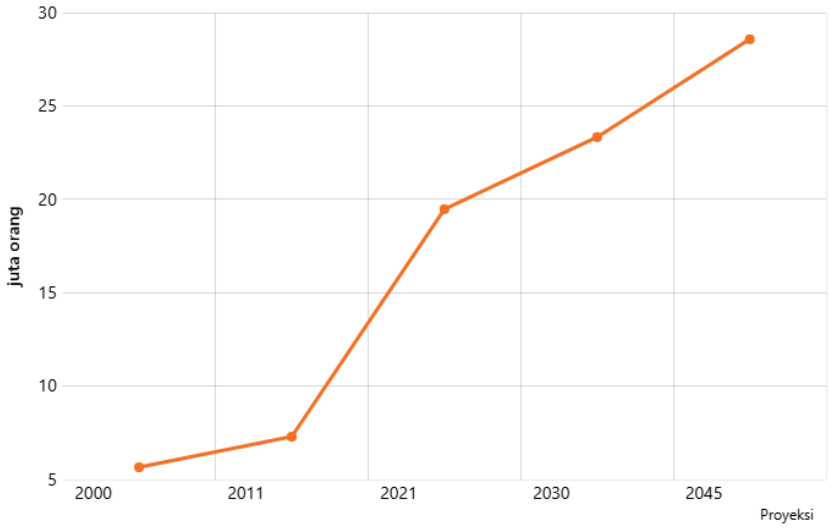
BAB I
PENDAHULUAN

1.1. Latar Belakang

Pada abad ke-21 ini, *diabetes mellitus* menjadi lebih umum, bahkan prevalensinya meningkat dalam beberapa tahun terakhir [1]. Data dari databoks sendiri menunjukkan bahwa di tahun 2030, penderita diabetes di Indoenesia akan mencapai lebih dari 20 Juta jiwa, seperti yang dapat dilihat di Gambar 1.1.



Gambar 1. 1 Prevalensi Diabetes di Indonesia [2]

Diabetes mellitus sendiri adalah sebuah penyakit yang tergolong kronis. Penyakit tersebut muncul ketika pankreas tidak mampu memproduksi insulin yang cukup (tipe 1) atau ketika tubuh tidak mampu menggunakan insulin dengan efektif (tipe 2)[3]. Akibatnya, terjadi peningkatan gula darah yang dapat berdampak buruk pada berbagai organ dalam tubuh. Meskipun seorang penderita *diabetes mellitus* mungkin menunjukkan gejala-gejala tertentu, mendiagnosis bahwa seorang tersebut menderita *diabetes mellitus* akan sangat sulit. Sebab gejala yang timbul sering kali sama dengan gejala penyakit non-kronis pada umumnya [4]. Sehingga, deteksi dini penyakit *diabetes mellitus* perlu dilakukan, supaya penderita penyakit ini bisa segera ditangani, sebelum penyakit menjadi makin parah dan berdampak buruk bagi kesehatan [5]. Salah satu pendekatan yang dapat dilakukan untuk melakukan deteksi dini adalah melalui pendekatan *machine learning*. Pada bulan

Januari lalu, telah dilaukan suatu penelitian untuk meningkatkan performa prediksi deteksi dini penyakit diabetes dengan pendekatan *machine learning* [6]. Penelitian dilakukan dengan melakukan imputasi pada data menggunakan KNN *Imputer*, serta menggunakan tri-ensemble model. Hasilnya, akurasi model meningkat menjadi 97,49%, yang kemungkinan besar dipengaruhi kuat oleh penggunaan tri-ensemble model selain dari imputasi KNN itu sendiri.

Dalam penelitian terkait analisis data, masalah *missing value* atau data yang hilang merupakan tantangan umum yang sering dihadapi [7]. *Missing value* dapat terjadi karena berbagai alasan, seperti kesalahan entri data, ketidaklengkapan informasi, atau faktor lainnya. Dalam dunia medis, *missing value* bisa terjadi karena pasien tidak mau memberi jawaban, pasien tidak melakukan tindak lanjut, error pada peralatan medis, atau faktor lainnya [8]. *Missing value* menyebabkan hilangnya informasi penting, bias dalam analisis, serta kesulitan dalam pemodelan, karena model tidak memiliki data yang cukup untuk memperoleh hasil prediksi yang maksimal [9]. Jika masalah *missing value* ini tidak ditangani, maka dapat mengurangi akurasi dari hasil analisis [10].

