BAB I

PENDAHULUAN

1.1 Latar Belakang

Diabetes mellitus adalah penyakit kronis yang terus meningkat jumlah penderitanya di seluruh dunia. Pada tahun 2021 sekitar 537 juta orang dewasa menderita diabetes, dan diperkirakan akan meningkat menjadi 643 juta pada tahun 2030 (Sun et al., 2022). Untuk memahami faktor risiko diabetes secara lebih mendalam, banyak penelitian yang menggunakan dataset populasi dalam model prediksi. Salah satu dataset yang sering digunakan adalah Pima Indians Diabetes Dataset (PIDD), yang dikembangkan oleh *National Institute of Diabetes and Digestive and Kidney Diseases*. Dataset ini berisi data medis dari 768 perempuan keturunan Pima Indian di Amerika Serikat, dengan 8 variabel independen dan 1 variabel dependen. Populasi ini dikenal memiliki risiko tinggi terhadap diabetes, sehingga dataset ini sering menjadi referensi dalam pengembangan model prediksi menggunakan *machine learning.*

Salah satu tantangan utama dalam mengurangi dampak diabetes adalah minimnya alat deteksi dini yang akurat dan mudah diakses, terutama di fasilitas kesehatan primer. Metode diagnostik konvensional seperti tes gula darah puasa dan tes toleransi glukosa oral seringkali memerlukan pemeriksaan laboratorium yang memakan waktu dan biaya. Oleh karena itu, penting untuk mengembangkan sistem prediksi yang lebih efisien demi mengidentifikasi individu berisiko tinggi, sehingga tindakan pencegahan dapat dilakukan lebih awal untuk mengurangi kemungkinan terjadinya komplikasi.

Seiring perkembangan teknologi, penerapan *machine learning* dalam bidang kesehatan semakin meluas dan digunakan untuk mendukung prediksi serta pengambilan keputusan klinis. Salah satu algoritma yang banyak diterapkan adalah *Extreme Gradient Boosting* (XGBoost), yaitu sistem boosting berbasis pohon yang sangat efisien dan efektif untuk dataset besar, dengan keunggulan utama pada kecepatan pelatihan, akurasi tinggi, kemampuan menangani *missing value*, serta teknik regularisasi untuk mencegah *overfitting* (Chen & Guestrin, 2016). Model XGBoost yang telah dioptimasi menunjukkan performa terbaik dibandingkan enam algoritma machine learning populer lainnya, dengan capaian akurasi, *precision*, *sensitivity*, dan *F1-score* tertinggi. Potensi XGBoost yang telah di-tuning sebagai alat yang kuat dan efisien untuk deteksi diabetes, sehingga dapat meningkatkan diagnosis medis secara efisien dan personal (Maulana et al., 2023).

Model XGBoost sering kali dianggap sebagai "*blackbox*" karena sulit untuk mendeteksi variabel independen mana yang berpengaruh kuat terhadap hasil prediksi. Untuk mengatasi tantangan ini, digunakan metode interpretasi model seperti *SHapley Additive exPlanations* (SHAP), yang mampu memberikan nilai kontribusi variabel independen terhadap prediksi model. SHAP sangat membantu dalam meningkatkan pemahaman terkait faktor-faktor yang mempengaruhi hasil model, terutama pada model-model yang sulit diinterpretasikan seperti XGBoost (Lundberg & Lee, 2017).

Salah satu tantangan pada Pima Indians Diabetes Dataset adalah ketidakseimbangan kelas, di mana jumlah pasien non diabetes jauh lebih banyak daripada yang menderita diabetes. Kondisi ini dapat membuat model cenderung bias terhadap kelas mayoritas. Untuk mengatasi hal tersebut, digunakan metode SMOTE

yang mensintesis data baru dari kelas minoritas agar distribusi data menjadi lebih seimbang. Penggunaan metode oversampling seperti SMOTE, ADASYN, dan Random Oversampling terbukti meningkatkan sensitivitas model terhadap kelas minoritas, meskipun hasilnya dapat berbeda tergantung algoritma yang dipilih. Selain itu, penerapan metode ini perlu didukung oleh proses validasi yang tepat agar tidak terjadi *overfitting* dan hasil klasifikasi menjadi lebih akurat dan dapat diandalkan (Aqsha & Sunusi, 2023.).

