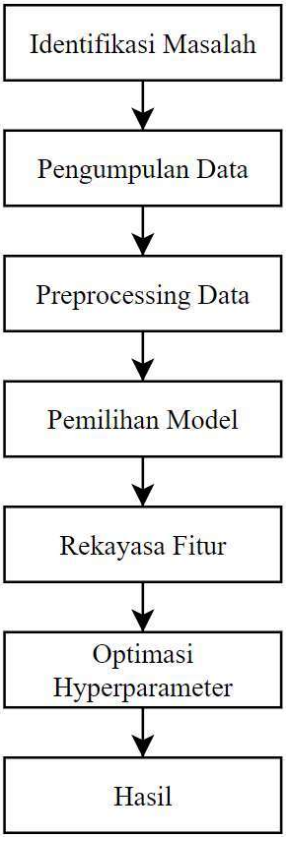
BAB III

METODE PENELITIAN

3.1. Desain Penelitian

Desain penelitian ini disusun untuk menjawab permasalahan terkait kebutuhan deteksi dini penyakit diabetes dengan pendekatan otomatis berbasis AI. Penelitian ini difokuskan pada pengembangan model prediksi yang dibangun menggunakan *platform* IBM Watson Studio dengan fitur *AutoAI*. Model menggunakan pendekatan algoritma *binary classification* untuk mengelompokkan data ke dalam dua kelas, yaitu 0 (tidak berisiko) dan 1 (berisiko).



Gambar 3. 1 Desain Peneilitian
Sumber: Data Penelitian, 2025

Langkah-langkah penelitian pada Gambar 3.1 dapat dijelaskan sebagai berikut:

1. Identifikasi Masalah

   Penelitian diawali dengan mengidentifikasi permasalahan utama, yaitu tingginya jumlah kasus diabetes dan belum optimalnya deteksi dini dengan metode manual. Dari permasalahan tersebut, dirumuskan tujuan untuk membangun model prediksi otomatis menggunakan teknologi kecerdasan buatan yang mampu memberikan hasil yang cepat dan akurat.

2. Pengumpulan Data

   Data yang digunakan dalam penelitian diperoleh dari *platform* Kaggle. Dataset ini berisi parameter medis penting yang berkaitan dengan diabetes, yaitu: *Pregnancies*, *Glucose*, *BloodPressure*, *SkinThickness*, Insulin, BMI, *DiabetesPedigreeFunction*, dan *Age*. Data ini menjadi dasar dalam proses pemodelan dan analisis.

3. *Pre-processing* Data

   Data yang telah dikumpulkan kemudian melalui tahap *pre-processing*. Data dibersihkan secara manual untuk menghapus duplikasi dan menangani nilai kosong. Dataset kemudian dibagi menjadi data *training* dan data *test* secara manual untuk keperluan pemodelan.

4. Pemilihan Model

   Proses ini dilakukan secara otomatis oleh *AutoAI*, yang akan mencoba berbagai algoritma *binary classification*, seperti *Logistic Regression*, *Random Forest*, dan SVM, untuk menemukan model terbaik berdasarkan akurasi.

5. Rekayasa Fitur

   AutoAI secara otomatis melakukan transformasi dan seleksi fitur untuk meningkatkan performa model. Tahapan ini bertujuan untuk mengubah fitur mentah menjadi bentuk yang lebih informatif dan relevan, sehingga dapat meningkatkan kinerja model klasifikasi secara otomatis.

6. Optimasi *Hyperparameter*

   Optimasi *hyperparameter* adalah proses penyetelan nilai-nilai eksternal algoritma sebelum pelatihan dimulai, guna memperoleh performa prediksi yang optimal.

7. Hasil

   Model yang terbentuk menghasilkan keluaran dalam bentuk *binary classification*, yaitu 0 untuk menunjukkan bahwa seseorang tidak berisiko menderita diabetes, dan 1 untuk menunjukkan bahwa seseorang berisiko menderita diabetes. Hasil prediksi ini ditampilkan oleh sistem berdasarkan *input* dari delapan parameter medis, dengan algoritma terbaik yang telah dipilih dan dioptimalkan secara otomatis oleh *AutoAI*.

3.2. Metode Pengumpulan Data

Dalam penelitian ini, data dikumpulkan secara tidak langsung melalui sumber dataset terbuka (*open dataset*) yang tersedia secara publik di *platform* Kaggle. Dataset tersebut berisi informasi medis yang relevan untuk mendeteksi risiko diabetes dan sering digunakan dalam studi klasifikasi kesehatan.

