandingkan kinerja LightGBM dengan berbagai algoritma lain. Pertama, penelitian oleh D. D. Rufo, dkk., membandingkan algoritma KNN, SVM, *Naive Bayes*, *Bagging*, *Random Forest*, *XGBoost*, dan LightGBM untuk melakukan *early recognition* pada *diabetes mellitus*. Data yang digunakan pada penelitian ini dikumpulkan dari Rumah Sakit Memorial Zewditu di Addis Ababa, Ethiopia. Sebagai hasil akhir, LightGBM terbukti mengungguli kinerja algoritma lain dalam analisis dataset tersebut dengan tingkat *accuracy* sebesar 98,1%, AUC sebesar 98,1%, *sensitivity* 99,9%, dan *specificity* 96,3% [32]. Kedua, penelitian oleh R. Ahsana, dkk., meneliti kinerja algoritma AdaBoost dan LightGBM terhadap *dataset* Pima Indians Diabetes yang terdiri dari total 768 entri data. Hasil penelitian menunjukkan bahwa algoritma LightGBM memiliki tingkat akurasi yang lebih tinggi dibandingkan dengan algoritma AdaBoost. Akurasi LightGBM mencapai 91,67%, sedangkan algoritma AdaBoost hanya mencapai 91,14%. Selain itu, nilai AUC untuk LightGBM adalah 97,04%, sedangkan untuk AdaBoost adalah 96,93% [33]. Kedua penelitian tersebut memiliki dua target *output*, yaitu apakah seseorang menderita diabetes atau tidak.

Ketiga, penelitian dilakukan oleh F. I. Kurniadi dan P. D. Larasati menggunakan dataset dari The Electronic Health Record yang terdiri dari 548 data penderita stroke dan 28.524 non-stroke untuk mendeteksi secara dini penyakit stroke. Dua skenario diterapkan terhadap dataset tersebut, yaitu skenario penelitian tanpa melakukan seleksi fitur dan skenario berikutnya adalah menggunakan seleksi fitur dengan metode *variance threshold*. Adapun algoritma yang digunakan adalah LightGBM, SVM, serta Random Forest. Pada skenario pertama, hasil evaluasi menggunakan LightGBM dan SVM memberikan hasil yang sama, yaitu *accuracy* sebesar 98%, *precision* sebesar 50%, dan *recall* sebesar 49%, sedangkan hasil evaluasi menggunakan Random Forest menghasilkan *accuracy* sebesar 98%, *precision* sebesar 51%, dan *recall* sebesar 56%. Hal tersebut berarti hasil *precision* dan *recall* tertinggi diperoleh oleh Random Forest. Kemudian, pada skenario kedua, hasil evaluasi menggunakan LightGBM, SVM, maupun Random Forest memberikan hasil yang sama, yaitu *accuracy* sebesar 98%, *precision* sebesar 50%, dan *recall* sebesar 49%. Hasil yang didapatkan seimbang antara ketiga algoritma. Namun,

model yang dihasilkan menunjukkan bias yang signifikan, yang terindikasi melalui perbedaan mencolok antara nilai *precision* dan *recall* dibandingkan dengan nilai *accuracy* yang diperoleh dan bahwa proses seleksi fitur tidak begitu memberikan perbedaan [34].

Keempat, penelitian yang dilakukan oleh L. Sari, dkk., memanfaatkan LightGBM dan Random Forest untuk mengklasifikasikan potensial pelanggan, menggunakan Superstore Marketing Campaign Data dari situs Kaggle sebagai dataset. *Dataset* ini menunjukkan distribusi data yang tidak seimbang, dengan rasio 85% untuk kelas mayoritas dan 15% untuk kelas minoritas. Hasil penelitian menunjukkan bahwa pada data tidak seimbang, LightGBM mencapai hasil yang lebih baik daripada Random Forest, dengan *accuracy* sebesar 85,49%, *recall* 99,5%, *specificity* 0,45%, dan *precision* 0,85%, sementara Random Forest tidak dapat membuat model. Setelah menerapkan metode SMOTE untuk menyeimbangkan kelas, LightGBM mencapai tingkat *accuracy* sebesar 53,4% dan *specificity* 100%, sedangkan presisi dan recall-nya 0%. Di sisi lain, Random Forest mencapai *accuracy* 91,5%, *recall* 95%, *specificity* 88%, dan *precision* 88,9%. Penggunaan metode SMOTE memengaruhi *accuracy* Random Forest, tetapi tidak memengaruhi *accuracy* LightGBM. Secara keseluruhan, nilai *accuracy*, *recall*, *specificity*, dan *precision* dari Random Forest menghasilkan nilai yang baik dibandingkan dengan LightGBM pada data yang seimbang. LightGBM, di sisi lain, terbukti mampu menangani dataset yang tidak seimbang [35].