Dalam kasus analisis data medis, data yang memiliki masalah *missing value* perlu dilakukan penanganan yang tepat untuk memastikan hasil prediksi model yang akurat dan dapat diandalkan. Karena diperlukan tingkat akurasi model yang tinggi karena model berkaitan dengan kesehatan pasien. Untuk menangani *missing value*, terdapat dua cara yang dapat dilakukan. Pertama yakni drop atau menghapus data. Pada metode ini, setiap data yang memiliki nilai hilang pada satu atau lebih atributnya akan dihapus dari tabel [11]. Kedua yakni impute atau imputasi. Pada metode ini, setiap nilai hilang akan diganti dengan nilai baru [12]. Di dalam metode imputasi ini terdapat berbagai macam teknik terkait bagaimana proses ‘nilai baru’ tersebut dibuat dan/atau dimasukkan. Beberapa diantaranya adalah KNN Imputation, Mean Imputation, dan masih banyak lagi [13].

Pada kasus prediksi *diabetes mellitus*, setiap fitur dalam dataset berpotensi memiliki kontribusi penting dalam memprediksi kondisi penyakit. Ketika mengunakan teknik drop, maka bisa terjadi kehilangan informasi berharga dari data yang dapat membantu meningkatkan akurasi prediksi. Di sisi lain, teknik imputasi *missing value* menawarkan pendekatan yang lebih baik dalam mengatasi masalah data hilang [14]. Dengan menggunakan metode imputasi yang tepat, kita dapat memperkirakan nilai-nilai yang hilang secara lebih akurat berdasarkan informasi yang tersedia dalam data, sehingga mempertahankan integritas dataset dan meningkatkan kualitas analisis serta akurasi prediksi *diabetes mellitus*.

Pada tahun 2023, sebuah penelitian menunjukkan hasil F1-score meningkat hingga 9% pada empat dataset yang berbeda setelah dilakukan imputasi pada data [15]. Pada tahun yang sama, penelitian lain menunjukkan hasil *accuracy score* meningkat hingga15% pada sembilan model yang berbeda setelah imputasi dilakukan [16]. Adapun pada tahun 2021, terdapat sebuah penilitian yang mengkaji tentang perbandingan performa enam jenis teknik imputasi yang berbeda, pada 31 dataset multiclass classification [17]. Hasilnya, pemberlakuan imputasi menghasilkan kurang lebih 10%-20% peningkatan akurasi pada model klasifikasi. Selain itu ditemukan juga bahwa imputasi dengan teknik *random forest* menghasilkan kualitas imputasi dan peningkatan akurasi model terbaik dibandingkan teknik imputasi lainnya. Adapun pada tahun 2023, telah dilaukan suatu penelitian untuk melakukan deteksi dini penyakit *diabetes mellitus* [18]. Data pada penelitian tersebut menggunakan ‘Pima Indians’ dimana data tersebut memiliki masalah *missing value*. Penelitian dilakukan dengan menerapkan KNN *Imputer* untuk menangani *missing value* dengan menggunakan model *random forest*. Hasilnya, model menunjukkan akurasi sebesar 77.06%. Dalam analisis data medis, tingkat akurasi model di bawah 90% belum bisa dikatakan andal, karena tingkat ketidak-tepatan prediksi masih terbilang cukup tinggi untuk bisa mengambil keputusan berdasarkan prediksi model [19].