Dataset ini terdiri dari 768 entri data dengan 8 fitur utama (parameter *input*) dan 1 label *output* (target klasifikasi). Delapan fitur tersebut merupakan variabel-variabel medis yang berpengaruh terhadap risiko diabetes, yaitu:

a. *Pregnancies*: Jumlah kehamilan
b. *Glucose*: Kadar glukosa darah
c. *Blood Pressure*: Tekanan darah diastolik (mm Hg)
d. *Skin Thickness*: Ketebalan lipatan kulit trisep (mm)
e. Insulin: Kadar insulin dalam darah (mu U/ml)
f. BMI: Indeks massa tubuh
g. *Diabetes Pedigree Function*: Riwayat keturunan diabetes
h. *Age*: Usia (tahun)

Sementara itu, kolom *Outcome* digunakan sebagai label atau variabel target, yang menunjukkan hasil:

a. Nilai 0: Tidak berisiko menderita diabetes
b. Nilai 1: Berisiko menderita diabetes

Tipe data dalam dataset ini adalah numerik dan terstruktur. Data tidak memuat informasi pribadi atau sensitif, sehingga aman digunakan dalam penelitian tanpa melanggar aspek etika atau privasi. Dataset disimpan dalam format CSV (*Comma Separated Values*) dan diolah melalui *platform* IBM Watson Studio menggunakan fitur *AutoAI* untuk proses pemodelan.

3.3. Analisis Data

Analisis data dilakukan untuk memahami pola nilai parameter medis yang berkaitan dengan risiko diabetes. Tujuan dari tahap ini adalah untuk menemukan

kecenderungan karakteristik pasien yang berisiko terhadap masing-masing jenis diabetes berdasarkan delapan parameter yang ditemukan dalam kumpulan data, sehingga AI dapat membantu proses klasifikasi secara otomatis.

Setiap jenis diabetes memiliki pola nilai parameter yang berbeda dan dapat dikenali melalui data medis. Berdasarkan hasil identifikasi karakteristik umum dari masing-masing kategori, diperoleh kesimpulan analisis sebagai berikut:

Tabel 3. 1 Analisis Berdasarkan Jenis Diabetes

Jenis Diabetes	Karakteristik Umum Berdasarkan Parameter	Keterangan Tambahan
Tipe 1	Glukosa >200 mg/dL, insulin rendah, usia <25 tahun, tekanan darah rendah (60–80 mmHg), kulit tipis, tidak hamil	Umum terjadi pada remaja dan anak-anak, menyerang pria dan wanita
Tipe 2	Glukosa ≥140 mg/dL, BMI ≥25, usia ≥40 tahun, tekanan darah ≥80 mmHg, kulit tebal, riwayat keturunan positif, kehamilan 0-5 (pada perempuan)	Berkembang perlahan, dipengaruhi pola hidup dan usia dewasa
MODY	Glukosa 126–200 mg/dL, insulin normal atau rendah, BMI normal, usia <25 tahun, tekanan darah normal	Genetik, muncul sejak usia muda, bisa terjadi pada siapa saja
Gestasional	Glukosa 140–199 mg/dL, insulin tinggi atau fluktuatif, BMI ≥25, usia 20–40 tahun, terjadi saat kehamilan	Khusus pada perempuan hamil, berisiko menjadi tipe 2 pasca melahirkan
Normal	Glukosa <140 mg/dL, insulin normal, tekanan darah 70–90 mmHg, BMI 18.5–24.9, usia <35 tahun, ketebalan kulit 10–20 mm, tanpa riwayat diabetes	Menunjukkan kondisi tubuh sehat berdasarkan nilai parameter

Sumber: Data Penelitian, 2025

Setiap jenis diabetes dibedakan oleh rentang nilai parameter, misalnya:

1. Diabetes tipe 1 sering ditandai dengan glukosa yang tinggi, insulin sangat rendah, dan kulit tipis.
2. Diabetes tipe 2 umumnya muncul pada usia ≥40 tahun dengan BMI tinggi, tekanan darah tinggi, dan ketebalan kulit lebih besar dari normal.
3. MODY muncul lebih awal (usia <25 tahun), tetapi memiliki nilai tekanan darah dan BMI normal.
4. Gestasional memiliki rentang glukosa dan BMI tinggi selama masa kehamilan dan dapat dikenali dari pola insulin yang fluktuatif.

