# BAB I
# PENDAHULUAN

## 1.1 Latar Belakang

Kesehatan merupakan aset yang berharga bagi setiap individu. Tubuh yang sehat dapat menjadi modal utama untuk menikmati momen penting bersama dengan orang yang kita sayangi. Oleh karena itu, penting untuk terhindar dari berbagai potensi penyakit dan gangguan kesehatan yang dapat menyerang tubuh. Data dari *International Diabetes Federation* (IDF) yang disusun oleh Magliano & Boyko (2025) dalam *IDF Diabetes Atlas 2025 – 11th Edition* mengungkapkan bahwa diabetes merupakan salah satu keadaan darurat kesehatan global yang paling cepat berkembang di abad ini. Pada laporan yang sama, IDF mengungkapkan pada tahun 2024, orang yang memiliki diabetes di seluruh dunia mencapai 588,7 juta jiwa. Angka tersebut bahkan tidak termasuk penderita dengan usia di bawah 20 tahun.

Diabetes Mellitus merupakan gangguan kesehatan kronis yang terjadi ketika kadar gula dalam darah tidak terkendali, disebabkan oleh tubuh yang tidak dapat memproduksi hormon insulin atau tidak bekerja secara efektif (Magliano & Boyko, 2025). Penderita diabetes cepat atau lambat akan merasakan gangguan pada aktivitas sehari-hari hingga dapat menimbulkan risiko kesehatan dengan berbagai komplikasi yang berujung kematian (Tomic dkk., 2022). Sekitar 3,4 juta orang dewasa meninggal karena diabetes per tahun 2024. Yang berbahaya adalah sekitar 251,7 juta orang memiliki diabetes yang tidak terdiagnosis dan 58,7% di antaranya hidup di negara berpendapatan rendah (Magliano & Boyko, 2025). Deteksi dini diperlukan agar dapat mengambil tindakan untuk mengurangi risiko diabetes agar tidak semakin parah. Salah satunya dengan *data mining* yang dapat digunakan untuk mendeteksi dan memperkirakan kondisi penyakit tertentu melalui prediksi berdasarkan pola data atau profil klinis (Kanakaraddi dkk., 2021). Dengan menggunakan salah satu teknik *data mining* yaitu prediksi melalui klasifikasi yang mendeteksi pola-pola yang ada pada data dan mengelompokkannya sesuai diagnosis yang diharapkan (Nazari Nezhad dkk., 2022).

Untuk dapat memanfaatkan potensi penuh algoritma pembelajaran mesin, penting untuk menganalisis perbedaan performa algoritma-algoritma pembelajaran mesin yang memiliki cara kerja berbeda agar menemukan algoritma yang paling sesuai dan andal untuk prediksi sebelum pengambilan keputusan (Alquthami dkk., 2022). Hasil tersebut memperlihatkan algoritma mana yang paling cocok untuk menangani kasus diagnosa diabetes dan implementasinya sebagai alat deteksi dini diabetes. Selain itu, *feature importance* dari hasil kinerja model dapat dianalisis untuk menunjukkan kontribusi fitur terhadap hasil prediksi, analisis ini sudah sering digunakan untuk menunjukkan faktor risiko pada kasus diagnosis tertentu sehingga informasi yang didapat dari data klinis tidak hanya berupa hasil prediksi saja (Oh dkk., 2022).

Penelitian terkait pengembangan model pembelajaran mesin dalam prediksi diabetes salah satunya oleh Kakoly dkk. (2023) yang menggunakan algoritma *Decision Tree*, *Random Forest*, *Support Vector Machine*, *Logistic Regression*, dan *K-Nearest Neighbors*, mendapat skor *accuracy* 82.2% dan AUC 87.2% untuk *Logistic Regression*. Namun Kakoly dkk. (2023) mengungkapkan keterbatasan penelitiannya mengenai performa model yang dapat ditingkatkan dengan teknik seleksi fitur dan optimasi lanjutan seperti *ensemble methods* atau *particle swarm optimization*. Selain itu, performa prediksi dapat ditingkatkan dengan menambahkan lebih banyak fitur klinis yang jumlahnya terbatas.

