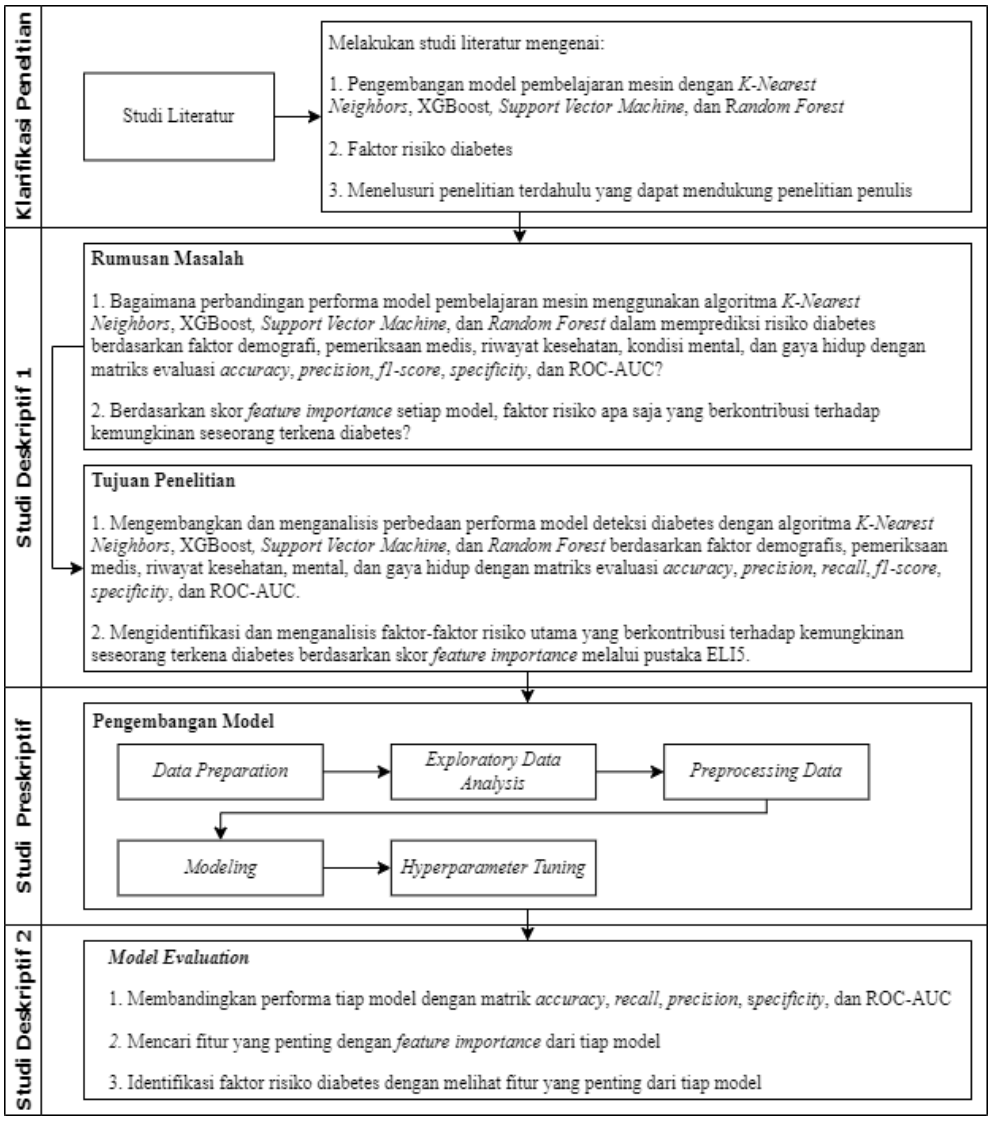
BAB III
METODE PENELITIAN

3.1 Desain Penelitian

Desain penelitian adalah konsep kerangka kerja untuk menentukan metode dan teknik penelitian yang bertujuan menjadi fondasi dan acuan dalam melakukan penelitian. Desain penelitian yang digunakan untuk penelitian ini adalah metode *Design Research Methodology* (DRM), terbagi ke dalam 4 tahap yaitu klarifikasi penelitian, studi deskriptif 1, studi preskriptif, dan studi deskriptif 2 (Blessing & Chakrabarti, 2009). Tahapan tersebut diilustrasikan pada Gambar 3.1 berikut.



Gambar 3.1 Desain Penelitian

3.1.1 Klarifikasi Penelitian

Tahap awal dari kerangka kerja DRM adalah klarifikasi penelitian yaitu melakukan studi literatur dengan melakukan kajian terhadap literatur yang membahas mengenai pengembangan model pembelajaran mesin dengan algoritma *K-Nearest Neighbors*, XGBoost, *Support Vector Machine*, dan *Random Forest*. Peneliti juga mengkaji sumber literatur mengenai fenomena diabetes yang mencakup faktor risiko yang dapat meningkatkan potensi seseorang terkena diabetes. Literatur yang dipakai sebagian besar dari berbagai jurnal dan skripsi yang melakukan penelitian dengan topik dan studi kasus yang berkaitan erat dengan topik diabetes, faktor risiko diabetes, pembelajaran mesin, *K-Nearest Neighbors*, XGBoost, *Support Vector Machine*, dan *Random Forest* yang diharapkan dapat membantu peneliti untuk melakukan penelitian.

3.1.2 Studi Deskriptif 1

Tahap kedua ketika peneliti memfokuskan penelitian berdasarkan permasalahan dari topik yang dibahas sehingga dapat menjadi acuan dalam menjalankan penelitian sehingga dapat memecahkan permasalahan tersebut tanpa keluar dari topik. Selanjutnya peneliti menentukan tujuan penelitian yang berasal dari permasalahan yang telah ditentukan sebelumnya. Tujuan penelitian merupakan hasil yang ingin diraih melalui penelitian yang dilakukan. Permasalahan dan tujuan penelitian merupakan acuan dan panduan peneliti dalam bagaimana peneliti membangun model pembelajaran mesin hingga dapat menghasilkan model dan analisis yang dapat menjawab permasalahan penelitian dan memenuhi tujuan penelitian.

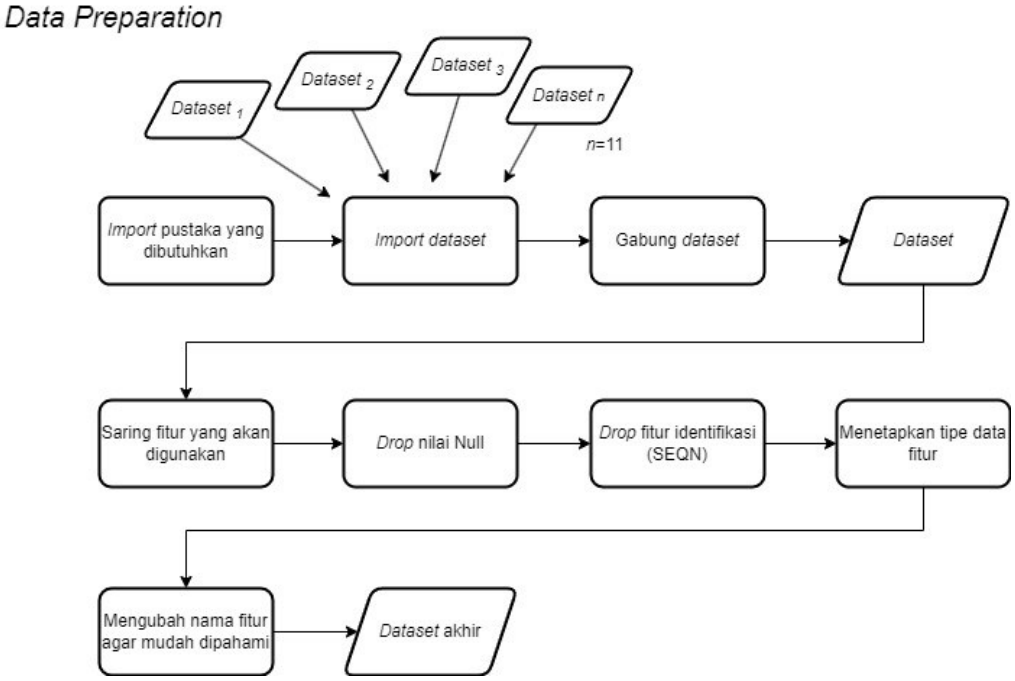
3.1.3 Studi Preskriptif

Tahap ketiga berisi proses pengembangan model pembelajaran mesin pada kasus deteksi diabetes. Tahapan pengembangan model dibagi menjadi *data preparation*, *exploratory data analysis*, *preprocessing data*, *modeling*, *hyperparameter tuning*, dan *model evaluation*.

1. *Data Preparation*

Data preparation merupakan tahapan identifikasi, ekstraksi, pembersihan, dan integrasi data agar siap digunakan untuk analisis data lebih lanjut. Tidak ada

urutan langkah yang disepakati secara universal pada proses *data preparation* (Fernandes dkk., 2023). Pada proses ini, peneliti menangani kumpulan *dataset* dalam beberapa tahap yaitu *import* pustaka yang dibutuhkan, *import dataset*, gabung *dataset*, saring fitur yang akan digunakan, *drop* nilai *null*, *drop* fitur identifikasi (SEQN), menetapkan tipe data fitur, dan mengubah nama fitur. Agar lebih jelas, perhatikan diagram alir pada Gambar 3.2.



Gambar 3.2 Diagram Alir *Data Preparation*

- *Import* Pustaka yang Dibutuhkan

Import pustaka yang dibutuhkan selama proses pengolahan data dan pengembangan. Pustaka-pustaka yang digunakan berdasarkan kebutuhan apa yang ingin dipenuhi untuk menangani karakteristik data yang dihadapi, di antaranya yaitu Pandas, Numpy, Matplotlib, Seaborn, Functools, Scikit-learn, Imblearn, Collections, XGBoost, dan ELI5.

