BAB I
PENDAHULUAN

A. Latar Belakang

Diabetes merupakan penyakit kronis yang menjadi salah satu penyebab utama meingkatnya angka morbiditas dan mortalitas secara global. Berdasarkan laporan *Internasional Diabetes Federation* (IDF), pada tahun 2021 ada sekitar 537 juga orang di seluruh dunia menderita diabetes dan jumlah ini diperkirakan akan meninggkat menjadi 783 juta pada tahun 2045 (IDF, 2021). Sekitar 90% dari seluruh kasus di dunia terjadi dan menjadi penyebab 6,7 juta kematian secara global per tahunnya. Ditambah dengan gaya hidup dan pola makan yang tidak sehat serta obesitas dan faktor keturunan (WHO, 2024).

Prevalensi dan jumlah kasus diabetes (total) di Indonesia dan di setiap provinsi diperkirakan meningkat cukup tinggi pada tahun 2020-2045. Secara nasional, prevalensi dabetes meningkat dari 9,19% pada tahun 2020 dengan 18,69 juta kasus menjadi 16,09% pada tahun 2045 dengan 40,7 juta kasus. Angka ini meningkat 75,1% dalam kurun waktu 25 tahun, dengan peningkatan rata-rata 3% dari prevalensi per tahun (Wahidin *et al.*, 2024). Faktor utama yang berkontribusi terhadap peningkatan ini meliputi pola makan yang tidak sehat, minimnya aktivitas fisik, serta faktor genetik.

Deteksi dini diabetes sangat diperlukan untuk mencegah komplikasi yang lebih serius, seperti gangguan *cardiovaskular,* gagal ginjal, serta neuropati (*American Association Diabetes*, 2022). Namun, gejalanya yang sering kali tidak spesifik pada tahap awal, banyak pasien baru menyadari kondisinya setelah mengalami komplikasi yang lebih parah.

Dalam perkembangan teknologi, khususnya bidang kecerdasan buatan dan *Data Mining*, telah memberikan peluang baru dalam dunia medis. *Machine Learning* dapat digunakan untuk menganalisis data pasien dalam skala besar guna mengidentifikasi pola serta faktor risiko yang

berkontribusi terhadap perkembangan diabetes. Dengan metode ini, memungkinkan sistem prediksi yang lebih akurat dengan mengidentifikasi pola dari data pasien.

Random Forest (RF) merupakan teknik dalam *Machine Learning* yang menggabungkan beberapa pohon keputusan untuk membuat prediksi. Setiap pohon dibangun secara acak dari data pelatihan dan kemudian prediksi dari setiap pohon digunakan unntuk membuat prediksi akhir. Dengan menggunakan *Random Forest*, didapatkan perolehan prediksi yang lebih stabil dan terpercaya (Salsabil, Lutvi and Eviyanti, 2024).

Beberapa peneliti telah melakukan sejumlah penelitian untuk memprediksi penyakit diabetes dengan pendekatan *data mining*. Seperti pada penelitian, menggunakan dua algoritma, salah satunya adalah algoritma *Support Vector Mechine (SVM)* untuk memprediksi penyakit diabetes berdasarkan *dataset* diabetes yang diambil dari dataset *Kaggle*, memperoleh hasil akurasi 78,04% (Maulidah et al., 2021). Sedangkan, dalam penelitian lainnya, algoritma *Random Forest (RF)* digunakan sebagai implementasi dalam melakukan prediksi penyakit diabetes berdasarkan data entri dan data yang diperoleh dari *Kaggle*, *Random Forest (RF)* memperoleh hasil akurasi 74% (Salsabil, Lutvi and Eviyanti, 2024). Dari kedua penelitian ini menunjukkan bahwa *Data Mining* dapat memberikan hasil akurat pada prediksi penyakit diabetes, namun masih harus dilakukan perbandingan lanjutan untuk dapat menentukan algoritma mana yang lebih optimal.