3.4. Metode Perancangan

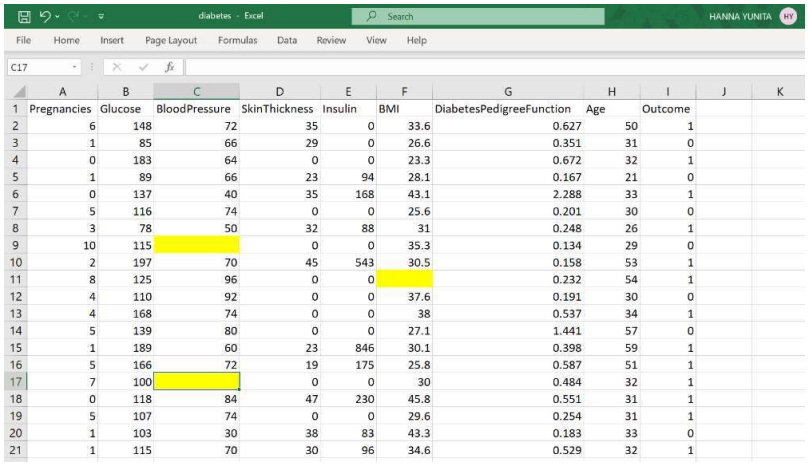
Penelitian ini merancang sistem yang dapat memprediksi risiko diabetes secara otomatis berbasis kecerdasan buatan (AI). Perancangan ini mencakup tahapan penting dari pengolahan data hingga pemodelan prediktif menggunakan algoritma *binary classification*. Agar dapat menghasilkan klasifikasi yang akurat dan efektif.

3.4.1. *Pre-Processing* Data

Tahap *pre-processing* merupakan langkah awal dalam perancangan sistem prediksi, yang bertujuan untuk memastikan bahwa data yang digunakan telah bersih, lengkap, dan siap untuk digunakan dalam proses pelatihan model.

Pada penelitian ini, proses *pre-processing* dilakukan secara manual tanpa bantuan skrip pemrograman. Langkah-langkah yang dilakukan meliputi:

1. Pemeriksaan nilai kosong: Kolom-kolom pada dataset diperiksa satu per satu. Baris yang memiliki nilai kosong pada parameter penting dihapus agar tidak mengganggu proses pelatihan.
2. Penghapusan data duplikat: Data yang teridentifikasi muncul lebih dari satu kali dihapus untuk mencegah duplikasi pengaruh terhadap hasil prediksi.
3. Pengecekan format data: Seluruh nilai dalam dataset diverifikasi agar sesuai dengan tipe numerik yang dibutuhkan.



	A	B	C	D	E	F	G	H	I
1	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
2	6	148	72	35	0	33.6	0.627	50	1
3	1	85	66	29	0	26.6	0.351	31	0
4	0	183	64	0	0	23.3	0.672	32	1
5	1	89	66	23	94	28.1	0.167	21	0
6	0	137	40	35	168	43.1	2.288	33	1
7	5	116	74	0	0	25.6	0.201	30	0
8	3	78	50	32	88	31	0.248	26	1
9	10	115		0	0	35.3	0.134	29	0
10	2	197	70	45	543	30.5	0.158	53	1
11	8	125	96	0	0		0.232	54	1
12	4	110	92	0	0	37.6	0.191	30	0
13	4	168	74	0	0	38	0.537	34	1
14	5	139	80	0	0	27.1	1.441	57	0
15	1	189	60	23	846	30.1	0.398	59	1
16	5	166	72	19	175	25.8	0.587	51	1
17	7	100		0	0	30	0.484	32	1
18	0	118	84	47	230	45.8	0.551	31	1
19	5	107	74	0	0	29.6	0.254	31	1
20	1	103	30	38	83	43.3	0.183	33	0
21	1	115	70	30	96	34.6	0.529	32	1

Gambar 3. 2 Data yang tidak valid
Sumber: Data Penelitian, 2025

Setelah proses pembersihan dilakukan, jumlah data yang semula terdiri dari 768 baris menjadi 724 baris. Hal ini menunjukkan bahwa sebanyak 44 baris data dianggap tidak layak digunakan karena mengandung nilai kosong atau tidak valid.

3.4.2. Pembagian Dataset

Setelah data melalui tahap pembersihan dan validasi, langkah selanjutnya adalah membagi dataset untuk keperluan pelatihan dan pengujian model. Tujuan dari pembagian ini adalah agar model dapat belajar dari sebagian data (data

training) dan diuji dengan data yang belum pernah dilihat sebelumnya (data *test*), sehingga hasil prediksi dapat dinilai secara objektif dan tidak bias.