- *Import Dataset*

Import 11 *dataset* yang telah diunduh pada situs NHANES yaitu *Demographic Variables and Sample Weights*, *Blood Pressure - Oscillometric Measurements*, *Body Measures*, *Glycohemoglobin*, *Cholesterol – Total*, *Alcohol Use*, *Blood Pressure & Cholesterol*, *Diabetes*, *Mental Health - Depression*

Screener, *Medical Conditions*, dan *Smoking - Cigarette Use*. Seluruh *dataset* tersebut mengandung fitur yang dibutuhkan dalam penelitian. Peneliti menggunakan *platform* Jupyter Lab agar dapat menggunakan dan mengolah *dataset* secara *offlline* dan tidak perlu mengunggah berkali-kali saat ingin digunakan.

- Gabung *Dataset*

Gabung 11 *dataset* yang dibantu oleh fungsi *reduce* dari pustaka Functools yang dikombinasikan dengan fungsi *lambda* untuk *merge* perulangan agar lebih efisien. Penggabungan *dataset* menggunakan fitur SEQN sebagai fitur identifikasi yang merepresentasikan urutan peserta survei.

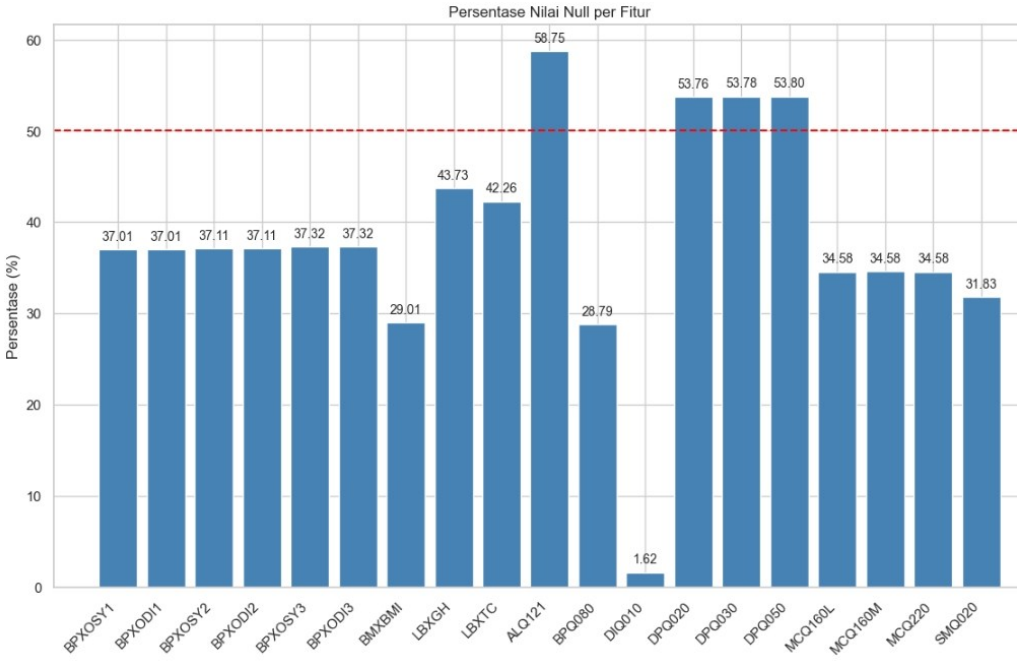
- Saring Fitur yang Akan Digunakan

Selanjutnya saring fitur mana saja yang akan digunakan dalam proses olah data dan pengembangan model pembelajaran mesin.

- *Drop* Nilai *Null*

Pada proses pengembangan, ditemukan banyak sekali nilai *null* pada setiap fitur. Sulit sekali untuk dapat memutuskan bagaimana nilai *null* ditangani pada *dataset* karena dihadapkan pada pilihan untuk *casewise deletion* atau imputasi nilai *null*. Sampai pada keputusan untuk *casewise deletion* setelah mempertimbangkan kelebihan dan kekurangannya. *Casewise deletion* merupakan salah satu teknik penanganan data yang hilang dengan menghapus atau membuang data pada observasi, dan hanya observasi yang lengkap yang tetap dipertahankan sebagai objek analisis. Kelebihan teknik ini adalah penerapannya yang mudah. Namun, *casewise deletion* atau *drop* nilai *null* pada *dataset* berisiko menghilangkan informasi yang ada dan dapat mengganggu representasi data apalagi pada penelitian ini, terdapat 4 fitur yang memiliki nilai *null* lebih dari 50% dari jumlah keseluruhan data seperti yang ditunjukkan pada Gambar 3.3. Namun, peneliti memperkirakan dengan data sebanyak 11933 akan memakan biaya komputasi yang besar untuk mengisinya. Selain itu, nilai *null* yang memiliki porsi besar tersebut tidak dapat ditangani hanya dengan mengisinya dengan nilai mean, median, atau modus saja. Teknik imputasi yang lebih lanjut dibutuhkan yang mana akan memakan biaya komputasi besar juga. Hal ini belum termasuk proses *resampling* dan *cross-*

validation yang membutuhkan biaya komputasi besar apalagi mengingat jumlah data yang ditangani. Jika mempertahankan jumlah data sebanyak sebelumnya, pilihan teknik *resample* akan terbatas pada teknik *undersampling* saja karena teknik *oversampling* akan memakan biaya komputasi yang besar juga, hal ini tentu mempersulit menemukan teknik *resample* yang cocok terhadap karakteristik data juga pengaruhnya terhadap hasil performa model. Dengan berbagai pertimbangan tersebut, peneliti memutuskan untuk mengorbankan data beserta informasi yang ada di dalamnya dengan *casewise deletion*, namun dengan begitu peneliti mampu melanjutkan proses pengembangan model tanpa mengkhawatirkan biaya komputasi yang mahal dan keterbatasan pilihan teknik *resample* yang dapat menghambat proses pengembangan model pembelajaran mesin.



Gambar 3.3 Visualisasi Persentase Nilai *Null* per Fitur

- *Drop* Fitur Identifikasi (SEQN)

Setelah menggabungkan 11 *dataset* dan menyaring mana fitur mana saja yang dibutuhkan, fitur SEQN sebagai fitur identifikasi yang mencerminkan nomor urutan responden menjalani tahap survei sudah tidak digunakan lagi dan harus di-*drop* agar pengolahan data menjadi lebih efisien.

- Menetapkan Tipe Data Fitur

Menetapkan tipe data fitur sebelum memasuki proses EDA dan *preprocessing data* agar memudahkan proses visualisasi dan pengambilan informasi. Penetapan tipe data ini berdasarkan ringkasan statistik yang ada pada *Doc File* dari *dataset* fitur tersebut berasal. Selain itu, model akan lebih mudah mengenali karakteristik data jika sudah ditentukan terlebih dahulu.

- Mengubah Nama Fitur

Ubah nama fitur agar memudahkan pembacaan, pengenalan, dan pemahaman konteks saat melakukan proses pengolahan data.

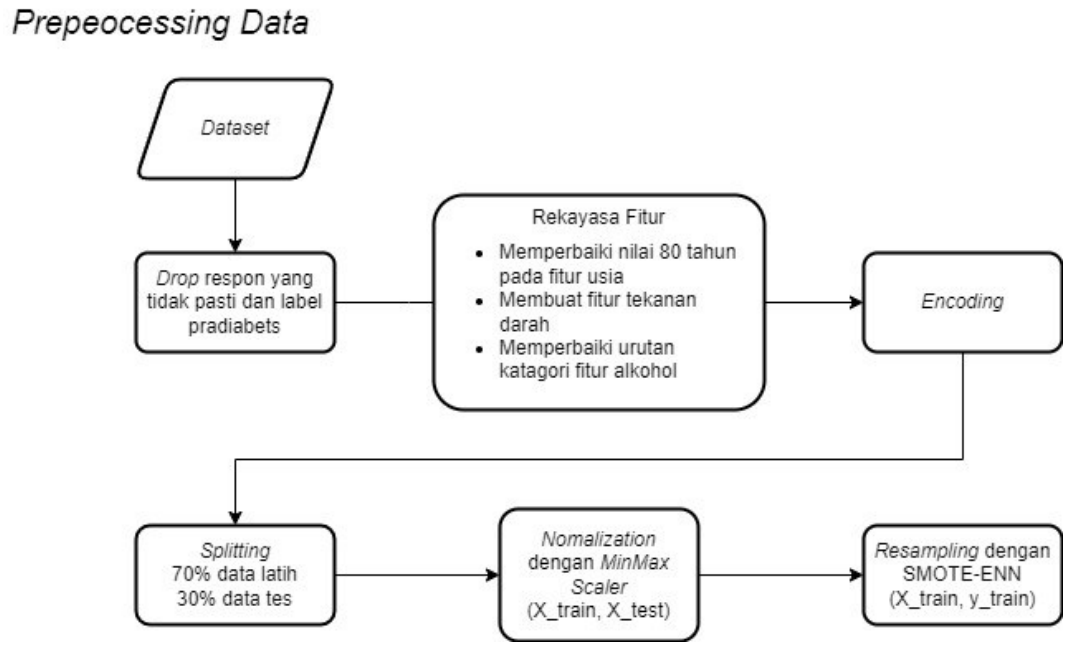
2. *Exploratory Data Analysis*

Exploratory Data Analysis (EDA) merupakan tahap penting pada penelitian apa pun yang melibatkan analisis data. Tujuan utamanya untuk memeriksa persebaran data, *outlier*, dan anomali data yang dapat mengarahkan peneliti agar uji hipotesis dilakukan berdasarkan informasi nyata yang ditemukan. Uji hipotesis ini dilakukan melalui visualisasi grafis agar lebih mudah memahami dan menganalisis informasi dalam data (Komorowski dkk., 2016). Peneliti melakukan EDA untuk menganalisis informasi *dataset* yang nantinya akan menentukan langkah lanjutan dalam menangani *dataset* pada tahap *preprocessing data*. Peneliti memeriksa karakteristik *dataset* dengan melakukan visualisasi statistik deskriptif dengan *bar plot*, *hist plot*, dan *box plot* untuk memeriksa persebaran data termasuk *outlier* pada fitur numerik. Juga visualisasi *count plot* pada fitur kategori. Semua visualisasi dilakukan menggunakan pustaka Matplotlib dan Seaborn.