Kelima, terdapat penelitian yang dilakukan dilakukan E. Al Daoud untuk menginvestigasi dan membandingkan kinerja dari tiga metode gradien (XGBoost, LightGBM, dan CatBoost) terhadap *home credit dataset*. Dataset terdiri dari 219 fitur dan 356.251 data. Hasil implementasi menunjukkan bahwa LightGBM lebih cepat dan akurat (Skor AUC 79%) dibandingkan dengan CatBoost dan XGBoost dengan menggunakan berbagai jumlah fitur dan data [36].

Selanjutnya, P. O. Odion dan E. O. Ogbonnia melakukan penelitian yang menghasilkan sebuah *website* untuk melakukan diagnosis tifus dan malaria

menggunakan algoritma XGBoost. Setelah itu, *classifiers* di-deploy menggunakan kerangka kerja web Flask. Klasifikasi yang dilakukan adalah klasifikasi *binary* (dua kelas target) dan klasifikasi *multiclass* untuk penyakit tifus dan malaria. Klasifikasi penyakit malaria (*Binary*) menghasilkan *accuracy* 98.6% dan *F1 Score* 99.2%. Klasifikasi penyakit malaria (*Multiclass*) menghasilkan *accuracy* 97.6% dan *F1 Score* 96.8%. Klasifikasi penyakit tifus (*Binary*) menghasilkan *accuracy* 96.1% dan *F1 Score* 98.5%. Klasifikasi penyakit tifus (*Multiclass*) menghasilkan *accuracy* 96.1% dan *F1 Score* 95.1% [37].

Kemudian, terdapat penelitian oleh K. Srivastava dan D. K. Choubey yang melakukan klasifikasi terhadap Cleveland Heart Disease Dataset menggunakan algoritma KNN. Penelitian menghasilkan *accuracy* sebesar 87%. Selain itu, terbentuk sebuah aplikasi web menggunakan Flask dalam bahasa pemrograman Python [38].

Tabel 1. Penelitian Terdahulu

No	Peneliti	Algoritma	Dataset	Hasil
1	R. Ali, A. Raheem, A. Kadhum, M. Al-Qurabat [30]	Decision Forest	Pima Indian Diabetes	<i>Accuracy</i> : 81,2% <i>Recall</i> : 58,1% <i>F1 Score</i> : 67,9% AUC : 83%
2	A. M. Argina [31]	KNN	Dataset penderita penyakit diabetes yang terdiri dari 77 data.	<i>Accuracy</i> : 39% <i>Precision</i> : 65% <i>Recall</i> : 36% <i>F-measure</i> : 46%
3	D. D. Rufo, T. G. Debelee, A. Ibenthal, W. G. Negera [32]	LightGBM	Data diabetes dari Rumah Sakit Memorial Zewditu	<i>Accuracy</i> : 98,1% AUC : 98,1% <i>Sensitivity</i> : 99,9% <i>Specitivity</i> : 96,3%
4	R. Ahsana, R. Rohmat S, V. P. Widartha [33]	AdaBoost, LightGBM	Pima Indian Diabetes	LightGBM <i>Accuracy</i> : 91,67% AUC : 97,04% AdaBoost <i>Accuracy</i> : 91,14% AUC : 96,93%



Tabel 1. Penelitian Terdahulu (Lanjutan)

No	Peneliti	Algoritma	<i>Dataset</i>	Hasil
5	F. I. Kurniadi, P. D. Larasati [34]	LightGBM, SVM, Random Forest	The Electronic Health Record	<p>Skenario 1 LightGBM, SVM: <i>Accuracy: 98%</i> <i>Precision: 50%</i> <i>Recall: 49%</i></p> <p>Random Forest: <i>Accuracy: 98%</i> <i>Precision: 51%</i> <i>Recall: 56%</i></p> <p>Skenario 2 LightGBM, SVM, dan Random Forest: <i>Accuracy: 98%</i> <i>Precision: 50%</i> <i>Recall: 49%</i></p>
6	L. Sari, A. Romadloni, R. Lityanigrum, H. D. Hastuti [35]	LightGBM, Random Forest	Superstore Marketing Campaign Data	<p><i>Imbalance Data:</i> LightGBM: <i>Accuracy: 85,6%</i> <i>Recall: 99,5%</i> <i>Specificity: 0,45%</i> <i>Precision: 0,85%</i></p> <p>Random Forest: <i>Accuracy: -</i> <i>Recall: -</i> <i>Specificity: -</i> <i>Precision: -</i></p> <p><i>Balanced Data:</i> LightGBM: <i>Accuracy: 53,4%</i> <i>Recall: -</i> <i>Specificity: 100%</i> <i>Precision: -</i></p>

Tabel 1. Penelitian Terdahulu (Lanjutan)