	A	B	C	D	E	F	G	H	I
1	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
2	2	112	66	22	0	25	0.307	24	0
3	3	113	44	13	0	22.4	0.14	22	0
4	7	83	78	26	71	29.3	0.767	36	0
5	0	101	65	28	0	24.6	0.237	22	0
6	5	137	108	0	0	48.8	0.227	37	1
7	2	110	74	29	125	32.4	0.698	27	0
8	8	106	72	54	0	36.6	0.178	45	0
9	2	100	68	25	71	38.5	0.324	26	0
10	7	136	70	32	110	37.1	0.153	43	1
11	1	107	68	19	0	26.5	0.165	24	0
12	1	80	55	0	0	19.1	0.258	21	0
13	4	123	80	15	176	32	0.443	34	0
14	7	81	78	40	48	46.7	0.261	42	0
15	4	134	72	0	0	23.8	0.277	60	1
16	2	142	82	18	64	24.7	0.761	21	0
17	6	144	72	27	228	33.9	0.255	40	0
18	2	92	62	28	0	31.6	0.13	24	0
19	1	71	48	18	76	20.4	0.323	22	0
20	6	93	50	30	64	28.7	0.356	23	0
21	1	122	90	51	220	49.7	0.325	31	1

Gambar 3. 3 Data *Training*
Sumber: Data Penelitian, 2025

	A	B	C	D	E	F	G	H	I
1	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	
2	6	148	72	35	0	33.6	0.627	50	
3	1	85	66	29	0	26.6	0.351	31	
4	0	183	64	0	0	23.3	0.672	32	
5	1	89	66	23	94	28.1	0.167	21	
6	0	137	40	35	168	43.1	2.288	33	
7	5	116	74	0	0	25.6	0.201	30	
8	3	78	50	32	88	31	0.248	26	
9	2	197	70	45	543	30.5	0.158	53	
10	4	110	92	0	0	37.6	0.191	30	
11	4	168	74	0	0	38	0.537	34	
12	5	139	80	0	0	27.1	1.441	57	
13	1	189	60	23	846	30.1	0.398	59	
14	5	166	72	19	175	25.8	0.587	51	
15	0	118	84	47	230	45.8	0.551	31	
16	5	107	74	0	0	29.6	0.254	31	
17	1	103	30	38	83	43.3	0.183	33	
18	1	115	70	30	96	34.6	0.529	32	
19	3	126	88	41	235	39.3	0.704	27	
20	8	99	84	0	0	35.4	0.388	50	
21	7	196	90	0	0	39.8	0.451	41	

Gambar 3. 4 Data *Test*
Sumber: Data Penelitian, 2025

Pada gambar 3.3 menampilkan seluruh parameter beserta kolom *Outcome*, yang berfungsi sebagai label dalam proses pelatihan model. Kolom ini sangat penting karena digunakan oleh sistem untuk mempelajari pola hubungan antara

input (fitur) dan *output* (kelas), sehingga model dapat mengenali karakteristik data yang termasuk kategori berisiko atau tidak berisiko diabetes. Sementara itu, gambar 3.4 hanya memuat delapan parameter medis tanpa menyertakan kolom *Outcome*. Hal ini dilakukan karena data *testing* digunakan untuk mengevaluasi kemampuan model dalam melakukan prediksi secara mandiri, tanpa bantuan label hasil yang sebenarnya.

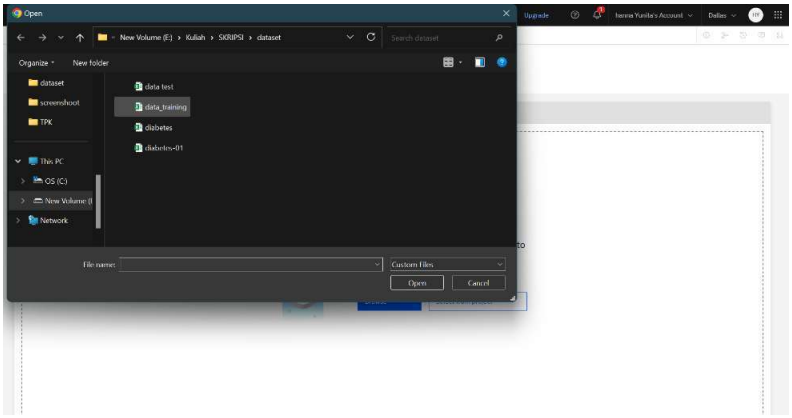
Dalam penelitian ini, dataset yang telah dibersihkan kemudian dibagi secara manual menjadi dua bagian, yaitu data *training* dan data *test*. Sebanyak 652 data digunakan sebagai data *training* yang berfungsi untuk melatih model dalam mengenali pola-pola hubungan antar parameter *input* dengan target klasifikasi. Sementara itu, 72 data lainnya digunakan sebagai data *test* untuk menguji performa model terhadap data yang belum pernah dikenali sebelumnya.