3. *Preprocessing Data*

Preprocessing data merupakan tahap yang paling berpengaruh untuk meningkatkan kualitas data agar performa model menjadi lebih baik. Pada tahap ini, seseorang dapat menghabiskan waktu 50% hingga 80% dari seluruh proses klasifikasi. Hal tersebut membuktikan seberapa penting *preprocessing data* dalam pengembangan model pembelajaran mesin (Maharana dkk., 2022). Pada penelitian ini, peneliti menyusun serangkaian tahap *preprocessing* berdasarkan penemuan informasi pada tahap EDA. Peneliti melakukan beberapa tahapan untuk meningkatkan kualitas data agar *modeling* dapat menghasilkan model dengan

performa yang baik. Tahapan tersebut antara lain yaitu *drop* respons yang tidak pasti, rekayasa fitur, *encoding*, *splitting*, *normalization*, dan *resampling* dengan SMOTE-ENN. Tahapan ini diilustrasikan melalui diagram alir pada Gambar 3.4.



Gambar 3.4 Diagram Alir *Preprocessing Data*

- *Drop* Respons yang Tidak Pasti dan Label Pradiabetes

Drop respons yang tidak informatif berdasarkan *Doc File* berupa jawaban 'tidak tahu' dan 'menolak menjawab' yang berasal dari jawaban partisipan survei. Selain itu, drop data dengan label pradiabetes agar pengembangan model terfokus pada klasifikasi biner diagnosis negatif dan positif diabetes.

- Rekayasa Fitur

Terdapat beberapa proses pada tahap ini. Pertama-tama, memperbaiki tipe data campuran pada fitur usia di mana umur 0 hingga 79 berupa numerik sedangkan 80 berupa kategori yang mewakili partisipan dengan umur 80 tahun ke atas. Berdasarkan *Doc_File*, pengelompokan nilai 80 tahun ke atas disebabkan karena kekhawatiran privasi. Oleh karena itu, peneliti mengganti nilai 80 dengan nilai 85 yang merupakan rerata tertimbang kelompok usia 80 tahun ke atas berdasarkan *Doc File* fitur terkait.

Selanjutnya, peneliti akan membuat fitur tekanan darah yang dibentuk dari fitur tiga pembacaan sistolik dan diastolik yaitu *Systolic - 1st oscillometric reading*,

Diastolic - 1st oscillometric reading, *Systolic - 2nd oscillometric reading*, *Diastolic - 2nd oscillometric reading*, *Systolic - 3rd oscillometric reading*, *Diastolic - 3rd oscillometric reading* dengan formula *Mean Arterial Pressure* (MAP) sebagai berikut.

$$\text{Diastolik} + \frac{1}{3}(\text{Sistolik - Diastolik}) \quad (1)$$

MAP merupakan salah satu pengukuran penting yang dapat menganalisis tekanan sistolik dan diastolik sekaligus, dua hal ini berperan penting dalam deteksi gangguan tekanan darah dan beberapa faktor risikonya. Dalam beberapa penelitian ditemukan bahwa dapat menjadi masalah jika seseorang memiliki tekanan sistolik yang normal namun memiliki tekanan diastolik yang tinggi, begitu juga sebaliknya. Dalam hal ini, MAP bisa sangat berguna untuk dapat mengumpulkan informasi dari tekanan sistolik dan diastolik sekaligus tanpa terpaku pada salah satunya. Selain itu, pengukuran MAP sangat berguna dalam data epidemiologi besar, untuk statistik inferensial yang lebih baik, dan interpretasi yang menghasilkan banyak informasi (Kundu dkk., 2017).

Proses terakhir pada tahap rekayasa fitur adalah memperbaiki urutan kategori fitur alkohol yang merupakan informasi seberapa sering partisipan konsumsi minuman beralkohol. Fitur ini merupakan fitur kategori ordinal karena adanya hubungan ordinal pada setiap nilainya, sehingga pengurutan kategori dibutuhkan agar dapat mempertahankan informasi yang berguna pada proses *modeling* (Poslavskaya & Korolev, 2023).

- *Encoding*

Sebagian besar algoritma pembelajaran mesin memerlukan *input* berupa numerik, sehingga data kategori perlu diubah terlebih dahulu ke dalam bentuk numerik. Proses perubahan tersebut disebut dengan *encoding,* yaitu perubahan bentuk data kategori menjadi bentuk numerik atau vektor numerik agar dapat melalui tahap pencocokan dengan algoritma (*modeling*) (Poslavskaya & Korolev, 2023). *Encoding* juga penting agar model dapat memahami dan mengambil informasi yang berguna dari data (Dahouda & Joe, 2021).

Terdapat tiga tipe fitur kategori dalam *dataset* yaitu kategori nominal pada fitur ras. Kategori ordinal pada fitur depresi, gangguan makan, gangguan tidur, dan alkohol. Sedangkan kategori biner pada fitur gender, riwayat tiroid, riwayat kanker, riwayat kolesterol tinggi, dan diabetes sebagai data target. Semua fitur kategori ordinal dalam *dataset* sudah dalam bentuk angka dan dalam urutan yang benar sehingga tidak perlu lagi dilakukan *encoding*. Selanjutnya, peneliti melakukan *one-hot encoding* khusus untuk fitur ras karena tiap nilai di dalamnya tidak berhubungan sama sekali. *One-hot encoding* digunakan saat fitur merupakan nominal yang tidak mempunyai urutan (Dahouda & Joe, 2021). Sedangkan untuk fitur kategori biner peneliti mengubah label 2 menjadi 0 untuk mewakili diagnosa negatif, sedangkan label 1 sudah benar melambangkan diagnosa positif.

- *Splitting*

Dalam *splitting*, peningkatan proporsi data latih dapat meningkatkan performa pelatihan dan membuat model lebih stabil. Peningkatan data latih dari 30% hingga 80% dapat meningkatkan performa saat *testing*, namun akan menurun saat ukuran data latih berada di proporsi 80% hingga 90% (Nguyen dkk., 2021). Oleh karena itu, peneliti memutuskan untuk melakukan *split data* dengan proporsi 70% data latih dan 30% data tes agar performa saat pelatihan optimal dan konsisten.

- *Normalization*

Normalization bertujuan agar *dataset* memiliki nilai dalam rentang yang mirip, hal ini penting terutama pada data yang kurang terstruktur yang berisi rentang nilai yang berbeda. Peneliti memutuskan untuk menggunakan *min-max scaler* pada pustaka Scikit-learn. *Min-max* adalah teknik *normalization* yang mengubah rentang data menjadi nilai antara 0 dan 1. Teknik ini unggul pada *dataset* yang berdimensi tinggi (Deepa & Ramesh, 2022). Berikut formula *min-max normalization*.

$$X_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (2)$$

- *Resampling*

Untuk mengatasi ketidakseimbangan data target, peneliti menggunakan teknik SMOTE-ENN. SMOTE-ENN merupakan teknik kombinasi *oversampling*

dengan SMOTE dan *undersampling* dengan ENN. SMOTE atau *Synthetic Minority Over-sampling Technique* menambah data buatan berdasarkan kemiripan data minoritas berdasarkan jarak tetangga terdekatnya sedangkan ENN atau *Edited Nearest Neighbors* mengurangi data kelas mayoritas yang dianggap sebagai *noise* pada penghitungan jarak observasi kelas mayoritas (Xu dkk., 2020). Teknik ini menggunakan fungsi SMOTEENN dari pustaka Imblearn. Peneliti memutuskan menggunakan teknik ini karena SMOTE-ENN menawarkan dua langkah solusi yaitu *oversampling* pada data minoritas (SMOTE), namun semua data termasuk data hasil *oversampling* ini diperiksa lagi dan akan dieliminasi jika data termasuk *noise* (ENN). Sehingga SMOTE tidak menimbulkan data sintetis yang dapat menimbulkan *overfitting* karena *noise* yang ada akan dicegah dan dieliminasi oleh ENN.

4. *Modeling*

Modeling merupakan tahap pelatihan data melalui pengenalan pola-pola data latih dengan perhitungan statistik algoritma tertentu agar dapat menghasilkan model pembelajaran mesin yang mampu melakukan prediksi akurat pada data baru. Peneliti pertama-tama melakukan *modeling* dengan algoritma *K-Nearest Neighbors*, XGBoost, *Support Vector Machine*, dan *Random Forest*. Keempat algoritma tersebut menggunakan parameter *default*. Untuk melihat performa keempat model tersebut, peneliti menggunakan *accuracy*, *precision*, *recall*, *f1-score*, *specificity* dan ROC-AUC sebagai matriks penilaian performa.