Pada penelitian lain oleh Islam dkk. (2023) dan Mao dkk. (2023) menemukan bahwa algoritma *Random Forest* unggul dibandingkan dengan algoritma lain yang digunakan. Namun Mao dkk. (2023) masih menggunakan *dataset* yang lama yaitu *dataset Chronic disease research database* of Wuyishan City, Fujian Province, China (REACTION *study*, 2011-2015). Selain itu, penelitian Mao dkk. (2023) hanya menggunakan satu matriks evaluasi saja (AUCROC). Begitu juga penelitian Islam dkk. (2023), meskipun sama-sama menggunakan *dataset* NHANES, namun masih menggunakan *dataset* yang lama (NHANES 2009-2012).

Pada penelitian ini, akan dilakukan perbandingan empat algoritma yang digunakan untuk memprediksi diabetes dengan cara kerja yang berbeda dengan

matriks evaluasi *accuracy*, *precision*, *recall*, *f1-score*, *specificity*, dan ROC-AUC. Algoritma yang dibandingkan yaitu *K-nearest neaighbors* (KNN) yang melakukan prediksi berdasarkan prinsip kedekatan tetangga (Primartha, 2021, hlm. 502). *Support Vector Machine* yang bekerja dengan berusaha mencari *hyperplane* optimal untuk memisahkan kelas-kelas data (Alzubi dkk., 2018). XGBoost yang melakukan prediksi dengan teknik *boosting* yang mengutamakan kecepatan dan performa dengan membuat iterasi banyak model yang mengoreksi kesalahan model sebelumnya (Asselman dkk., 2023). Sedangkan *Random Forest* merupakan varian teknik *bagging* yang melakukan prediksi dengan membuat kombinasi pohon keputusan secara paralel lalu mengambil keputusan dengan *vote* terbanyak dari masing-masing pohon (Primartha, 2021, hlm. 555). Proses pengembangan model akan menggunakan *dataset NHANES August 2021-August 2023* dengan memilih fitur-fitur yang berhubungan dengan faktor risiko diabetes yang dikelompokkan menjadi faktor demografis (gender, umur, dan ras), pemeriksaan medis (tekanan darah, BMI, HbA1c, dan kadar kolesterol), riwayat kesehatan (kolesterol tinggi, liver, tiroid, dan kanker), kesehatan mental (depresi, gangguan tidur, dan gangguan makan), dan gaya hidup (konsumsi alkohol dan merokok) dengan justifikasi masing-masing fitur yang relevan secara klinis. Nantinya Fitur-fitur ini akan dianalisis melalui *feature importance* dengan pustaka Python ELI5 untuk menunjukkan faktor risiko mana saja yang paling berpengaruh dalam meningkatkan potensi diabetes.

## 1.2 Rumusan Masalah

Berikut adalah rumusan masalah yang diajukan:

1. Bagaimana perbandingan performa model pembelajaran mesin menggunakan algoritma *K-Nearest Neighbors*, XGBoost, *Support Vector Machine*, dan *Random Forest* dalam memprediksi risiko diabetes berdasarkan faktor demografis, pemeriksaan medis, riwayat kesehatan, kondisi mental, dan gaya hidup dengan matriks evaluasi *accuracy*, *precision*, *recall*, *f1-score*, *specificity*, dan ROC-AUC?
2. Berdasarkan skor *feature importance* setiap model, faktor risiko apa saja yang berkontribusi terhadap kemungkinan seseorang terkena diabetes?