3.4.3. Pemilihan Model Otomatis

Pemilihan model dilakukan menggunakan fitur *AutoAI* yang tersedia di *platform* IBM Cloud Pak for Data. Proses ini mencakup beberapa tahapan utama, yaitu *input* data pelatihan, pemilihan fitur target dan *input*, serta seleksi algoritma *binary classification* terbaik berdasarkan performa.

a. *Input* data *training*

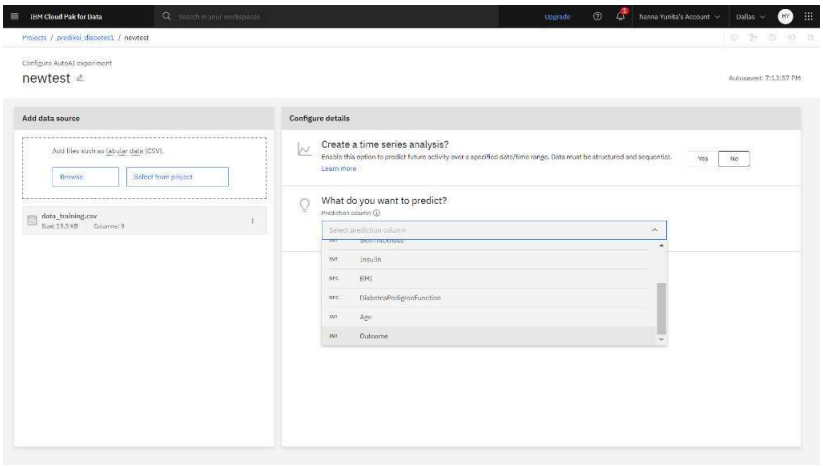
Langkah pertama adalah mengunggah data *training* ke dalam proyek *AutoAI*. Data yang telah dibersihkan sebelumnya dimasukkan dalam bentuk *file* CSV, dan sistem akan secara otomatis membaca struktur dataset tersebut.



Gambar 3. 5 *Input* Data *Training*
Sumber: Data Penelitian, 2025

b. Penentuan kolom target

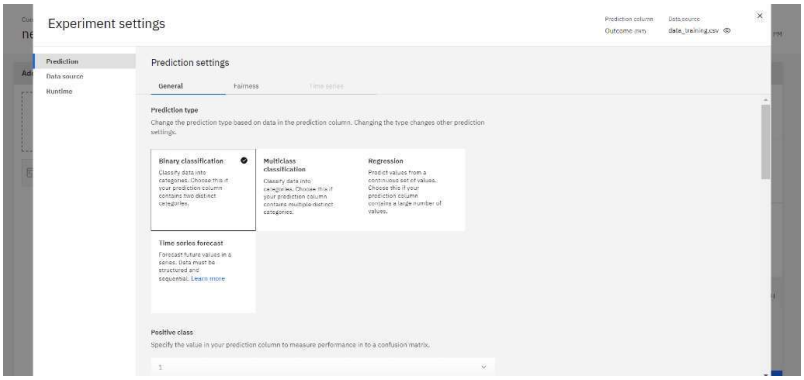
Setelah data dimuat, *AutoAI* meminta pengguna untuk menentukan kolom target, yaitu kolom yang akan diprediksi. Dalam penelitian ini, kolom *Outcome* dipilih sebagai target klasifikasi, karena berisi nilai 0 (tidak berisiko) dan 1 (berisiko) terhadap diabetes. Sementara itu, kolom-kolom lainnya seperti *Glucose*, BMI, dan *Age* ditetapkan sebagai fitur *input* yang digunakan untuk memprediksi nilai target tersebut.



Gambar 3. 6 Menentukan kolom target
Sumber: Data Penelitian, 2025

c. *Eksperiment settings*

Setelah target dan fitur *input* ditentukan, tahap selanjutnya adalah melakukan *eksperiment settings*. Dalam *AutoAI*, pengguna memiliki kebebasan untuk menyesuaikan beberapa parameter penting sebelum menjalankan proses pelatihan model.



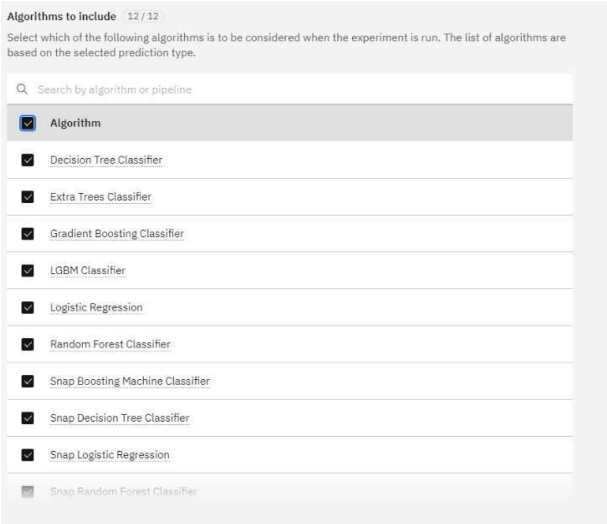
Gambar 3. 7 *Experiment Settings* (1)
Sumber: Data Penelitian, 2025