No	Peneliti	Algoritma	Dataset	Hasil
				Random Forest: <i>Accuracy</i> : 91,5% <i>Recall</i> : 95% <i>Specificity</i> : 88% <i>Precision</i> : 88,9%
7	E. Al Daoud [36]	XGBoost, CatBoost, LightGBM	<i>Home credit dataset</i>	LightGBM: AUC : 79%
8	P. O. Odion, E. O. Ogbonnia [37]	XGBoost	Diagnostic Centre Newni, Anambra, Nigeria (Malaria Dataset)	Malaria ( <i>Binary</i> ): <i>Accuracy</i> : 98,6% <i>F1 Score</i> : 99,2%  Malaria ( <i>Muticlass</i> ): <i>Accuracy</i> : 97,6% <i>F1 Score</i> : 96,8%  Tifus ( <i>Binary</i> ): <i>Accuracy</i> : 96,1% <i>F1 Score</i> : 98,5%  Tifus ( <i>Multiclass</i> ): <i>Accuracy</i> : 96,1% <i>F1 Score</i> : 95,1%
9	K. Srivastava, D. K. Choubey [38]	KNN	Cleveland Heart Disease Dataset	<i>Accuracy</i> : 87%

Dengan mempertimbangkan kelebihan dan kekurangan dari penelitian terdahulu yang disajikan pada narasi maupun Tabel 1, pada penelitian ini dilakukan klasifikasi penyakit *diabetes mellitus* dengan menggunakan algoritma LightGBM terhadap Ghanaian Diabetes Dataset (data dari suatu fasilitas kesehatan Ghana) karena pada penelitian terdahulu, LightGBM sering kali memiliki akurasi yang tinggi. Kemudian, mengimplementasikan *framework* Flask untuk membuat antarmuka *website* yang dapat mengklasifikasikan penyakit diabetes berdasarkan *input* yang dimasukan oleh pengguna. Setelah itu, terdapat dua kemungkinan *output* yang disajikan, yaitu tidak menderita diabetes dan terindikasi diabetes tipe II.

### III. METODOLOGI PENELITIAN

#### 3.1. Waktu dan Tempat Penelitian

Adapun waktu dan tempat untuk pengerjaan penelitian ini adalah:

Waktu Penelitian : September 2023 sampai dengan Februari 2024

Tempat Penelitian : Laboratorium Komputer Jurusan Teknik Elektro  
Universitas Lampung

Tabel 2. Jadwal Penelitian

No	Aktivitas	2023				2024	
		Sep	Okt	Nov	Des	Jan	Feb
1	Studi Literatur						
2	Persiapan Alat dan Bahan						
3	<i>Business Understanding</i>						
4	<i>Data Understanding</i>						
5	<i>Data Preparation</i>						
6	<i>Modeling</i>						
7	<i>Evaluation</i>						
8	<i>Deployment</i>						
9	Analisis Hasil						

### 3.2. Alat dan Bahan Penelitian

Adapun alat yang digunakan dalam penelitian ini terdiri dari *hardware* dan *software* sebagaimana dijelaskan pada Tabel 3 berikut:

Tabel 3. Alat Penelitian

Jenis	No	Perangkat	Spesifikasi	Kegunaan
<i>Hardware</i>	1	Laptop	Intel i5-8265U RAM 8 GB	Perangkat dalam pembuatan serta pengujian sistem.
	2	Python	Versi 3.8.10	Bahasa pemrograman pembuatan sistem.
<i>Software</i>	3	Google Colaboration		<i>Text editor</i> untuk pengembangan sistem.
	4	Flask	Versi 2.3.3	<i>Web framework</i> yang digunakan dalam pembangunan <i>project</i> .
	5	Visual Studio Code	Versi 1.74.3	<i>Text editor</i> untuk pengembangan sistem.
	6	lightgbm	Versi 4.0.0	<i>Library</i> untuk menjalankan klasifikasi LightGBM.
	7	sklearn	Versi 1.2.2	<i>Library</i> yang memiliki berbagai fungsi dan kelas untuk melakukan proses <i>modeling</i> , <i>feature selection</i> , dan <i>evaluation</i> .
	8	joblib	Versi 1.3.2	<i>Library</i> untuk menyimpan model klasifikasi.
	9	pyarrow	Versi 1.23.5	<i>Library</i> untuk menyimpan maupun membaca <i>dataset</i> berekstensi '.feather'.