Perbandingan antara algoritma *Support Vector Mechine (SVM)* dan *Random Forest (RF)* untuk memprediksi risiko terjadinya diabetes berdasarkan data yang diperoleh. *Random Forest (RF)* dipilih karena kemampuannya dalam menangani data dengan banyak variabel serta mengurangi *overfitting*, sedangkan *Support Vector Mechine (SVM)* lebih dikenal dengan keunggulanya dalam menangani data *non linear* dengan dimensi tinggi dan dapat memberikan hasil klasifikasi yang akurat.

Keunggulan kedua algoritma ini telah dibuktikan dalam berbagai penelitian. Sebuah studi menunjukkan bahwa *Random Forest* menghasilkan performa terbaik dibandingkan *Naïve Bayes* dan *Support Vector Machin* (SVM) dalam tugas klasifikasi prediksi ketidakhadiran kerja, dengan nilai akurasi sebesar 99,38%, *precision* sebesar 99,42%, dan *recall* sebesar 99,39%. Hasil ini memperkuat keyakinan bahwa RF mampu memberikan hasil prediksi yang sangat akurat dan konsisten pada berbagai domain data (Nalatissifa *et al.*, 2021).

Studi lain yang membandingkan tiga algoritma *Support Vector Machin* (SVM), *Random Forest* (RF), dan *K-Nearest Neighbors* (KNN) menunjukkan bahwa *Support Vector Machin* (SVM) memberikan nilai *overall accuracy* (OA) tertinggi, dengan sensitivitas paling rendah terhadap variasi jumlah sampel pelatihan. *Random Forest* (RF) berada di posisi berikutnya, diikuti oleh *K-Nearest Neighbors* (KNN). Ketiganya mampu mencapai akurasi di atas 93,85% ketika ukuran data pelatihan lebih dari 750 sampel per kelas, baik pada dataset yang seimbang maupun tidak seimbang. Temuan ini menunjukkan bahwa *Support Vector Machine* (SVM) dan *Random Forest* (RF) tidak hanya unggul dari segi akurasi, tetapi juga tangguh terhadap variasi ukuran data, menjadikannya cocok untuk diterapkan dalam prediksi penyakit berbasis data terbatas seperti kasus diabetes(Thanh Noi and Kappas, 2017).

Berdasarkan uraian di atas, peneliti tertarik untuk melakukan penelitian tentang Perbandingan Akurasi Algoritma *Support Vector Machine* (SVM) dan *Random Forest* (RF) untuk Prediksi Penyakit Diabetes. Peneliti menggunakan dataset *Pima Indian Diabetes Database*, namun dalam dataset tersebut terdapat salah satu tantangan utama yang dihadapi adalah ketidakseimbangan *(imbalanced dataset),* di mana jumlah pasien non-diabetes secara signifikan lebih banyak dibandingkan dengan pasien diabetes. Ketidakseimbangan ini berisiko membuat model lebih memihak pada kelas mayoritas. Untuk mengatasi hal tersebut, digunakan teknik penyeimbangan data seperti SMOTE *(Synthetic Minority*

Oversampling Technique) yang berfungsi menambahkan sampel sintetis pada kelas minoritas, sehingga model dapat lebih akurat dalam mengenali dan memprediksi kasus diabetes.

Selain itu, kinerja algoritma klasifikasi seperti *Support Vector Machine* (SVM) dan *Random Forest* (RF) sangat bergantung pada pengaturan *hyperparameter* yang tepat. Jika parameter tidak diatur secara optimal, model dapat mengalami *overfitting* atau *underfitting*. Oleh karena itu, penelitian ini menerapkan proses tuning *hyperparameter* menggunakan *GridSearchCV* dengan pendekatan *cross-validation*, guna menemukan kombinasi parameter terbaik yang mampu meningkatkan performa model, terutama dalam hal akurasi, *presisi, recall*, dan *f1-score* pada tugas prediksi diabetes.

B. Perumusan Masalah

Berdasarkan latar belakang, maka didapatkan rumusan masalah sebagai berikut:

1. Algoritma mana yang lebih akurat berdasarkan dataset yang digunakan?
2. Bagaimana performa algoritma *Support Vector Machine* (SVM) dan *Random Forest* (RF) dalam prediksi diabetes?