Pada gambar 3.7 memperlihatkan tampilan pengaturan eksperimen, di mana salah satu komponen utamanya adalah pemilihan *prediction type* atau tipe prediksi.

a. *Binary Classification*: Digunakan untuk memprediksi data dengan dua kategori, seperti "ya/tidak" atau "0/1". Cocok untuk masalah klasifikasi dua kelas. Digunakan dalam penelitian ini.
b. *Multiclass Classification*: Digunakan ketika target memiliki lebih dari dua kategori. Contohnya klasifikasi jenis produk atau kategori penyakit.
c. *Regression*: Untuk memprediksi nilai kontinu, seperti harga, suhu, atau jumlah penjualan.

d. *Time Series Forecast*: Digunakan untuk memprediksi nilai berdasarkan urutan waktu, seperti ramalan cuaca atau penjualan mingguan.

Dalam penelitian ini, tipe prediksi yang dipilih adalah *binary classification*, karena tujuan sistem adalah mengklasifikasikan data ke dalam dua kelas, yaitu 0 (tidak berisiko diabetes) dan 1 (berisiko diabetes).

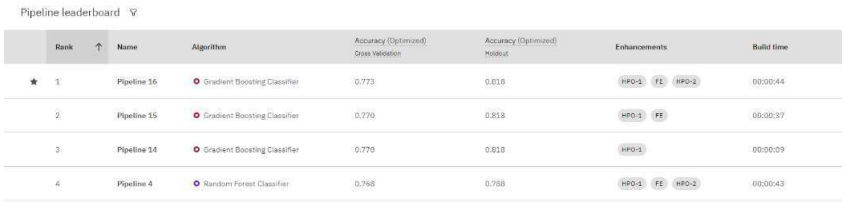


Gambar 3. 8 *Eksperiment Setting* (2)
Sumber: Data Penelitian, 2025

Selain itu, pada gambar 3.8 pengguna juga dapat menentukan algoritma yang akan disertakan dalam proses pelatihan. *AutoAI* menyediakan beberapa pilihan algoritma *binary classification* seperti *Logistic Regression*, *Random Forest*, *Support Vector Machine* (SVM), *Gradient Boosted Classifier* dan lainnya. Pemilihan algoritma ini berdasarkan rekomendasi sistem, namun pengguna tetap dapat menyesuaikannya sesuai kebutuhan. Setelah seluruh pengaturan dikonfirmasi, proses eksperimen dapat dijalankan dan *AutoAI* akan mulai membangun serta mengevaluasi *pipeline* model secara otomatis.

d. Pemilihan algoritma dan *pipeline* model

Setelah pengaturan eksperimen selesai, *AutoAI* akan menjalankan proses pelatihan model menggunakan algoritma-algoritma *binary classification* yang telah dipilih. Proses ini sepenuhnya otomatis dan dilakukan tanpa perlu campur tangan pengguna.


Pipeline leaderboard

	Rank	Name	Algorithm	Accuracy (Optimized) Cross Validation	Accuracy (Optimized) Holdout	Enhancements	Build time
★	1	Pipeline 16	Gradient Boosting Classifier	0.773	0.818	HPO-1 FE HPO-2	00:00:44
	2	Pipeline 15	Gradient Boosting Classifier	0.770	0.818	HPO-1 FE	00:00:37
	3	Pipeline 14	Gradient Boosting Classifier	0.770	0.818	HPO-1	00:00:09
	4	Pipeline 4	Random Forest Classifier	0.768	0.788	HPO-1 FE HPO-2	00:00:43

Gambar 3. 9 *Pipeline Leaderboard*
Sumber: Data Penelitian, 2025

Setelah seluruh *pipeline* model selesai dilatih dan dievaluasi, *AutoAI* akan menyajikan hasil dalam bentuk *leaderboard*. Pada gambar 3.9 *Leaderboard* ini menampilkan peringkat setiap *pipeline*. Berdasarkan *leaderboard* pada gambar di atas, *pipeline* dengan peringkat tertinggi adalah *Pipeline* 16, yang menggunakan algoritma *Gradient Boosting Classifier*. *Pipeline* ini memiliki akurasi tertinggi yaitu 0.818 (*holdout*) dan 0.773 (*cross-validation*). *Pipeline* ini diikuti oleh *Pipeline* 15 dan 14 yang menggunakan algoritma serupa namun memiliki akurasi dan konsistensi sedikit lebih rendah. Sementara itu, Pipeline 4 menggunakan algoritma *Random Forest Classifier* dengan akurasi paling rendah yaitu 0.788 (*holdout*) dan 0.768 (*cross-validation*). Berdasarkan hasil tersebut, *Pipeline* 16 dipilih sebagai model final karena memberikan performa terbaik. *Pipeline* dengan performa tertinggi secara otomatis dipilih sebagai model akhir untuk digunakan dalam klasifikasi data *testing*.