5. *Hyperparameter Tuning*

Penelitian ini menggunakan teknik *grid search* dan *cross-validation* yaitu melalui fungsi GridSearchCV yang dapat diakses pada pustaka *scikit-learn* untuk mengimplementasikan *hyperparameter tuning* yang optimal. *Cross-validation* ditentukan sebanyak 10 *fold* melalui fungsi StratifiedKFold.

Berikut Tabel 3.1 yang berisi ruang parameter yang nantinya akan menjadi *input* fungsi GridSearchCV untuk melakukan *hyperparameter tuning* pada setiap model.

Tabel 3.1 Parameter dan Ruang Parameter yang Digunakan pada Proses *Hyperparameter Tuning*

Model	**Parameter**	**Nilai/Rentang Nilai**
KNN	*n_neighbors*	3, 5, 7, 9, ... 19
	weights	*uniform*, *distance*
	metric	*euclidean*, *manhattan*, *minkowski*
	p	1, 2
XGBoost	*n_estimators*	400, 500
	learning_rate	0.05, 0.1, 0.2
	max_depth	7, 8, 9
	min_child_weight	1, 3, 5
	gamma	0, 0.1, 0.2
	subsample	0.7, 0.8
	colsample_bytree	0.7, 0.8
SVM	*kernel*	*linear*, *rbf*, *poly*
	C	0.01, 0.1, 1
	gamma	scale, 0.01, 0.1 (khusus *poly* dan *rbf*)
	coef0	0, 0.1, 0.2, 0.3, ... 0.9 (khusus *poly*)
	degree	2, 3 (khusus *poly*)
Random Forest	*n_estimators*	100, 200, 300, 400
	max_depth	*None*, 5, 10, 15
	min_samples_split	2, 5, 10
	min_samples_leaf	1, 2, 4
	max_features	sqrt, log2, 0.5

Berikut penjelasan parameter dan *input* nilai/rentang nilai setiap model pada Tabel 3.1.

- *K-Nearest Neighbors*

Pemilihan rentang nilai *n_neighbors* dimulai dari 3 untuk menghindari *overfitting* termasuk *noise*. Nilai yang lebih tinggi membuat prediksi tidak agresif dan menghindari *overfitting*, namun tetap dalam rentang yang tidak terlalu tinggi (kurang dari 20) untuk menghindari *underfitting* (Bruce dkk., 2020, hlm. 246). Selanjutnya, peneliti memasukkan setiap parameter dengan ruang parameter yang tersedia dengan pada KNN yaitu *weights* (*distance* dan *uniform*) dan *metric* (*euclidean*, *manhattan*, dan *minkowski*) untuk memaksimalkan pencarian *hyperparameter*. Sisanya parameter p mewakili *minkowski.*

- XGBoost

Dari 7 parameter yang digunakan, 6 di antaranya digunakan untuk mencegah agar prediksi tidak *overfitting*. Hanya *n_estimators* yang berfokus untuk menggali kedalaman informasi dan memperkecil *error* untuk setiap iterasi *boosting*. Hal ini sangat wajar karena *hyperparameter* pada XGBoost memang sering dipakai untuk menyeimbangkan fenomena *overfitting* dengan *accuracy* dan kerumitan komputasi yang dapat memakan waktu. Oleh karena itu, setiap parameter diberikan ruang parameter yang dibatasi agar biaya komputasi tidak melonjak (Bruce dkk., 2020, hlm. 272).

Jumlah ruang parameter pada *n_estimators* yang sedikit namun dengan nilai yang besar (400 dan 500) merupakan *trade-off* antara biaya komputasi dengan performa. 6 parameter lainnya berfokus pada pencegahan *overfitting*, hal ini dilakukan agar prediksi tidak cenderung menghasilkan diagnosa negatif yang merupakan label mayoritas pada data target. Seperti pada *learning_rate* yang mengurangi bobot penggalian data tiap iterasi dengan ruang parameter 0.05, 0.1, dan 0.2 agar menghindari *overfitting* sekaligus mengatasi *noise* yang dapat diatasi dengan nilai yang lebih kecil dari nilai *default* (0.3) (Bruce dkk., 2020, hlm. 272). Begitu juga dengan *max_depth* dengan rentang nilai di bawah 10 (7, 8, dan 9) yang mengatur kedalaman pohon tiap iterasi agar menghindari model menjadi terlalu kompleks yang dapat memicu *overfitting* (Bruce dkk., 2020, hlm. 279).

Pada *min_child_weight*, *overfitting* dicegah dengan pilihan menerapkan nilai *default* dan lebih dari *default* yaitu 1, 3, dan 5. Nilai tersebut merupakan bobot minimum yang mencegah percabangan, semakin besar nilainya makan model semakin konservatif (XGBoost Developers, 2022). Parameter *gamma* juga memiliki peran agar model sedikit lebih konservatif dengan *treshold* rendah (0, 0.1, 0.2) terhadap penurunan *error* yang harus dicapai agar suatu percabangan dilakukan.

Selanjutnya ruang parameter *subsample* ditentukan dengan nilai mendekati 1 (0.7, 0.8, dan 0.9) agar proporsi *sample* pada *fold* yang diambil sedikit berbeda tiap iterasinya, sama halnya dengan *colsample_bytree* yang memiliki ruang parameter mendekati 1 (0.7, 0.8, 0.9) agar hanya 70-90% fitur yang dipakai tiap iterasinya. Dua parameter terakhir ini sama-sama digunakan untuk mencegah *overfitting* (Bruce dkk., 2020, hlm. 281)

- *Support Vector Machine*

Peneliti memakai semua *kernel* yaitu *linear*, *rbf*, dan *poly* kecuali *sigmoid* agar menghemat biaya komputasi. Ruang parameter C yaitu 1, 0.1, dan 0.01 digunakan untuk mencegah *overfitting*. Semakin kecil C, semakin banyak data yang boleh melewati batas *hyperplane* yang membuat prediksi lebih *general* (Suyanto, 2018, hlm. 114). Sama halnya dengan C, menghindari *overfitting* dapat dilakukan dengan mengurangi nilai *gamma*, maka pada penelitian ini nilai *gamma* yang kurang dari 1 digunakan (0.01, 0.1, *scale*) (Géron, 2019, hlm. 162).

Selanjutnya parameter *coef0* yang mana referensi penggunaan praktis mengenai rentang nilai parameter ini sangat sedikit, namun terdapat temuan pada penelitian yang dilakukan oleh Wu dkk. (2024) bahwa performa model menurun ketika nilai *coef0* lebih kecil dari -1 dan lebih besar dari 1. Oleh karena itu, peneliti memutuskan menggunakan rentang nilai 0, 0.1, 0.2, 0.3, dan seterusnya sampai 0.9.

Terakhir adalah parameter *degree* yang mewakili pangkat pada kernel *poly.* Selain C dan *gamma*, *degree* juga digunakan untuk mencegah *overfitting* dengan menurunkan nilainya. Peneliti menggunakan dua nilai (2 dan 3) agar prediksi tidak *overfitting* namun tidak *underfitting* juga (Géron, 2019, hlm. 160).

- *Random Forest*

Sama dengan XGBoost, ruang parameter *n_estimators* dan *max_depth* pada *Random Forest* yang sedikit namun dengan nilai yang besar digunakan untuk menghasilkan performa yang maksimal namun tetap dalam biaya komputasi yang tidak besar dan membatasi kedalaman model setiap iterasi agar tidak *overfitting*. Namun *max_depth* pada *Random Forest* tidak ketat dibandingkan dengan XGBoost dan memiliki opsi *max_depth* tak terbatas untuk pengurangan *error* yang maksimal, yang mana dapat menimbulkan masalah biaya komputasi. Hal itu dapat dihindari karena parameter yang di *tuning* beserta ruang parameternya tidak sebanyak yang di *tuning* pada XGBoost.

Masalah *overfitting* pun dapat muncul jika tidak adanya batas iterasi. Oleh karena itu, peneliti menggunakan *min_samples_split* (2, 5, dan 10) dan *min_samples_leaf* (1, 2, dan 4) sebagai *stopping rules* yang mencegah penggalian informasi lebih dalam (Géron, 2019, hlm. 182). Begitu juga dengan penggunaan *max_features* atau jumlah fitur maksimal dipertimbangkan ketika mencari *split* terbaik, nilai *max_features* yang lebih besar membuat prediksi terhindar dari *overfitting* karena korelasi yang rendah tercipta dari penggunaan fitur yang berbeda dari pohon yang berbeda. Namun *max_features* bergantung pada jumlah fitur sehingga diterapkan nilai (sqrt, log2, dan 0.5) agar tidak terlalu rendah atau terlalu tinggi (Shreyas dkk., 2016).

3.1.4 Studi Deskriptif 2

Tahap keempat yaitu studi deskriptif 2 yang berupa *model evaluation*. *Model evaluation* merupakan tahap validasi model melalui analisis baik atau tidaknya performa model yang memberikan prediksi tidak bias pada data baru yang belum pernah dilihat sebelumnya dengan matriks evaluasi tertentu. (Varoquaux & Colliot, 2023). Matriks evaluasi yang digunakan yaitu *accuracy*, *precision*, *recall*, *f1-score*, *specificity* dan ROC-AUC. Tahap ini memiliki beberapa bagian yaitu matriks evaluasi, *feature importance*, dan identifikasi faktor risiko diabetes.

3.1.4.1 Matriks Evaluasi

Matriks yang digunakan untuk menilai performa model yaitu *accuracy*, *precision*, *recall*, *f1-score*, *specificity* dan ROC-AUC. Keenam matriks ini

digunakan untuk membandingkan performa setiap model, baik itu pada *modeling* awal dengan parameter bawaan juga pada model yang telah melewati proses *hyperparameter tuning*.