C. Tujuan

1. Membandingkan kinerja algoritma *Random Forest* (RF) dan *Support Vector Machine* (SVM) dalam memprediksi risiko diabetes dengan menggunakan metrik evaluasi untuk menentukan performa terbaik dengan penerapan teknik SMOTE pada dataset *Pima Indian Diabetes Database.*

2. Mengimplementasikan algoritma *Support Vector Mechine* (SVM) dan *Random Forest* (RF) untuk prediksi diabetes, sehingga dapat memudahkan dalam pencegahan risiko diabetes.

D. Manfaat

1. Manfaat bagi Peneliti
   a. Memberikan pemahaman lebih mendalam mengenai penerapan algoritma *Support Vector Mechine* (SVM) dan *Random Forest* (RF) dalam bidang Kesehatan.
   b. Dapat mengasah kemampuan peneliti dalam pengolahan data hingga evaluasi performa model. Melalui pemrograman, *Data Mining*, dan *Machine Learning*.
   c. Menambah ilmu dan wawasan penelitian penulis.
2. Manfaat bagi Instansi Pendidikan
   a. Dapat digunakan sebagai referensi untuk mahasiswa dan dosen dalam bidang teknologi informasi dan kesehatan.
   b. Sebagai dasar penelitian selanjutnya dalam pengembangan model prediksi penyakit.
3. Manfaat bagi Tenaga Kesehatan
   a. Digunakan sebagai pertimbangan pengambilan keputusan.
   b. Memberikan hasil indentifikasi pasien diabetes dengan cepat.

E. Keaslian Penelitian

Penelitian ini dilakukan sebagai upaya untuk mengembangkan metode prediksi penyakit diabetes yang lebih akurat dan aplikatif melalui pendekatan *komparatif* algoritma *Support Vector Machine (SVM)* dan *Random Forest (RF)*. Meskipun topik serupa telah dikaji dalam beberapa penelitian terdahulu, penelitian ini memiliki keaslian pada aspek metodologi, yaitu penggunaan teknik penyeimbangan data *Synthetic*

Minority Oversampling Technique (SMOTE), penerapan evaluasi performa model secara komprehensif, serta integrasi hasil model ke dalam antarmuka aplikasi web berbasis *Streamlit*. Dengan pendekatan tersebut, penelitian ini diharapkan memberikan kontribusi ilmiah yang orisinal dan relevan dalam pengembangan teknologi prediktif di bidang Kesehatan.

Tabel 1. 1 Keaslian Penelitian

No.	Judul Penelitian dan Nama Penulis	Metode Penelitian	Hasil Penelitian	Perbedaan
1.	Klasifikasi menggunakan metode *Support Vector Machine* dan *Random Forest* untuk Deteksi Awal Risiko Diabetes Melitus, Chea Zahrah Vaganza Junus; Tarno; Puspita Kartikasari; (Junus, Tarno and Kartikasari, 2023)	*Random Forest (RF)* dan *Support Vector Machine (SVM)*	Performa klasifikasi dari metode *Support Vector Machine* menghasilkan nilai akurasi sebesar 91%, *recall* sebesar 86%, *precision* sebesar 97% dan *F1-Score* sebesar 91%. Performa klasifikasi dari metode *Random Forest* menghasilkan nilai akurasi sebesar 98%%, *recall* sebesar 98%, *precision* sebesar 99% *dan F1-Score* sebesar 98%.	Peneliti menggunakan teknik SMOTE untuk menyeimbangkan data
2.	Prediksi Penyakit Diabetes Melitus Menggunakan	*Metode Support Vector Machine*	Nilai akurasi untuk model metode *Support Vector Machine*	Peneliti menggunakan data sekunder dari *Pima Indian Diabetes*