Untuk memverifikasi hasil performa model, dilakukan perhitungan manual berdasarkan *confusion matrix* yang dihasilkan diperoleh:

a. *True Positive* (TP) = 20
b. *False Negative* (FN) = 2
c. *False Positive* (FP) = 10
d. *True Negative* (TN) = 34
e. Total data = 66

Berdasarkan Rumus 2. 1 ini, nilai akurasi manual yang diperoleh adalah:

$$Accuracy = \frac{TP + TN}{Total} = \frac{20 + 34}{66} = 0.818$$

Nilai akurasi manual ini identik dengan nilai *holdout score* yang dilaporkan oleh *AutoAI*, yaitu 0.818, yang mengindikasikan konsistensi antara evaluasi manual dan sistem. Sementara itu, nilai *cross-validation score* yang diperoleh sebesar 0.773 menunjukkan stabilitas model yang baik ketika diuji pada beberapa subset data berbeda. Perbedaan skor ini wajar terjadi dan mencerminkan kemampuan model dalam menghadapi variasi data di luar data pelatihan.

Gradient Boosting membentuk model secara bertahap dengan memperbaiki kesalahan dari iterasi sebelumnya. Salah satu komponen utama yang dihitung adalah residual, yaitu selisih antara nilai aktual dan prediksi sebelumnya. Misalnya, untuk suatu data, jika nilai aktual y adalah 1, dan prediksi sebelumnya $\hat{y}$ adalah 0,72, maka berdasarkan **Rumus 2. 6** residual-nya dihitung sebagai:

$$Residual = y - \hat{y} = 1 - 0.72 = 0.28$$

Residual ini kemudian dipelajari oleh model lemah, dan hasilnya digunakan untuk memperbarui prediksi. Misalnya, dengan learning rate ν=0,1 dan hasil prediksi residual adalah 0,28, maka nilai prediksi baru akan dihitung berdasarkan **Rumus 2. 8** sebagai:

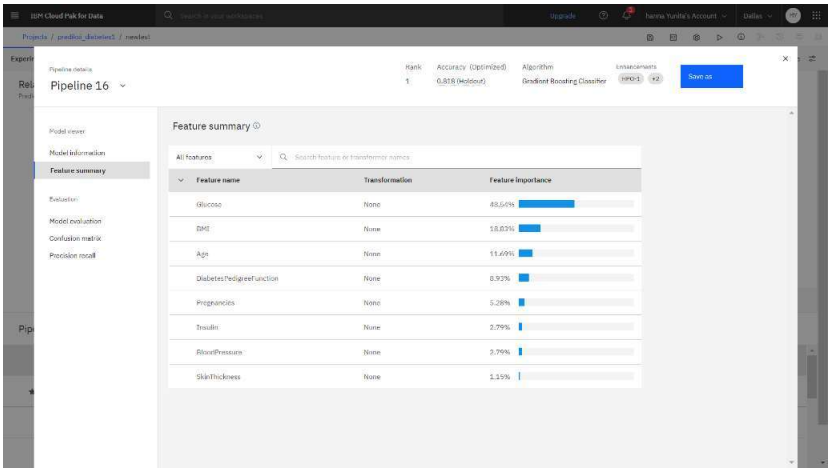
$$\hat{y}_{baru} = \hat{y}_{lama} + v \times Predicted\ Residual = 0.72 + 0.1 \times 0.28 = 0.748$$

Dengan pembaruan ini, model menjadi sedikit lebih dekat dengan nilai aktual. Proses ini diulang terus menerus hingga prediksi cukup akurat. Pendekatan inilah yang digunakan pada *Pipeline* 16 yang menghasilkan akurasi tinggi pada data penelitian ini.

3.4.4. Rekayasa Fitur Otomatis

Tahap rekayasa fitur (*feature engineering*), tahap ini bertujuan untuk meningkatkan kualitas representasi data dengan mengubah fitur-fitur mentah menjadi bentuk yang lebih informatif dan sesuai dengan kebutuhan model klasifikasi yang dipilih.