3.1.4.2 *Feature Importance*

Feature Importance merupakan tahap untuk mengukur kontribusi individu fitur atau *independent variabel* terhadap kinerja prediksi model. Pengukuran ini terbagi menjadi dua yaitu *modular global* yang mengukur seberapa penting fitur untuk seluruh model, sedangkan *local importance* hanya mengukur pada ruang observasi yang lebih spesifik (Saarela & Jauhiainen, 2021). Salah satu teknik *modular global* ialah *permutation importance*, merupakan salah satu teknik *feature importance* yang bersifat model-agnostik atau dapat diterapkan di berbagai jenis model pembelajaran mesin (Fumagalli dkk., 2023). Peneliti memilih teknik tersebut untuk menunjukkan seberapa besar kontribusi tiap fitur dalam proses prediksi diabetes. Implementasi *permutation importance* menggunakan pustaka ELI5.

3.1.4.3 Identifikasi Faktor Risiko

Tahap ini menganalisis *feature importance* pada setiap model pembelajaran yang efektif untuk mengetahui faktor risiko suatu penyakit. Informasi dari *feature importance* telah umum digunakan untuk menjelaskan bagaimana model bergantung pada faktor risiko tertentu dalam membuat prediksi. Studi terbaru juga mengungkapkan bahwa faktor risiko utama penyakit kardiovaskular di dapatkan dengan teknik *permutation feature importance* (PFI) (Oh dkk., 2022).

Peneliti akan membandingkan *feature importance* dari setiap model untuk melihat urutan fitur-fitur mana saja yang paling berkontribusi membuat prediksi dari yang paling besar hingga terendah. Lalu, peneliti akan melihat fitur mana saja yang paling banyak mendominasi urutan teratas skor *feature importance* tiap modelnya.

3.2 Sumber Himpunan Data

3.2.1 Pemilihan *Dataset*

Peneliti menggunakan data yang bersumber dari *National Health and Nutrition Examination Survey* (NHANES) yaitu data *NHANES August 2021-*

August 2023. *Dataset* dihasilkan dari badan kesehatan yang sangat kredibel. Selain itu, dengan melihat kesenjangan data pada penelitian Islam dkk. (2023) yang sama-sama menggunakan *dataset* NHANES namun dengan perbedaan rentang waktu satu dekade (periode 2009-2012), *dataset* ini memberikan kebaruan terkini lebih dari satu dekade mengenai perilaku kesehatan termasuk penyesuaian pasca pandemi.

NHANES sendiri merupakan survei yang dilakukan oleh *National Center for Health Statistics* (NCHS) di bawah naungan *Centers for Disease Control and Prevention* (CDC) atau Pusat Pengendalian dan Pencegahan Penyakit di Amerika Serikat yang bertujuan untuk menyediakan data kondisi kesehatan masyarakat dan faktor risiko penyakit terkait, memantau tren penyakit, perilaku masyarakat, dan paparan lingkungan. Selain itu, hasil survei juga dapat memenuhi kebutuhan kesehatan masyarakat yang muncul dan mempertahankan informasi dasar representatif skala nasional tentang kesehatan dan gizi di Amerika Serikat. Survei telah dilakukan sejak tahun 1959 dan menjadi siklus rutin dua tahunan sejak 1999, namun pada periode 2019-2020 survei dihentikan karena penyebaran virus COVID-19 dan dilanjutkan kembali sebagai siklus survei baru pada Agustus 2021.

Survei periode baru bernama *NHANES August 2021-August 2023* terbagi ke dalam empat tahap yaitu pemeriksaan kelayakan rumah tangga, wawancara rumah tangga, pemeriksaan *Mobile Examination Center* (MEC), dan wawancara pasca-MEC yang sudah melalui perubahan protokol pelaksanaan untuk mencegah penyebaran virus COVID-19. Data hasil survei kemudian dimodifikasi untuk meningkatkan keakuratan, konsistensi dan untuk melindungi kerahasiaan peserta. Data dikelompokkan ke dalam kategori yaitu *Demographics Data*, *Dietary Data*, *Examination Data*, *Laboratory Data*, *Questionnaire Data*, dan *Limited Access Data* yang dapat diakses di NHANES Questionnaires, Datasets, and Related Documentation. Untuk setiap *dataset*-nya memiliki dua berkas yang dapat diakses yaitu *Doc File* dan *Data File*. *Doc File* berisi penjelasan lengkap *dataset* dan *summary statistics* setiap fitur yang membantu peneliti memilih *dataset* dan fitur mana saja yang digunakan untuk penelitian. Lalu ada *Data File*, merupakan berkas *dataset* itu sendiri yang dapat diunduh dengan ekstensi xpt.

3.2.2 Pemilihan Atribut

Peneliti memilih 21 fitur berdasarkan faktor risiko diabetes yang dikelompokkan menjadi data demografis, pemeriksaan medis, riwayat kesehatan, kesehatan mental, dan gaya hidup. Fitur-fitur tersebut dapat dilihat pada Tabel 3.2.

Tabel 3.2 Fitur Terpilih dan Pengelompokannya Berdasarkan Faktor Risiko

Nama Fitur	Deskripsi Fitur	Kelompok Fitur
RIAGENDR	*Gender*	Data Demografis
RIDAGEYR	*Age in years at screening*	
RIDRETH3	*Race/Hispanic origin w/ NH Asian*	
BPXOSY1	*Systolic - 1st oscillometric reading*	Pemeriksaan Medis
BPXODI1	*Diastolic - 1st oscillometric reading*	
BPXOSY2	*Systolic - 2nd oscillometric reading*	
BPXODI2	*Diastolic - 2nd oscillometric reading*	
BPXOSY3	*Systolic - 3rd oscillometric reading*	
BPXODI3	*Diastolic - 3rd oscillometric reading*	
BMXBMI	*Body Mass Index*	
LBXGH	*Glycohemoglobin*	
LBXTC	*Total Cholesterol*	
BPQ080	*Doctor told you - high cholesterol level*	Riwayat Kesehatan
MCQ160L	*Ever told you had any liver condition*	
MCQ160M	*Ever told you had thyroid problem*	
MCQ220	*Ever told you had cancer or malignancy*	
DPQ020	*Feeling down, depressed, or hopeless*	Kesehatan Mental
DPQ030	*Trouble sleeping or sleeping too much*	
DPQ050	*Poor appetite or overeating*	
ALQ121	*Alcohol Past 12 Month*	Gaya Hidup
SMQ020	*Smoked at Least 100*	

3.2.2.1 Data Demografis

Faktor risiko berdasarkan data demografis yaitu fitur gender, umur, dan ras yang peneliti pilih pada *Demographic Variables and Sample Weights dataset*. Gender, umur, dan ras masuk ke dalam kategori faktor risiko diabetes yang tidak dapat diubah. Bertambahnya usia meningkatkan risiko intoleransi glukosa, hal ini berpeluang lebih besar terjadi pada wanita karena berkaitan dengan mudahnya mereka mengalami peningkatan indeks massa tubuh (Widiasari dkk., 2021).

Meskipun begitu, penelitian yang dilakukan oleh Phan dkk. (2022) pada partisipan dari Rumah Sakit Nasional Endokrinologi di Vietnam menunjukkan bahwa umur merupakan faktor risiko langsung pradiabetes dan diabetes tanpa memandang gender. Hasil penelitian menunjukkan grup umur lebih tua memiliki proporsi yang lebih besar dalam status diabetes, dari yang paling besar yaitu grup umur 60-69 tahun sebesar 10%, 50-59 tahun sebesar 8,8%, 40-49 tahun sebesar 6,4%, dan yang paling rendah 30-39 tahun sebesar 2,7%.

Masih pada penelitian yang sama, Phan dkk. (2022) menemukan bahwa pasien diabetes cenderung didominasi oleh laki-laki. Hal ini berlawanan dengan hasil penelitian yang dilakukan di Cina dan Iran yang didominasi oleh perempuan. Phan dkk. (2022) menyimpulkan perbedaan terjadi karena latar belakang etnis yang berbeda dengan negara lain, begitu juga dengan perbedaan resistensi insulin laki-laki yang lebih tinggi dibandingkan dengan perempuan. Hal ini ditegaskan oleh (Alam dkk., 2021) yang mengatakan bahwa gender dan ras termasuk ke dalam faktor genetik yang merupakan salah satu faktor risiko pradiabetes dan diabetes.

3.2.2.2 Pemeriksaan Medis

Faktor risiko berdasarkan pemeriksaan medis yaitu tekanan darah, BMI, HbA1c, dan kadar kolesterol. Fitur tekanan darah diambil dari beberapa fitur yang berisi nilai sistolik dan diastolik untuk mengukur tekanan darah partisipan dengan tiga kali pembacaan lalu di rata-ratakan. Peneliti menghitung tekanan darah dengan metode *Mean Arterial Pressure* (MAP). MAP merupakan pengukuran tekanan arteri rata-rata selama satu siklus jantung, termasuk fase sistolik dan diastolik (DeMers & Wachs, 2023). Memiliki formula sebagai berikut.

$$\text{Diastolik} + \frac{1}{3}(\text{ Sistolik - Diastolik}) \quad (3)$$

Mengukur tekanan darah partisipan mempermudah deteksi hipertensi atau tekanan darah tinggi yang merupakan salah satu faktor risiko diabetes. Hipertensi membuat sel dalam tubuh kehilangan kesensitifan pada insulin yang berujung resistensi insulin. Aktivitas insulin yang biasanya mengatur metabolisme dengan mengambil glukosa pada sel pun terganggu, sehingga resistensi insulin akan mengakibatkan gangguan pada kadar gula dalam darah (Pratama Putra dkk., 2019). WHO dalam bukunya mengenai hipertensi mengatakan bahwa seseorang dikatakan hipertensi jika tekanan darah sistolik ≥ 140 mmHg dan tekanan darah diastolik ≥ 90 mmHg.