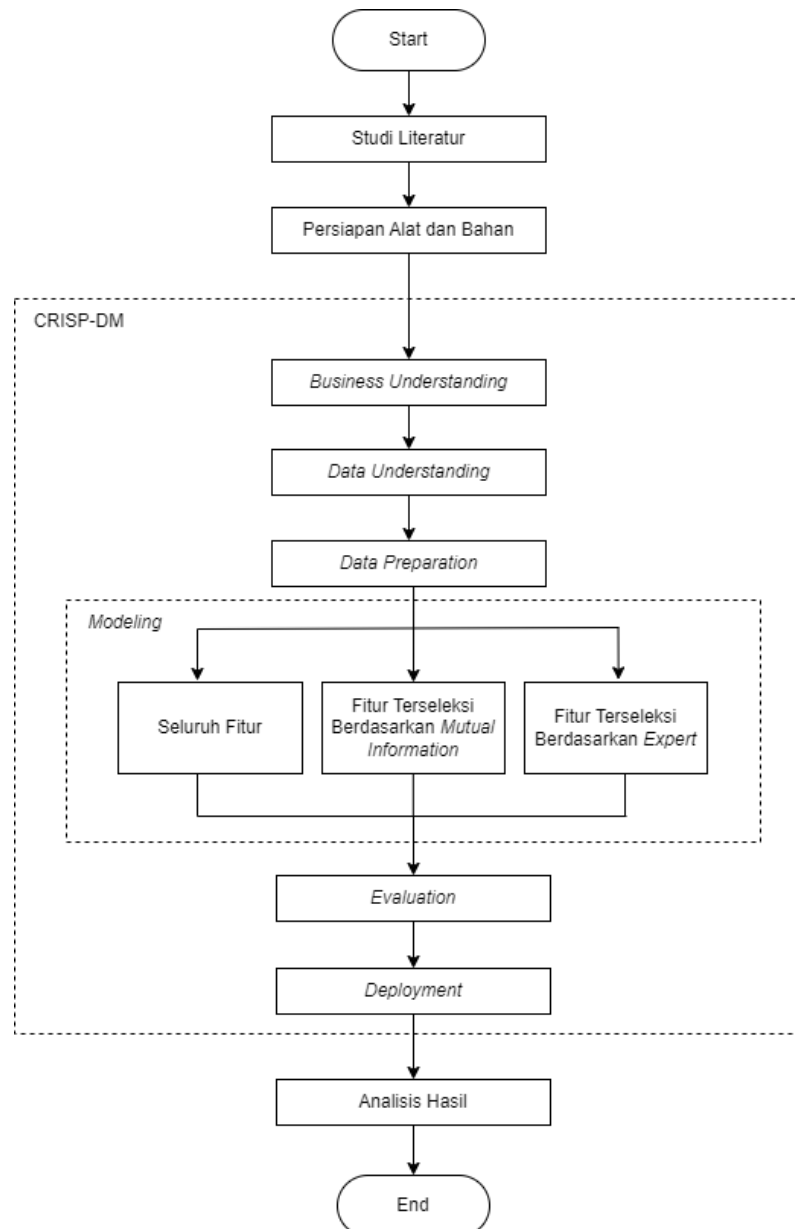
Bahan penelitian yang digunakan adalah *dataset* berupa 3.415 data dari Ghanaian Diabetes Dataset (data dari suatu fasilitas kesehatan Ghana) yang dapat diakses melalui situs kaggle.com (<https://www.kaggle.com/datasets/devkyle/diabetes-dataset?select=Diabetes+%281%29.csv>) [39] dengan distribusi data sebagai berikut:

Tabel 4. Distribusi Jumlah *Dataset*

Kelas	Jumlah	Persentase
0 (Tidak menderita diabetes)	1.176	34.40%
2 (Diabetes tipe II)	2.239	65.6%
<b>Total</b>	<b>3.415</b>	<b>100%</b>

### 3.3. Tahapan Penelitian

Tahapan penelitian diawali dengan melakukan studi literatur mengenai permasalahan terkait. Kemudian, menyiapkan berbagai alat (baik *hardware* maupun *software*) dan bahan yang digunakan pada penelitian. Setelah itu, menggunakan metode tahapan CRISP-DM untuk penelitian. Kemudian, dilanjutkan dengan analisis hasil dari penelitian. *Flowhcart* tahapan ditunjukkan pada Gambar 11.