Berdasarkan penelitian-penelitian tersebut, diketahui bahwa metode imputasi mampu meningkatkan akurasi prediksi model secara signifikan. Namun, pada kasus data *diabetes mellitus*, meski sudah dilakukan imputasi dengan KNN, akurasi yang didapat masih cukup rendah. Sehingga, untuk mengatasi potensi risiko terkait *missing value* yang dapat memengaruhi akurasi prediksi model pada kasus *diabetes mellitus*, diperlukan teknik imputasi data yang tepat untuk menagangani *missing value*. Dengan demikian, dibuatlah penelitian berjudul 'ANALISIS PERBANDINGAN KINERJA TEKNIK IMPUTASI UNTUK PENANGANAN *MISSING VALUE* (STUDI KASUS DATASET *DIABETES MELLITUS*)'. Penelitian ini bertujuan untuk mengeksplorasi dan membandingkan kinerja berbagai teknik imputasi dalam menangani *missing value* pada dataset *diabetes mellitus* secara khusus. Besar harapannya hasil penelitian ini dapat memberikan rekomendasi berbasis bukti mengenai teknik imputasi yang paling efektif untuk menangani *missing value* pada dataset diabetes. Rekomendasi ini dapat menjadi landasan bagi penelitian selanjutnya untuk membangun model prediksi dengan akurasi yang lebih tinggi.

1.2.Rumusan Masalah

Teknik imputasi KNN untuk penanganan missing value masih kurang andal untuk meningkatkan akurasi prediksi model deteksi dini penyakit *diabetes mellitus*. Sementara itu, terdapat berbagai teknik imputasi lain yang berpotensi menghasilkan performa lebih baik, tetapi belum dieksplorasi secara komprehensif pada kasus *diabetes mellitus*.

1.3. Tujuan Penelitian

1. Menganalisis dampak berbagai teknik terhadap kualitas dan karakteristik dataset *diabetes mellitus* yang mengandung *missing value*, dengan fokus pada bagaimana setiap teknik mengubah atau mempertahankan distribusi data asli.

2. Menganalisis perbandingan kinerja teknik imputasi KNN *Imputer*, *MissForest*, dan MICE dalam menangani masalah *missing value* pada kasus *diabetes mellitus*, menggunakan metrik akurasi, presisi, *f1-score*, dan *recall*.
3. Menentukan teknik imputasi terbaik dengan mengevaluasi signifikansi peningkatannya dibandingkan dengan baseline KNN *Imputer*.

1.4. Batasan Masalah

1. Data yang digunakan adalah dataset 'Pima Indians Diabetes Database' yang diambil dari kaggle.
2. Teknik imputasi yang akan dibandingkan adalah KNN *Imputer*, MICE, dan *MissForest*. MICE dan *MissForest* karena mewakili pendekatan yang berbeda: imputasi berganda berbasis model linear (MICE), dan imputasi berganda berbasis model ensemble (*MissForest*).
3. Model klasifikasi utama yang digunakan untuk mengevaluasi dampak dari berbagai teknik imputasi adalah *RandomForestClassifier*.
4. Hiperparameter untuk teknik imputasi yang diuji dalam penelitian ini menggunakan nilai yang telah ditentukan berdasarkan praktik umum atau studi pendahuluan, dan bukan merupakan hasil dari proses optimasi hiperparameter khusus untuk *imputer* itu sendiri.
5. Biaya komputasi dari masing-masing teknik imputasi tidak menjadi fokus utama dalam analisis perbandingan.

1.5. Manfaat Penelitian

1. Meningkatkan Deteksi Dini Penyakit (Target SDG 3.4): Temuan pada penelitian ini berpotensi menjadi landasan untuk meningkatkan akurasi model prediksi, sehingga sistem deteksi dini *diabetes mellitus* menjadi lebih andal dan pada gilirannya dapat membantu mengurangi angka kematian prematur akibat penyakit tidak menular.
2. Memperkuat Sistem Kesehatan dan Analisis Data Medis (Target SDG 3.d): Hasil penelitian ini dapat berfungsi sebagai referensi teknis untuk meningkatkan kualitas serta keandalan data medis yang krusial dalam diagnosis.

3. Mendorong Inovasi di Bidang Kesehatan (Target SDG 3.8 & 3.d): Penelitian ini turut memberikan kontribusi pada pengembangan ilmu *machine learning* dan penerapannya secara praktis dalam sektor kesehatan. Dengan demikian, temuan ini dapat mendorong terciptanya layanan kesehatan masa depan yang berbasis data, yang lebih baik dan lebih akurat.