Selanjutnya fitur BMI atau *body mass index* yang bertujuan untuk mengukur lemak tubuh berdasarkan proporsi tinggi dan berat badan terutama untuk partisipan yang obesitas. Obesitas merupakan faktor risiko utama diabetes dengan kontribusi 55% dari semua kasus diabetes tipe 2 (Hardianto, 2020). Obesitas berpengaruh besar mengganggu pengaturan tubuh dalam menjaga glukosa dalam darah di keadaan stabil (homeostatis glukosa sistemik). Hal ini dikarenakan meningkatnya resistensi insulin oleh obesitas yang dapat memicu diabetes tipe 2. Hubungan resistensi insulin karena obesitas dengan diabetes menjadikan mayoritas penderita diabetes tipe 2 memiliki kondisi berat badan berlebih atau obesitas. Meskipun begitu, tidak semua penderita diabetes tipe 2 mengalami obesitas (Banday dkk., 2020). Pengelompokan BMI oleh CDC dapat dilihat pada Tabel 3.3.

Tabel 3.3 Kategori BMI oleh CDC Amerika Serikat

Kategori	**Kisaran BMI (kg/m^2)**
Berat badan kurang	Kurang dari 18.5
Berat badan sehat	18.5 hingga 24
Berat badan berlebih	25 hingga 29
Obesitas tingkat 1	30 hingga 34
Obesitas tingkat 2	35 hingga 39
Obesitas tingkat 3	40 atau lebih

(Sumber: Centers for Disease Control and Prevention, 2024a)

Lalu ada fitur HbA1c yang merujuk pada tes Hemoglobin A1c yang bertujuan untuk mengetahui kadar gula dalam darah rata-rata seseorang pada 2-3 bulan terakhir dengan mengukur persentase Glycohemoglobin (hemoglobin terglikasi). Fitur ini bukan hanya faktor risiko diabetes, namun merupakan indikasi utama untuk mendeteksi seseorang terkena diabetes atau tidak melalui hiperglikemia atau meningkatnya kadar gula darah dalam tubuh karena gangguan insulin yang tidak normal dari biasanya. Lebih lanjut, untuk diabetes tipe 1 hiperglikemia akan muncul karena penghancuran sel T melalui sel β pankreas yang mengakibatkan defisiensi insulin, berbeda dengan diabetes tipe 2 yang disebabkan karena resistensi insulin (Banday dkk., 2020). Menurut Al-hussein dkk. (2025), tes HbA1c sangat umum digunakan untuk diagnosis diabetes karena lebih mudah dan akurat jika dibandingkan tes lainnya seperti tes gula darah puasa terganggu (GDPT) dan tes toleransi glukosa oral (TTGO). Hal ini diperkuat oleh artikel pada situs resmi *American Diabetes Association* (ADA), tes A1c merupakan salah satu cara untuk mengetahui diagnosa awal diabetes pada seseorang dengan detail pada Tabel 3.4 berikut.

Tabel 3.4 Kategori Diagnosis Diabetes Berdasarkan Tes A1c

Kategori Diagnosis	**Persentase A1c**
Normal	Kurang dari 5,7%
Pradiabetes	5,7% hingga 6,4%
Diabetes	6,5% atau lebih

(Sumber: American Diabetes Association, 2025)

Terakhir yaitu fitur Kolesterol yang mengacu pada total kolesterol yang ada dalam tubuh partisipan. Salah satu faktor yang dapat meningkatkan risiko diabetes adalah dislipidemia dengan kadar *high-density lipoprotein* (HDL) kurang dari 35 mg/dL dan/atau kadar trigliserida melebihi 250 mg/dL (Widiasari dkk., 2021). Untuk Total Kolesterol dengan kadar 200 mg/dL atau lebih sudah masuk olesterol tinggi (Centers for Disease Control and Prevention, 2024c). Dislipidemia merupakan gangguan kadar lipid dalam darah tidak normal. Gangguan kadar lipid atau lemak ini ditandai dengan meningkatnya kadar kolesterol total yang terdiri dari

naiknya kadar *low-density liporpotein* (LDL), naiknya kadar trigliserida, dan turunnya *high-density liporpotein* (HDL) (Adityas Trisnadi dkk., 2021).

3.2.2.3 Riwayat Kesehatan

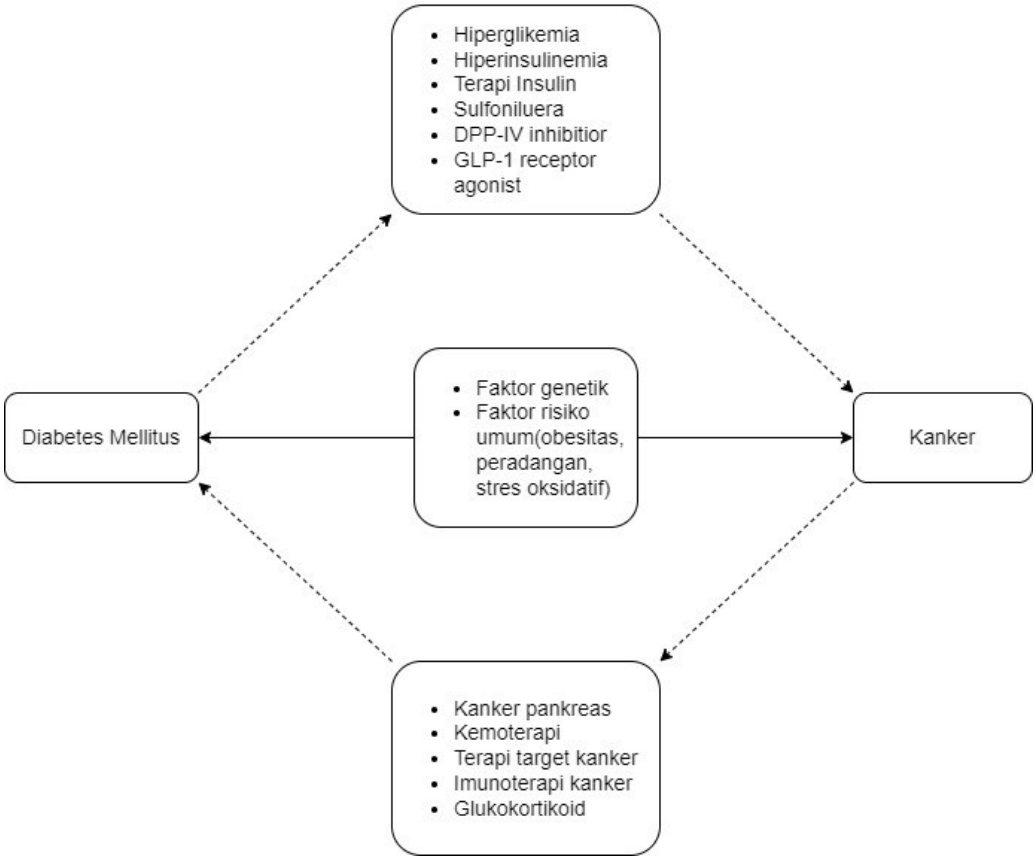
Faktor risiko berdasarkan riwayat kesehatan yaitu riwayat kolesterol tinggi, liver, tiroid, dan kanker. Fitur riwayat kolesterol tinggi erat kaitannya dengan fitur kadar kolesterol pada kategori pemeriksaan medis. Berbeda dengan kategori pemeriksaan medis yang memang didasarkan pada pemeriksaan medis yang dijalani partisipan di MEC saat survei, kategori ini berasal dari kuesioner yang menanyakan apakah partisipan pernah diberitahu oleh dokter atau profesional kesehatan memiliki riwayat penyakit tertentu atau tidak. Oleh karena itu, fitur-fitur pada riwayat kesehatan berfokus pada historis medis partisipan dan semuanya bertipe kategori.

Lalu ada gangguan liver yang memiliki hubungan erat dengan diabetes. *Alanine Aminotransferase* (ALT) dan *Aspartate Aminotransferase* (AST) merupakan dua enzim pada liver yang dilepaskan ke aliran darah saat terjadi kerusakan sel hati, dua enzim ini sering digunakan sebagai indikator penyakit hati berlemak non-alkohol (NAFLD). De Silva dkk. (2019) menemukan bahwa genetik predisposisi terhadap kadar ALT dan AST yang lebih tinggi dapat meningkatkan risiko diabetes.

Fitur selanjutnya yaitu gangguan tiroid. Tiroid atau kelenjar gondok berfungsi dalam produksi hormon-hormon yang penting dalam metabolisme tubuh. Prevalensi disfungsi tiroid terhadap diabetes tipe 1 dan 2 lebih tinggi dibandingkan dengan gangguan non-diabetes menunjukkan hubungan disfungsi tiroid dan diabetes yang dekat. Diabetes tipe 1 berhubungan erat dengan penyakit-penyakit *autoimmune* tiroid. Sedangkan hipertiroidisme dan hipotiroidisme berhubungan erat dengan diabetes tipe 2 (Nishi, 2018). Kelebihan hormon tiroid atau hipertiroidisme terbukti meningkatkan resistensi insulin dan hiperglikemia yang mungkin terjadi karena hipertiroidisme meningkatkan penyerapan glukosa dan sekresi glukosa pada liver. Sedangkan hipotiroidisme terjadi ketika berkurangnya jumlah hormon yang diproduksi dari biasanya. Hal ini dapat menyebabkan berkurangnya laju metabolisme, resistensi insulin, obesitas, dan berbagai faktor

risiko kardiovaskular, hingga memicu munculnya diabetes tipe 2 (Roa Dueñas dkk., 2022).