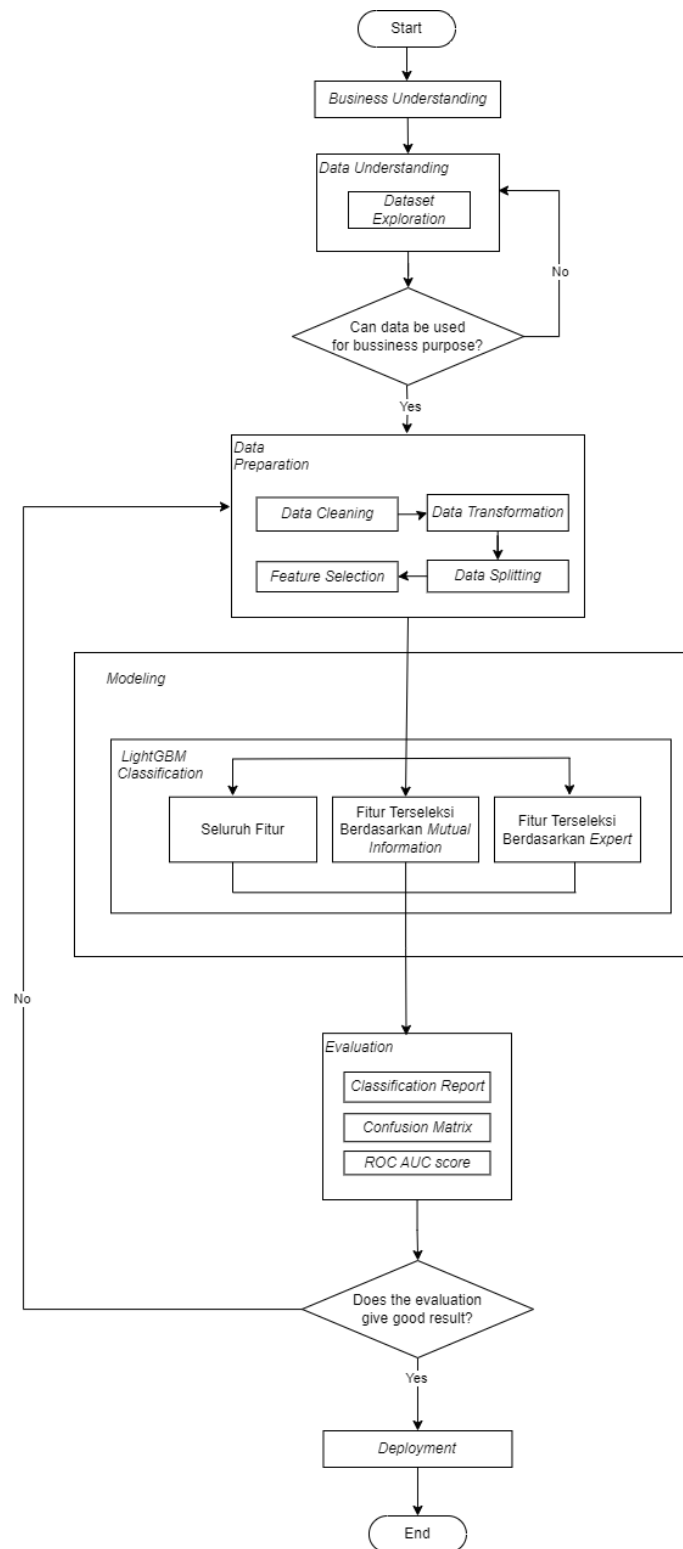


Gambar 11. *Flowchart* Tahapan Penelitian

### 3.4. Tahapan CRISP-DM

Data penelitian berupa Ghanaian Diabetes Dataset (data dari suatu fasilitas kesehatan Ghana). Data diperoleh dari situs kaggle.com dan diolah dengan LightGBM. Adapun metode yang diterapkan adalah *Cross Industry Standard Process for Data Mining* (CRISP-DM). Berdasarkan CRISP-DM, berikut

merupakan diagram tahapan dalam proses pengembangan sistem dalam penelitian ini:



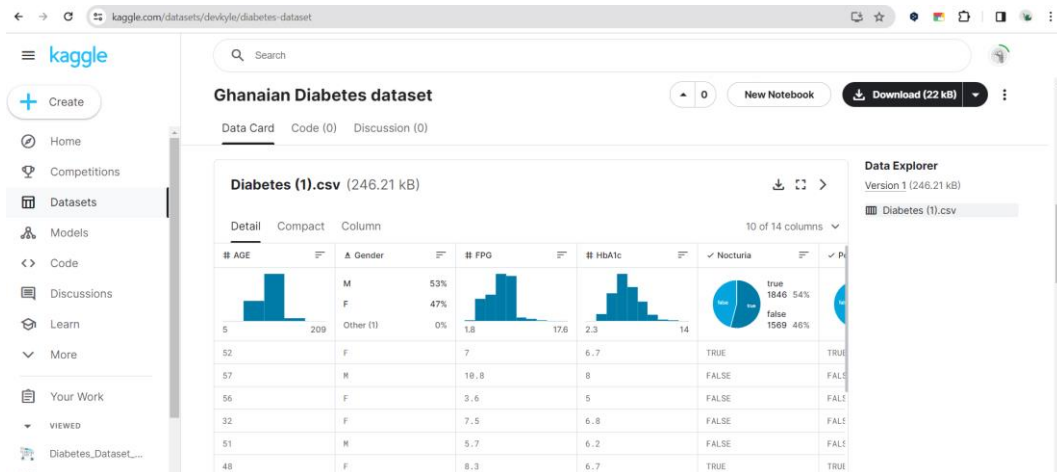
Gambar 12. Tahapan CRISP-DM

### 3.4.1. Business Understanding

Penelitian ini memiliki tujuan untuk membuat antarmuka yang memfasilitasi pengguna dalam melakukan klasifikasi penyakit diabetes berdasarkan suatu *dataset* menggunakan algoritma LightGBM. Selain itu, dilakukan verifikasi parameter *dataset* dilakukan dengan melibatkan seorang *expert* melalui wawancara.