Dalam penelitian ini, hasil dari proses rekayasa fitur divisualisasikan dalam bentuk *Feature Summary*, yang menampilkan kontribusi relatif setiap fitur terhadap hasil prediksi model. Meskipun pada gambar 3.10 tidak ditemukan fitur baru atau transformasi tambahan (ditunjukkan oleh kolom *Transformation* yang bernilai *None*), sistem tetap menjalankan proses seleksi otomatis dan menyusun peringkat fitur berdasarkan pengaruhnya terhadap klasifikasi.

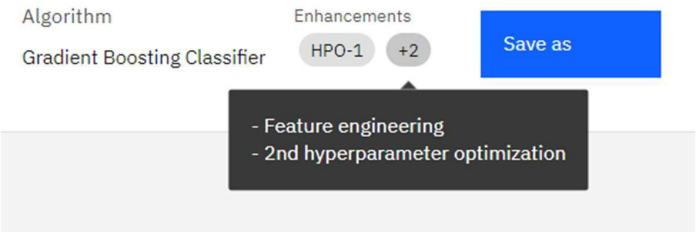


Gambar 3. 10 *Feature Summary*
Sumber: Data Penelitian, 2025

Dari Gambar 3.10 terlihat bahwa fitur *Glucose* memberikan kontribusi tertinggi sebesar 48.54%, diikuti oleh BMI (18.83%), *Age* (11.69%), dan fitur-fitur lainnya. Gambar ini menunjukkan bahwa meskipun tidak dilakukan rekayasa fitur tambahan, sistem tetap melakukan evaluasi pentingnya setiap fitur terhadap model akhir. Algoritma yang digunakan dalam *pipeline* terbaik adalah *Gradient Boosting Classifier*, yang sangat sensitif terhadap kualitas dan struktur data *input*, sehingga pemilihan fitur yang optimal menjadi langkah penting dalam membentuk performa model.

3.4.5. Optimasi *Hyperparameter*

Pada tahap optimasi *hyperparameter*, yaitu penyetelan parameter-parameter utama dari algoritma pembelajaran mesin untuk memperoleh performa prediksi terbaik. Berbeda dengan parameter model yang dipelajari langsung dari data, *hyperparameter* adalah nilai-nilai eksternal yang perlu ditentukan sebelum proses pelatihan dimulai.



Gambar 3. 11 *Enhancements*
Sumber: Data Penelitian, 2025

Dalam penelitian ini, *AutoAI* memilih *Gradient Boosting Classifier* sebagai algoritma terbaik berdasarkan hasil evaluasi model. Proses pemilihan dan peningkatan kualitas model dilakukan melalui tiga tahap *enhancements*, yaitu:

1. HPO-1 (*First Hyperparameter Optimization*): Proses awal pencarian konfigurasi parameter terbaik untuk setiap algoritma yang diuji.
2. *Feature Engineering*: Transformasi dan penyempurnaan fitur dilakukan secara otomatis untuk meningkatkan representasi data.
3. HPO-2 (*Second Hyperparameter Optimization*): Penyesuaian ulang *hyperparameter* untuk menyesuaikan dengan struktur data hasil rekayasa fitur.

Proses *hyperparameter optimization* dalam *AutoAI* dirancang untuk menyempurnakan model dengan performa terbaik. Sistem ini menggunakan algoritma optimasi khusus yang mampu menangani fungsi evaluasi kompleks, seperti pelatihan dan pengujian model, yang umum terjadi dalam pembelajaran mesin. Meskipun setiap iterasi evaluasi memerlukan waktu yang relatif lama, pendekatan ini tetap dapat secara efisien mengidentifikasi kombinasi *hyperparameter* paling optimal.

3.5. Lokasi dan Jadwal Penelitian

Penelitian ini dilakukan secara mandiri di rumah dengan memanfaatkan *platform* IBM Cloud Pak for Data. Seluruh proses penelitian seperti pemilihan model, pelatihan model, hingga evaluasi dilakukan secara *online*, sehingga tidak memerlukan lokasi fisik khusus. Penggunaan *AutoAI* memungkinkan peneliti untuk melakukan eksperimen secara fleksibel, efisien, dan terukur tanpa harus menulis kode secara manual.

Jadwal penelitian ini meliputi beberapa tahap penyusunan. Rincian jadwal pelaksanaan penelitian disajikan dalam bentuk tabel sebagai berikut:

Tabel 3. 2 Jadwal Penelitian

No	Kegiatan	Bulan																			
		Maret				April				Mei				Juni				Juli			
		1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
1	Pengajuan Judul																				
2	Penyusunan Bab I																				
3	Penyusunan Bab II																				
4	Penyusunan Bab III																				
5	Penyusunan Bab IV																				
6	Penyusunan Bab V																				
7	Revisi																				
8	Pengumpulan Skripsi																				

Sumber: Data Penelitian, 2025