## 1.3 Tujuan Penelitian

Berdasarkan latar belakang dan rumusan masalah di atas, berikut adalah tujuan dari penelitian ini:

1. Mengembangkan dan menganalisis perbedaan performa model deteksi diabetes dengan algoritma *K-Nearest Neighbors*, XGBoost, *Support Vector Machine*, dan *Random Forest* berdasarkan faktor demografis, pemeriksaan medis, riwayat kesehatan, mental, dan gaya hidup dengan matriks evaluasi *accuracy*, *precision*, *recall*, *f1-score*, *specificity*, dan ROC-AUC.
2. Mengidentifikasi dan menganalisis faktor-faktor risiko utama yang berkontribusi terhadap kemungkinan seseorang terkena diabetes berdasarkan skor *feature importance* melalui pustaka ELI5.

## 1.4 Manfaat Penelitian

Berikut adalah manfaat dari penelitian ini:

1. Memberikan informasi mengenai risiko kesehatan dan faktor risiko pada penyakit diabetes.
2. Memberikan informasi mengenai proses pengembangan dan perbandingan performa model pembelajaran mesin menggunakan algoritma *K-Nearest Neighbors*, XGBoost, *Support Vector Machine*, dan *Random Forest*.
3. Menjadi bahan kajian dan referensi bagi peneliti yang ingin melakukan analisis perbandingan performa model pembelajaran mesin dengan algoritma *K-Nearest Neighbors*, XGBoost, *Support Vector Machine*, dan *Random Forest*.

## 1.5 Batasan Penelitian

Batasan masalah dari penelitian ini:

1. Peneliti berfokus pada proses pengembangan dan perbandingan model deteksi diabetes berdasarkan performa model dan *feature importance*, penelitian ini tidak mencakup implementasi model menjadi sebuah aplikasi.
2. Peneliti berfokus pada klasifikasi biner dengan diagnosis negatif dan positif diabetes, tidak mencakup diagnosis pradiebetes.
3. Algoritma yang dipakai dalam pengembangan model yaitu *K-Nearest*

*Neighbors*, XGBoost, *Support Vector Machine*, dan *Random Forest*.

4. *Dataset* yang digunakan oleh peneliti merupakan gabungan dari kumpulan *dataset NHANES August 2021-August 2023* 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* dengan fitur identifikasi *SEQN* yang merupakan nomor urutan responden. Peneliti memilah fitur yang akan digunakan pada setiap *dataset* berdasarkan faktor risiko diabetes sehingga menghasilkan *dataset* akhir yang memiliki 22 fitur.

## 1.6 Struktur Organisasi Skripsi

BAB I PENDAHULUAN

Bab ini memuat gambaran umum penelitian yang akan dilakukan. Bab pendahuluan terdiri dari latar belakang, rumusan masalah, tujuan penelitian, manfaat penelitian, batasan masalah, dan struktur organisasi skripsi.

BAB II TINJAUAN PUSTAKA

Bab ini berisi pemaparan mengenai topik yang dibahas dengan landasan teori dan penjelasan dari penelitian sebelumnya yang relevan terhadap isu yang diangkat.

BAB III METODE PENELITIAN

Bab ini menjelaskan mengenai metode-metode penelitian yang akan digunakan untuk dapat menyelesaikan rumusan masalah. Terdiri dari desain penelitian, sumber himpunan data, alat dan bahan yang dibutuhkan dalam penelitian, instrumen penelitian, prosedur penelitian, dan analisis data.

BAB IV HASIL DAN PEMBAHASAN

Bab ini berisi hasil dan uraian dari pengembangan dan perbandingan performa model. Bab ini memaparkan hasil perbandingan antara model pembelajaran mesin menggunakan *K-Nearest Neighbors*, XGBoost, *Support Vector Machine*, dan *Random Forest* berdasarkan performa model dengan matriks evaluasi dan *feature importance* untuk mengidentifikasi

faktor risiko diabetes.

BAB V PENUTUP

Bab ini menjelaskan mengenai simpulan dan saran dari hasil penelitian yang akan dimanfaatkan sebagai rujukan, inspirasi, dan ide untuk penelitian selanjutnya.