### 3.4.2. Data Understanding

*Dataset* yang digunakan pada penelitian ini berupa 3.415 data dari Ghanaian Diabetes Dataset (data dari suatu fasilitas kesehatan Ghana). Kelas target memiliki dua kelas, 0 yang mengindikasikan tidak menderita diabetes dan 1 yang mengindikasikan *diabetes mellitus* tipe II. Pada tahap ini juga dilakukan *dataset exploration* yang mencakup pengecekan informasi *dataset*, *missing values*, *zero count*, dan distribusi kelas target.



Gambar 13. Ghanaian Diabetes Dataset dari kaggle.com [39]



### 3.4.3. *Data Preparation*

Pada tahap ini dilakukan penanganan *missing values* dan *zero count*, transformasi data berupa *label encoding* dan pengubahan tipe data dari object dan boolean menjadi category, serta terakhir melakukan *feature selection* menggunakan *mutual information*, dan *data splitting* dengan rasio 80 *data training* : 20 *data testing* untuk digunakan pada tahap *modeling*.

### 3.4.4. *Modeling*

Pada tahap ini digunakan LightGBM untuk melatih *dataset*. Lalu, *data testing* akan diuji dalam model *classifier* yang telah terbentuk. Model diujicobakan pada data yang menggunakan seluruh fitur, fitur-fitur terseleksi dari *mutual information*, dan fitur-fitur rekomendasi dari *expert*.

### 3.4.5. *Evaluation*

Setelah proses pengklasifikasian menggunakan algoritma LightGBM selesai dilakukan, maka akan dilanjutkan ke tahap evaluasi. Evaluasi pada penelitian ini dilakukan menggunakan *classification report* (*recall*, *precision*, *f1-score*, *accuracy*), *confusion matrix*, dan ROC AUC *score*.

### 3.4.6. *Deployment*

Model *final* yang telah didapat dari tahapan-tahapan yang dilakukan selanjutnya di-*deploy* menggunakan *web framework* Flask. Pada tahap ini terdapat pengaturan tampilan antarmuka *website*, pengambilan *input* dari pengguna, proses konversi tipe data *input* pengguna agar sesuai dengan data pelatihan yang digunakan, dan pengaturan untuk melakukan prediksi *input* serta menampilkan hasil prediksi pada antarmuka.

## V. KESIMPULAN DAN SARAN

### 5.1. Kesimpulan

Berdasarkan hasil penelitian yang telah dilakukan, maka didapatkan kesimpulan sebagai berikut:

1. Klasifikasi *diabetes mellitus* tipe II dengan LightGBM terhadap Ghanaian Diabetes Dataset dapat dilakukan menggunakan dua skenario, yaitu menggunakan fitur hasil tes laboratorium (skenario 1) dan tanpa fitur hasil tes laboratorium (skenario 2). Pada masing-masing skenario, terdapat 3 konfigurasi fitur, yaitu konfigurasi seluruh fitur (dua belas fitur pada skenario 1 dan sebelas fitur pada skenario 2), fitur terseleksi berdasarkan *mutual information* (delapan fitur pada skenario 1 dan tujuh fitur pada skenario 2), dan fitur rekomendasi berdasarkan *expert* (delapan fitur pada skenario 1 dan tujuh fitur pada skenario 2).
2. Berdasarkan evaluasi, model memiliki kinerja yang baik pada skenario 1 dan 2 ketika menggunakan seluruh fitur (dua belas fitur pada skenario 1 dan sebelas fitur pada skenario 2), fitur terseleksi berdasarkan *mutual information* (delapan fitur pada skenario 1 dan tujuh fitur pada skenario 2), dan fitur rekomendasi berdasarkan *expert* (delapan fitur pada skenario 1 dan tujuh fitur pada skenario 2). Dengan begitu, fitur rekomendasi berdasarkan *expert* pada skenario 2 digunakan untuk pemodelan karena lebih relevan dengan ilmu kedokteran.
3. Antarmuka untuk prediksi *diabetes mellitus* tipe II berhasil dibangun dengan menggunakan *web framework* Flask. Antarmuka digunakan dengan cara

melakukan *input* data yang diperlukan yang kemudian hasil prediksi akan ditampilkan pada antarmuka.