Fitur terakhir yaitu riwayat kanker. Terdapat hubungan yang kompleks antara diabetes dengan kanker. Risiko kanker dan kematian akibat kanker meningkat karena diabetes tipe 1 dan 2. Sedangkan, beberapa jenis kanker dan terapi kanker berkaitan erat dengan meningkatnya risiko diabetes (Zhu & Qu, 2022). Selain itu, faktor genetik, obesitas, peradangan, stres oksidatif, hiperglikemia, hiperinsulinemia, terapi kanker, insulin, dan beberapa obat hipoglikemik oral berperan dalam hubungan timbal balik antara diabetes dan kanker (Zhu & Qu, 2022). Lebih jelasnya dapat dilihat pada Gambar 3.5.



Gambar 3.5 Mekanisme Hubungan antara Diabetes dengan Kanker

(Sumber: Zhu & Qu, 2022)

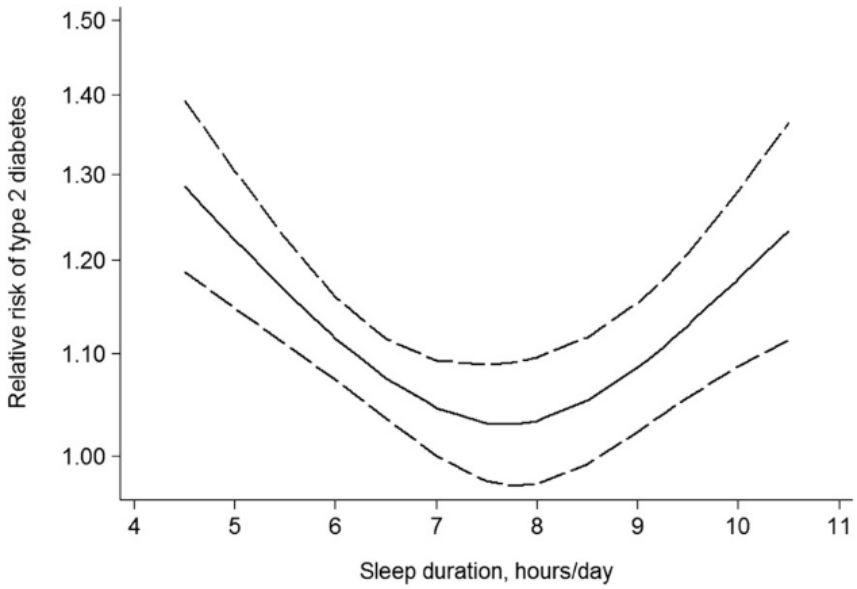
3.2.2.4 Kesehatan Mental

Faktor risiko berdasarkan kesehatan mental yaitu depresi, gangguan tidur, dan gangguan makan. Keempat fitur ini diambil dari *dataset* yang sama yaitu

Mental Health - Depression Screener. C. Zheng dkk. (2024) menganalisis hubungan antara depresi dan risiko diabetes pada *dataset* NHANES dari tahun 2005 hingga 2020 yang menemukan adanya peningkatan tren prevalensi diabetes, depresi, dan masing-masing komorbidnya setiap tahun selama 2005 hingga 2020. Selain itu, peningkatan indikator metabolik seperti trigliserida, insulin, HbA1c, gula darah puasa, dan penurunan kolesterol baik (HDL) muncul seiring dengan meningkatnya tingkat keparahan depresi. C. Zheng dkk. (2024) juga menjelaskan prevalensi diabetes lebih tinggi pada kelompok yang mengalami depresi dan meningkat pada kelompok dengan tingkat depresi yang lebih parah.

Selanjutnya yaitu fitur gangguan tidur. Insomnia yang merupakan salah satu gangguan tidur secara spesifik, merupakan sindrom kronis yang ditandai terganggunya fase memulai tidur dan/atau menjaga kelangsungan fase tidur dapat menjadi indikasi peningkatan risiko gangguan mental. Hasil meta-analisis menemukan bahwa insomnia merupakan indikator signifikan untuk depresi, kecemasan, dan penyalahgunaan konsumsi alkohol (Hertenstein dkk., 2019).

Sebuah penelitian yang dilakukan pada pasien rawat jalan RSU. Pancaran Kasih Manado menemukan bahwa terdapat hubungan yang signifikan antara diabetes dengan kualitas tidur (Tentero dkk., 2016). Selain itu, meta-analisis yang dilakukan oleh Shan dkk. (2015) pada studi prospektif menemukan adanya hubungan antara durasi tidur dan potensi diabetes tipe 2 yang berbentuk pola U. Pola ini menjelaskan bahwa tidur yang terlalu singkat maupun terlalu lama, sama-sama berkaitan dengan peningkatan risiko diabetes tipe 2 secara signifikan. Jika perbandingan dilakukan dengan orang yang tidur dengan durasi 7 jam per hari, maka penurunan satu jam durasi tidur berkaitan dengan peningkatan risiko diabetes sebesar 9%, sedangkan peningkatan satu jam durasi dikaitkan dengan peningkatan risiko diabetes sebesar 14% pada populasi umum. Agar lebih jelasnya dapat dilihat pada Gambar 3.6.



Gambar 3.6 Hubungan antara Durasi Tidur dan Risiko Diabetes Tipe 2
(Sumber: Shan dkk., 2015)

Fitur terakhir yaitu gangguan makan berdasarkan pertanyaan *poor appetite or overeating* (nafsu makan buruk atau makan berlebihan). Rahmayunita dkk. (2023) dalam penelitiannya menjelaskan salah satu faktor pemicu diabetes tipe 2 merupakan faktor lingkungan yang di dalamnya termasuk obesitas, makan berlebihan, kurang makan, olahraga, stres, dan penuaan. Lebih lanjut, peningkatan prevalensi diabetes disebabkan karena perubahan gaya hidup atau konsumsi makanan yang tidak sehat. Diet atau pola makan merupakan sebuah gambaran dari jenis, jumlah, dan komposisi makanan yang dimakan seseorang tiap harinya. Asupan yang mengandung zat seperti karbohidrat atau gula, protein, lemak, termasuk energi berlebih dapat menjadi faktor risiko diabetes (Asyumdah dkk., 2020).

3.2.2.5 Gaya Hidup

Faktor risiko berdasarkan gaya hidup yaitu konsumsi alkohol dan merokok. Fitur yang pertama yaitu konsumsi alkohol yang merujuk pada seberapa sering partisipan mengonsumsi minuman beralkohol dalam 12 bulan terakhir. Studi kohor retrospektif yang dilakukan oleh Cao dkk. (2024) menemukan bahwa terdapat korelasi independen antara konsumsi alkohol berlebih dengan diabetes. Seseorang

dengan konsumsi alkohol berlebih menunjukkan peningkatan potensi diabetes sekitar 73%. Lebih lanjut, pada analisis sub-kelompok BMI, terdapat korelasi antara konsumsi alkohol berlebih dengan diabetes lebih kuat pada seseorang dengan BMI lebih dari 24 kg/m^2.

Selanjutnya yaitu fitur merokok, merujuk pada pertanyaan apakah partisipan pernah merokok minimal 100 kali seumur hidup. Kandungan yang umum pada rokok yaitu nikotin, memperburuk proses homeostasis glukosa dengan cara mengganggu fungsi sel beta, meningkatkan resistensi insulin, dan menyebabkan hiperglikemia secara hormonal (Szwarcbard dkk., 2020).

3.3 Alat dan Bahan Penelitian

Alat dan bahan penelitian yang digunakan pada penelitian ini meliputi perangkat keras dan perangkat lunak. Berikut Tabel 3.5 yang berisi daftar spesifikasi perangkat keras yang digunakan.

Tabel 3.5 Spesifikasi Perangkat Keras

No.	Nama Komponen	Spesifikasi
1	Proccesor	Intel® Core™ i5-12400F
2	Memory	16 GB DDR5
3	Kartu Grafis	GeForce RTX 3060 12 GB
4	Penyimpanan Data	SSD Apacer 512 GB

Berikut Tabel 3.6 yang berisi daftar perangkat lunak yang digunakan.

Tabel 3.6 Perangkat Lunak

No.	Nama	Deskripsi
1	Python 3.12.5	Bahasa pemrograman yang digunakan dalam pengembangan model pembelajaran mesin adalah Python versi 3.12.5.
2	PIP 24.3.1	PIP adalah *Package Installer* untuk bahasa pemrograman Python yang dapat memudahkan pengguna untuk mengintall pustaka Python yang

		dibutuhkan. Versi yang digunakan pada penelitian ini yaitu 24.0.
3	Jupyter Lab	*Platform* interaktif berbasis web untuk menjalankan komputasi dari berbagai bahasa pemrograman. Jupyter Lab sangat unggul dan memudahkan dalam pengembangan model pembelajaran mesin. Peneliti menggunakan Jupyter Lab untuk menulis dan menjalankan kode bahasa pemrograman Python.

Berikut Tabel 3.7 yang berisi daftar pustaka Python yang digunakan.

Tabel 3.7 Pustaka Python

No.	Nama Pustaka	Deskripsi
1	Pandas	Pustaka pada Python yang memiliki banyak kegunaan dalah mengolah data terutama dalam strukturisasi data, analisa data, dan memanipulasi bentuk data.
2	Numpy	Pustaka pada Python yang bertugas untuk menangani data yang berbentuk matriks juga berguna dalam menangani perhitungan saintifik dan matematika tingkat lanjut.
3	Matplotlib	Pustaka pada Python yang berfokus pada visualisasi data agar informasi data dapat dengan mudah dicerna berbagai dengan grafik dasar termasuk kostumisasinya sehingga visualisasi yang dapat dihasilkan sangat fleksibel.
4	Seaborn	Pustaka pada Python digunakan untuk visualisasi data yang hampir sama dengan Matplotlib. Seaborn memiliki sintaks yang lebih ringkas dan varian grafik yang lebih beragam dibandingkan dengan Matplotlib.