Berbagai penelitian telah mengembangkan model prediksi diabetes menggunakan algoritma XGBoost maupun pendekatan penyeimbangan data seperti SMOTE, namun masih sedikit penelitian yang secara menyeluruh menggabungkan XGBoost dengan interpretasi berbasis SHAP untuk memberikan pemahaman yang mendalam terhadap hasil prediksi. Sebagian besar studi hanya fokus pada akurasi model tanpa mengkaji aspek interpretabilitas, padahal pemahaman terhadap faktor risiko yang berkontribusi terhadap hasil prediksi sangat penting dalam konteks klinis. Selain itu, masih terbatas pula penelitian yang mengevaluasi performa model prediktif dengan mempertimbangkan masalah ketidakseimbangan data dan interpretasi variabel secara bersamaan.

Mempertimbangkan keunggulan XGBoost dalam hal performa prediksi dan kemampuan SHAP dalam interpretasi model, penelitian ini bertujuan untuk menggabungkan kedua pendekatan tersebut dalam membangun sistem prediksi risiko diabetes yang tidak hanya akurat, tetapi juga dapat dipahami dengan baik. Data yang digunakan berasal dari rekam medis pasien atau dataset terbuka dari Pima Indians Diabetes Dataset, yang mencakup variabel-variabel kesehatan penting seperti usia,

BMI, kadar gula darah, tekanan darah, dan riwayat keluarga. Dengan pendekatan ini, diharapkan penelitian ini dapat memberikan gambaran yang lebih mendalam tentang faktor-faktor yang mempengaruhi risiko diabetes, serta membantu tenaga kesehatan dalam menyusun strategi pencegahan yang lebih efektif dan berdasarkan data.

1.2 Rumusan Masalah

Rumusan masalah yang akan dibahas pada penelitian ini adalah:

1. Bagaimana algoritma XGBoost dapat digunakan untuk memprediksi penyakit diabetes menggunakan Pima Indians Diabetes Dataset?
2. Bagaimana kinerja model XGBoost dalam memprediksi penyakit diabetes berdasarkan evaluasi performa model?
3. Bagaimana interpretasi kontribusi variabel independen terhadap prediksi penyakit diabetes menggunakan SHAP?

1.3 Batasan Masalah

Permasalahan dalam penelitian ini dibatasi dengan:

1. Penelitian ini hanya menggunakan variabel yang tersedia dalam Pima Indians Diabetes Dataset, seperti kadar glukosa darah, usia, dan faktor kesehatan lainnya. Variabel tambahan, seperti faktor gaya hidup, tidak disertakan karena keterbatasan dalam dataset yang digunakan.
2. Penelitian ini hanya menggunakan algoritma XGBoost tanpa membandingkan dengan model lain, seperti *random forest* atau *neural network*. Proses *hyperparameter tuning* dilakukan dengan metode Grid Search, tanpa mencoba semua teknik yang tersedia. Selain itu hanya beberapa *hyperparameter tuning* utama, seperti kedalaman maksimum pohon, jumlah minimum data pada setiap

cabang pohon, jumlah total pohon yang digunakan, kecepatan model belajar, proporsi variabel yang digunakan untuk setiap pohon, dan proporsi data latih yang digunakan dalam setiap iterasi yang disesuaikan sehingga tidak mencakup seluruh kemungkinan kombinasi *hyperparameter.*

3. Interpretasi variabel dalam penelitian ini terbatas pada model yang dibangun menggunakan algoritma XGBoost, dengan pendekatan SHAP untuk menjelaskan kontribusi variabel independen terhadap hasil prediksi. Metode interpretasi lainnya tidak digunakan dalam analisis ini.

1.4 Tujuan Penelitian

Tujuan dari penelitian ini yaitu:

1. Membangun model prediksi penyakit diabetes menggunakan algoritma XGBoost, serta pelatihan model untuk menghasilkan prediksi risiko diabetes yang akurat.
2. Mengevaluasi performa model XGBoost dalam memprediksi penyakit diabetes menggunakan evaluasi performa model yang telah ditentukan, yaitu metrik evaluasi, *confusion matrix,* dan kurva ROC untuk menilai efektivitas model dalam melakukan klasifikasi
3. Menganalisis kontribusi variabel independen terhadap hasil prediksi model XGBoost menggunakan SHAP, untuk mengidentifikasi variabel independen mana yang paling berpengaruh terhadap hasil prediksi model.