## 5.2. Saran

Saran untuk penelitian lebih lanjut adalah sebagai berikut:

1. Klasifikasi *diabetes mellitus* tipe II dapat menambahkan atribut sesuai dengan rekomendasi dari dokter, yang mencakup pola makan (seperti frekuensi makan sayuran dan buah-buahan), aktivitas fisik minimal 3x/minggu dengan durasi minimal 30 menit, dan gaya hidup (kebiasaan minum alkohol dan merokok).
2. Pemodelan klasifikasi *diabetes mellitus* tipe II terhadap Ghanaian Diabetes Dataset menggunakan algoritma lain, seperti algoritma XGBoost atau CatBoost untuk membandingkan hasil kinerjanya dengan LightGBM.

## **DAFTAR PUSTAKA**

## DAFTAR PUSTAKA

- [1] International Diabetes Federation, *IDF Diabetes Atlas 10th Edition*. 2021, 2021. [Online]. Available: [www.diabetesatlas.org](http://www.diabetesatlas.org)
- [2] World Health Organization, "Diabetes." Accessed: May 05, 2023. [Online]. Available: <https://www.who.int/news-room/fact-sheets/detail/diabetes>
- [3] The Institute for Health Metrics and Evaluation, "Indonesia Health Research." Accessed: May 05, 2023. [Online]. Available: <https://www.healthdata.org/indonesia>
- [4] World Health Organization, *Classification Of Diabetes Mellitus 2019*. World Health Organization, 2019. [Online]. Available: <http://apps.who.int/bookorders>.
- [5] P. R. Febrinasari, T. A. Sholikah, D. N. Pakha, and S. E. Putra, *Buku Saku Diabetes Melitus untuk Awam*, 1st ed. Surakarta: UNS PRESS, 2020.
- [6] Pusat Data dan Informasi Kementerian Kesehatan RI, *Infodatin 2020 Diabetes Melitus*. Jakarta Selatan: Kementerian Kesehatan RI, 2020.
- [7] A. Rostamizadeh and A. Talwalkar, *Foundations of Machine Learning*, 2nd ed. Massachusetts: The MIT Press, 2018.
- [8] A. C. Müller and S. Guido, *Introduction to Machine Learning with Python*. Sebastopol: O'Reilly, 2017.

- [9] P. Jahangiry, “Boosting Models,” *Diakses melalui* [https://github.com/PJalgotrader/Machine\\_Learning-USU/blob/main/Lectures%20and%20codes/Module%2010-%20Bagging%20and%20Boosting/Module%2010-%20Part%202-%20Boosting%20models.pdf](https://github.com/PJalgotrader/Machine_Learning-USU/blob/main/Lectures%20and%20codes/Module%2010-%20Bagging%20and%20Boosting/Module%2010-%20Part%202-%20Boosting%20models.pdf). Utah State Univeristy, Logan, Aug. 2023.
- [10] Gautam Kunapuli, *Ensemble Methods for Machine Learning*. New York: Manning Publications Co., 2023.
- [11] Microsoft Corporation, “LightGBM Features.” Accessed: Jul. 12, 2023. [Online]. Available: <https://lightgbm.readthedocs.io/en/stable/Features.html#references>
- [12] B. Quinto, *Next-generation machine learning with spark: Covers XGBoost, LightGBM, Spark NLP, distributed deep learning with keras, and more*. Apress Media LLC, 2020. doi: 10.1007/978-1-4842-5669-5.
- [13] F. Wang, H. Cheng, H. Dai, and H. Han, “Freeway Short-Term Travel Time Prediction Based on LightGBM Algorithm,” in *IOP Conf. Series: Earth and Environmental Science*, 2020.
- [14] A. Martelli, A. M. Ravenscroft, S. Holden, and P. McGuire, *Python in a Nutshell*, 4th ed. Sebastopol: O’Reilly, 2023.
- [15] M. Lutz, *Learning Python*, 5th Edition. Sebastopol: O’Reilly, 2013.
- [16] J. M. Ortega, *Mastering Python for Networking and Security*. Brimingham: Packt Publishing, 2018.
- [17] Jason Brownlee, *Imbalanced Classification with Python*. Machine Learning Mastery, 2020.