5	Functools	Pustaka pada Python yang digunakan untuk melakukan *merge* berulang pada banyak *dataset* dengan fungsi *reduce*.
6	Scikit-learn	Pustaka pada Python yang difokuskan pada pembelajaran mesin. Pustaka ini menyediakan berbagai paket modul yang dapat membantu pengguna dalam pengembangan model pembelajaran. Pada penelitian ini beberapa modul dari Scikit-learn digunakan di antaranya untuk *encoder*, *scaler*, *splitting*, *import* algortima (KNN, SVM, *Random Forest*), matriks untuk evaluasi model, *feature importance*, dan modul optimasi dengan *RandomizedSearchCV*.
7	Imblearn	Pustaka yang masih bagian dari Scikit-learn. Pada penelitian ini, imblearn digunakan untuk memanggil fungsi SMOTEENN.
8	Collections	Pustaka yang digunakan untuk memanggil fungsi counter()
9	XGBoost	Pustaka yang menyediakan akses menyeluruh algoritma *extreme gradient boosting*
10	ELI5	Pustaka yang digunakan untuk melihat *feature importance* melalui teknik *permutation importance*.

3.4 Instrumen Penelitian

Instrumen Penelitian yang digunakan untuk menguji performa model menggunakan matriks-matriks tertentu yaitu *accuracy*, *precision*, *recall*, *f1-score*, *specificity*, dan ROC-AUC. Keenam matriks ini mudah dipahami dengan konsep *confusion matrix* yang merupakan alat ukur performa model klasifikasi yang membandingkan label prediksi dengan label sebenarnya. Lebih jelasnya dapat dilihat pada Tabel 3.8 berikut.

Tabel 3.8 *Confusion Matrix*

Label Kelas	Prediksi Positif	Prediksi Negatif
Aktual Positif	*True Positive* (TP)	*False Negative* (TN)
Aktual Negatif	*False Positive* (FP)	*True Negative* (TN)

- *Accuracy*

Accuracy mengukur proporsi prediksi yang benar dari semua prediksi yang dilakukan. *Accuracy* mengevaluasi seberapa akurat sebuah model mengelompokkan individu ke dalam kategori yang tepat (Miller dkk., 2024). Berikut adalah formula untuk menghitung *accuracy*.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

- *Precision*

Precision mengukur proporsi prediksi berlabel positif yang memang benar-benar positif. *Precision* juga sering disebut *positive predictive value* (PPV) (Miller dkk., 2024). Berikut adalah formula untuk menghitung *precision*.

$$Precision = \frac{TP}{TP + FP} \quad (5)$$

- *Recall*

Recall mengukur proporsi sampel yang sebenarnya positif yang berhasil diprediksi sebagai label positif. Dalam konteks medis, *recall* sering disebut dengan *sensitivity* merupakan parameter uji diagnostik untuk mengenali dengan tepat individu yang memiliki penyakit (Marselin dkk., 2024). Berikut adalah formula untuk menghitung *recall.*

$$Recall = \frac{TP}{TP + FN} \quad (6)$$

- *F1-Score*

F1-Score mengukur hubungan keseimbangan atau harmoni antara matriks *precision* dan *recall*. Skor matriks ini akan tinggi jika *precision* dan *recall* sama-sama memiliki skor yang tinggi. *F1-Core* berfungsi untuk mengukur kemampuan model dalam mengecilkan peluang salah diagnosis (*false negative* dan *false positive*) (Miller dkk., 2024). Berikut adalah formula untuk menghitung *f1-score*.

$$F1\text{-}Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (7)$$

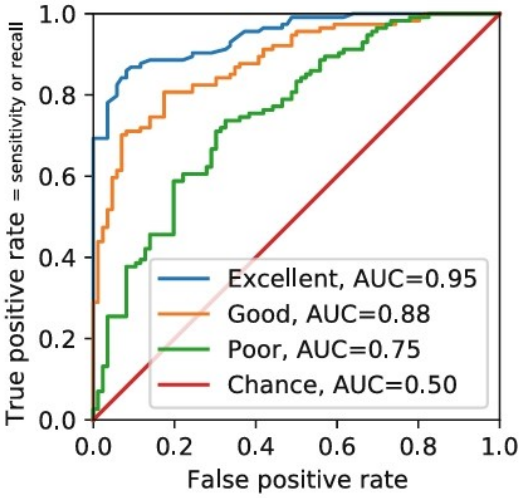
- *Specificity*

Specificity mengukur proporsi sampel yang sebenarnya negatif yang berhasil diprediksi sebagai label negatif. Dalam konteks medis, *specificity* merupakan parameter uji diagnostik untuk mengenali dengan tepat individu yang tidak memiliki penyakit (Marselin dkk., 2024). Berikut adalah formula untuk menghitung *specificity*.

$$Specificity = \frac{TN}{TN + FP} \quad (8)$$

- ROC-AUC

ROC-AUC merupakan kombinasi antara kurva yang menunjukkan seberapa baik model membedakan kelas positif dan kelas negatif pada semua ambang batas klasifikasi (ROC) dengan luas area di bawah kurva (AUC) pada rentang nilai 0 hingga 1. Kurva ROC merepresentasikan nilai *true positive rate* (TPR) dan *false positive rate* (FPR). Ambang batas yang biasa digunakan untuk penilaian klasifikasi biner adalah 0,5. Berdasarkan Gambar 3.7, jika kurva mendekati garis diagonal berwarna merah artinya model cenderung melakukan prediksi yang acak (AUC≈0,5), semakin besar nilai AUC artinya semakin akurat model melakukan prediksi (Varoquaux & Colliot, 2023).

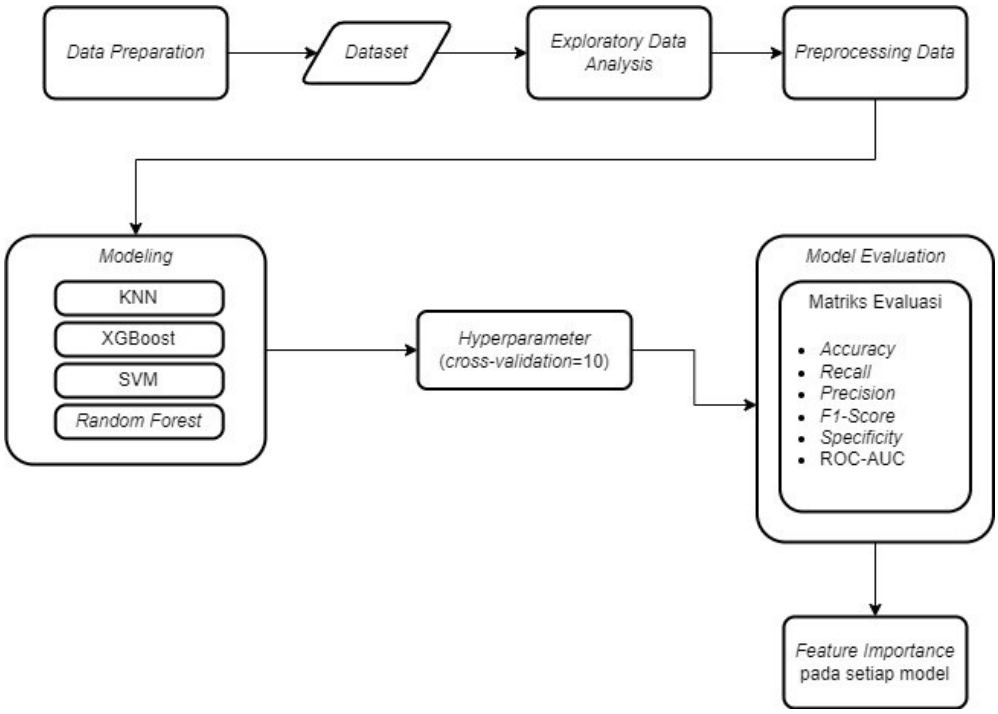


Gambar 3.7 Grafik ROC-AUC

(Sumber: Varoquaux & Colliot, 2023)

3.5 Prosedur Penelitian

Prosedur penelitian dibuat dalam daur hidup model pembelajaran mesin yang setiap prosesnya telah disesuaikan dan dimodifikasi berdasarkan kebutuhan dan efisiensi penelitian yang dilakukan. Setelah melewati daur hidup model pembelajaran mesin, empat model pembelajaran mesin akan memasuki tahap evaluasi model menggunakan matriks *accuracy*, *precision*, *recall*, *f1-score*, *specificity*, dan ROC-AUC. Terakhir adalah melakukan analisa *feature importance* pada setiap model. Untuk lebih jelasnya digambarkan oleh diagram alir pada Gambar 3.8.



Gambar 3.8 Diagram Alir Prosedur Penelitian

3.6 Analisis Data

Hasil performa model yang diperoleh dari empat model dengan algoritma yang berbeda akan dianalisis lebih lanjut dengan analisis komparatif agar mendapatkan kesimpulan dari penelitian yang dilakukan. Instrumen pendukung yang digunakan untuk analisis dan eksplorasi *dataset* adalah bahasa pemrograman Python. Untuk melihat karakteristik data melalui visualisasi digunakan pustaka Seaborn dan Matplotlib. Untuk menentukan *feature importance* pada setiap model dengan algoritma yang berbeda juga menggunakan visualisasi Matplotlib.