	Metode Support Vector Machine dan *Naive Bayes*, Nurlaelatul Maulidah; Riki Supriyadi; Dwi Yuni Utami; Fuad Nur Hasan; Ahmad Fauzi; Ade Christian (Maulidah *et al.*, 2021)	dan *Naive Bayes*	adalah 78,04% dan nilai akurasi untuk metode *Naive Bayes* 76,98%. Perbedaan akurasinya adalah 1,06%.	*Database* dengan 768 baris dan 8 atribut.
3.	*Comparison Support Vector Machine and Random Forest Algorithms in Detect Diabetes*, Habib Alrasyid1; Ahmad Homaidi; M. Kom.2, Zaehol Fatah, M. Kom. (Alrasyid *et al.*, 2024)	*Support Vector Machine* dan *Random Forest*	Algoritma SVM menghasilkan nilai akurasi 77% dan algoritma *Random Forest* menghasilkan akurasi 75%. Sehingga, algoritma SVM dinyatakan kebih cocok dalam mendeteksi diabetes.	Peneliti menggunakan Teknik SMOTE untuk menyeimbangkan data dengan pembagian rasio data 70:30 dan memperoleh akurasi tinggi pada algoritma *Random Forest*.

Berdasarkan Tabel Keaslian Penelitian, dapat diketahui bahwa beberapa studi sebelumnya telah menerapkan algoritma *Support Vector Machine* (SVM) dan *Random Forest* (RF) dalam prediksi penyakit diabetes. Penelitian yang dilakukan oleh Junus, Tarno, dan Kartikasari (2023) menggunakan dataset diabetes sebanyak 520 data dan menunjukkan bahwa algoritma *Random Forest* memberikan hasil

klasifikasi yang lebih tinggi dibandingkan dengan SVM. *Random Forest* menghasilkan akurasi dan *F1-score* masing-masing sebesar 98%, sementara SVM mencapai akurasi sebesar 91% dan *F1-score* 91% . Penelitian tersebut juga menerapkan proses *hyperparameter tuning* untuk memperoleh akurasi optimal (Junus et al., 2023). Perbedaan dengan penelitian ini terletak pada penggunaan dataset serta penerapan teknik SMOTE dalam penelitian ini untuk menangani ketidakseimbangan data.

Selanjutnya, penelitian oleh Nurlaelatul Maulidah et al. (2021) membandingkan algoritma SVM dan *Naïve Bayes* menggunakan dataset berjumlah 2000 data dari basis data kesehatan Diabetes Dataset. Hasil penelitian menunjukkan bahwa algoritma SVM memberikan akurasi sebesar 78,04%, sedikit lebih tinggi dibandingkan *Naïve Bayes* yang menghasilkan akurasi sebesar 76,98%. Namun, penelitian tersebut tidak menerapkan teknik *balancing data* seperti SMOTE, serta tidak menyertakan algoritma *Random Forest* dalam perbandingan, sehingga belum memberikan evaluasi menyeluruh terhadap algoritma yang lebih kompleks dan banyak digunakan saat ini (Maulidah et al., 2021).

Penelitian oleh Habib Alrasyid et al. (2024) membandingkan performa SVM dan *Random Forest* dalam prediksi diabetes dengan menggunakan teknik SMOTE untuk penyeimbangan data. Berdasarkan dataset berjumlah 768 data, hasil menunjukkan bahwa algoritma SVM menghasilkan akurasi sebesar 77%, sedikit lebih unggul dibandingkan *Random Forest* yang memperoleh akurasi sebesar 75%. Penelitian tersebut membagi data latih dan data uji dengan rasio 80:20 (Alrasyid et al., 2024). Adapun dalam penelitian ini digunakan rasio pembagian data sebesar 70:30, dengan penerapan SMOTE dan proses *hyperparameter tuning* yang bertujuan untuk mengoptimalkan kinerja model klasifikasi yang dibangun.

Dengan demikian, keaslian penelitian ini terletak pada perbandingan langsung antara algoritma SVM dan *Random Forest* dalam prediksi penyakit diabetes menggunakan dataset *Pima Indian Diabetes Database*,

serta penerapan SMOTE dan *hyperparameter tuning* secara terintegrasi guna memperoleh hasil prediksi yang lebih optimal dan representatif.