- [18] Scikit Learn, “sklearn.feature\_selection.mutual\_info\_classif.” Accessed: Oct. 11, 2023. [Online]. Available: [https://scikit-learn.org/stable/modules/generated/sklearn.feature\\_selection.mutual\\_info\\_classif.html#r50b872b699c4-2](https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.mutual_info_classif.html#r50b872b699c4-2)
- [19] Scikit Learn, “sklearn.feature\_selection.SelectKBest.” Accessed: Oct. 11, 2023. [Online]. Available: [https://scikit-learn.org/stable/modules/generated/sklearn.feature\\_selection.SelectKBest.html#sklearn.feature\\_selection.SelectKBest](https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.SelectKBest.html#sklearn.feature_selection.SelectKBest)
- [20] Scikit Learn, “Mutual Info Score,” [https://scikit-learn.org/stable/modules/generated/sklearn.metrics.mutual\\_info\\_score.html](https://scikit-learn.org/stable/modules/generated/sklearn.metrics.mutual_info_score.html).
- [21] Scikit Learn, “Classification Report.” Accessed: Jul. 12, 2023. [Online]. Available: [https://scikit-learn.org/stable/modules/model\\_evaluation.html#classification-report](https://scikit-learn.org/stable/modules/model_evaluation.html#classification-report)
- [22] P. Singh, *Deploy Machine Learning Models to Production*. Karnataka: Apress, 2021.
- [23] J. Webb, *Web Development and Design for Beginners*. Alberta Inc., 2020.
- [24] M. Grinberg, *Flask Web Development: Developing Web Applications with Python*, 2nd ed. Sebastopol: O’Reilly, 2018.
- [25] M. Kantardzic, *Data Mining Concepts, Models, Methods, and Algorithms*, 3rd ed. New Jersey: John Wiley & Sons, Inc., 2020.
- [26] A. Fortino, *Data Mining and Predictive Analytics for Business Decisions*. Mercury Learning and Information LLC, 2023.

- [27] F. Provost and T. Fawcett, *Data Science for Business*. Sebastopol: O'Reilly, 2013.
- [28] P. Naik, G. Naik, and M. B. Patil, *Conceptualizing Python in Google Colab*. Bilaspur: Shashwat Publication, 2021.
- [29] A. Speight, *Visual Studio Code for Python Programmers*. Hoboken: John Wiley & Sons, Inc., 2021.
- [30] R. Ali, A. Raheem, A. Kadhum, M. Al-Qurabat, الرحيم عبد, and علي رشا, "Developing a Predictive Health Care System for Diabetes Diagnosis as a Machine Learning-Based Web Service," *Journal of University of Babylon for Pure and Applied Science (JUBPAS)*, vol. 30, no. 1, pp. 1–32, Mar. 2022.
- [31] A. M. Argina, "Penerapan Metode Klasifikasi K-Nearest Neighbor pada Dataset Penderita Penyakit Diabetes," *Indonesian Journal of Data and Science*, vol. 1, no. 2, pp. 29–33, Jul. 2020.
- [32] D. D. Rufo, T. G. Debelee, A. Ibenthal, and W. G. Negera, "Diagnosis of Diabetes Mellitus using Gradient Boosting Machine (LightGBM)," *Diagnostics*, vol. 11, no. 9, Sep. 2021.
- [33] R. Ahsana, R. Rohmat Saedudin, and V. P. Widartha, "Perbandingan Akurasi Algoritma AdaBoost dan Algoritma LightGBM untuk Klasifikasi Penyakit Diabetes," *e-Proceeding of Engineering*, vol. 8, no. 5, pp. 9738–9748, Oct. 2021.
- [34] F. I. Kurniadi and P. Larasati, "Light Gradient Boosting Machine untuk Deteksi Penyakit Stroke," *Jurnal Sistem Komputer dan Kecerdasan Buatan*, vol. VI, no. 1, pp. 67–72, Sep. 2022.



- [35] L. Sari, A. Romadloni, R. Lityaningrum, and H. D. Hastuti, "Implementation of LightGBM and Random Forest in Potential Customer Classification," *TIERS Information Technology Journal*, vol. 4, no. 1, pp. 43–55, Jun. 2023.
- [36] E. Al Daoud, "Comparison between XGBoost, LightGBM and CatBoost Using a Home Credit Dataset," *International Journal of Computer and Information Engineering*, vol. 13, no. 1, pp. 6–10, 2019.
- [37] P. O. Odion and E. O. Ogbonnia, "Web-Based Diagnosis of Typhoid and Malaria using Machine Learning," *NDA Journal of Military Science and Disciplinary Studies*, vol. 1, no. 2, pp. 85–99, Jul. 2022.
- [38] K. Srivastava and D. K. Choubey, "Heart Disease Prediction using Machine Learning and Data Mining," *International Journal of Recent Technology and Engineering (IJRTE)*, vol. 9, no. 1, pp. 212–219, May 2020.
- [39] Devkyle, "Ghanaian Diabetes Dataset," <https://www.kaggle.com/datasets/devkyle/diabetes-dataset?select=Diabetes+%281%29.csv>.
- [40] Perkumpulan Endokrinologi Indonesia (PERKENI), *Pedoman Pengelolaan dan Pencegahan Diabetes Melitus Tipe 2 di Indonesia*. PB. PERKENI, 2021.
- [41] Siloam Hospitals, "Cara Menghitung BMI." Accessed: Jul. 15, 2023. [Online]. Available: <https://www.siloamhospitals.com/informasi-siloam/artikel/cara-menghitung-bmi>
- [42] Larxel, "Early Classification of Diabetes," <https://www.kaggle.com/datasets/andrewmvd/early-diabetes-classification>.