

BAB I

PENDAHULUAN

1.1 LATAR BELAKANG

Diabetes merupakan penyakit metabolisme jangka panjang yang berdampak pada tubuh manusia dengan cara mengubah glukosa darah menjadi energi. Individu yang telah didiagnosis dengan diabetes tidak dapat mengontrol level gula dalam darah, yang akan menyebabkan peningkatan kadar gula darah dan tekanan darah. Jika diabetes tidak dikenali, didiagnosis, dan tidak diobati dengan tepat sejak awal, Diabetes akan berdampak pada kondisi jangka panjang yang serius dan berdampak besar besar yang berdampak pada kehidupan dan kesejahteraan individu, keluarga, dan masyarakat di seluruh dunia. Ini adalah salah satu dari 10 penyebab utama kematian pada orang dewasa. dan diperkirakan telah menyebabkan empat juta kematian secara global pada tahun 2017 dan pada tahun 2017 pengeluaran kesehatan global untuk diabetes diperkirakan mencapai USD 727 miliar [1].

Tantangan utama dalam pengendalian diabetes adalah keterbatasan akses terhadap layanan kesehatan dan kurangnya kesadaran masyarakat tentang faktor resiko diabetes. lebih dari 50% kasus diabetes tidak terdiagnosis karena minimnya *screening* dan deteksi dini. Kondisi ini diperburuk dengan terbatasnya sumber daya kesehatan, terutama di daerah *rural* dan *remote*.

Seiring dengan perkembangan teknologi, machine learning telah menjadi alat yang potensial dalam mendukung diagnosis dan prediksi penyakit diabetes.

Berbagai model *machine learning* seperti *Random Forest*, *Logistic Regression*, *Support Vector Classification* (SVC), dan *Gradient Boosting* telah menunjukkan hasil yang menjanjikan dalam memprediksi diabetes dengan tingkat akurasi yang tinggi [2]. Namun, salah satu tantangan utama dalam penerapan model *machine learning* di bidang kesehatan adalah bagaimana mengkomunikasikan hasil prediksi secara efektif kepada para praktisi medis dan pasien.

Penelitian sebelumnya yang berjudul “Perbandingan Support Vector Machine dan Modified Balanced Random Forest dalam Deteksi Pasien Penyakit Diabetes ” Penelitian ini bertujuan untuk membandingkan berbagai algoritma Machine Learning dalam mendeteksi penyakit diabetes. Model yang akan dibandingkan meliputi *Support Vector Machine* (SVM), *Modified Balanced Random Forest* (MBRF), dan *Random Forest*. Evaluasi dilakukan dengan menggunakan metrik seperti akurasi, presisi, *recall*, dan *F1-score* guna menentukan metode yang paling efektif dalam mendeteksi diabetes berdasarkan dataset yang digunakan. Namun, belum banyak penelitian yang secara komprehensif membandingkan berbagai algoritma *Machine Learning* pada dataset yang sama untuk memprediksi penyakit diabetes. Oleh karena itu, penelitian ini berupaya mengisi celah tersebut dengan melakukan perbandingan kinerja beberapa algoritma Machine Learning pada data prediksi penyakit diabetes. Urgensi dan Signifikansi Penelitian Penelitian ini memiliki nilai penting karena: Memastikan akurasi prediksi yang lebih baik dengan mengevaluasi keunggulan masing-masing algoritma *Machine Learning*. Mengatasi ketidakseimbangan kelas dengan teknik resampling, sehingga meningkatkan deteksi pasien diabetes. Membantu pengambilan keputusan medis dengan

menyediakan model prediksi yang lebih andal untuk dokter dan tenaga kesehatan. Dengan adanya penelitian ini, diharapkan dapat memberikan kontribusi dalam pemilihan algoritma terbaik untuk meningkatkan kualitas prediksi diabetes, sehingga diagnosis dapat dilakukan lebih cepat dan akurat.

Sebagai peneliti di bidang sistem informasi kesehatan, peneliti menemukan bahwa *Random Forest* adalah metode klasifikasi dalam statistika yang berbasis komputasi. Metode klasifikasi digunakan untuk pembelajaran fungsi-fungsi berbeda yang memetakan masing-masing data terpilih ke dalam salah satu dari kategori kelas yang telah ditetapkan[3]. *Support Vector Classification* (SVC) memberikan pendekatan yang unik dalam klasifikasi diabetes melalui konsep *hyperplane* dan margin maksimal. algoritma *Support Vector Classifier* (SVC) untuk mengklasifikasi pesan spam dengan menggunakan atribut data training berupa numerik dan teks[4]. Menurut website resmi *Scikit-learn*, algoritma *SVC* merupakan algoritma berbasis *library SVM*. Adapun atribut numerik berisi panjang teks (*length*) dan jumlah tanda baca (*punct*), sedangkan atribut teks (*message*) adalah isi keseluruhan pesan. Regresi Logistik (Logit) merupakan suatu metode analisis statistika yang mendeskripsikan hubungan antara peubah respon (*dependent variable*) yang bersifat kualitatif memiliki dua kategori atau lebih dengan satu atau lebih peubah penjelas (*independent variable*) berskala kategori atau interval[5]. *Gradient Boosting* (GBR) adalah sebuah algoritma yang terdiri dari beberapa pohon keputusan untuk melakukan tugas klasifikasi dan regresi[6]

Penelitian ini bertujuan untuk menganalisis dan membandingkan visualisasi data dari empat model machine learning (*Random Forest*, *Logistic Regression*, *SVC*, dan *Gradient Boosting*) dalam prediksi penyakit diabetes. Hasil penelitian ini diharapkan dapat memberikan panduan dalam pemilihan teknik visualisasi yang optimal untuk mendukung interpretasi hasil prediksi diabetes berbasis *machine learning*, serta meningkatkan efektivitas komunikasi hasil prediksi kepada praktisi medis dan pasien. Penelitian ini menjadi penting mengingat kebutuhan akan sistem prediksi diabetes yang tidak hanya akurat tetapi juga dapat dipahami dengan baik oleh pengguna akhir. Dengan membandingkan berbagai teknik visualisasi dari model yang berbeda, penelitian ini diharapkan dapat memberikan kontribusi signifikan dalam pengembangan sistem prediksi diabetes yang lebih efektif dan *user-friendly*.

Berdasarkan latar belakang yang sudah dijelaskan maka peneliti tertarik untuk membuat laporan penelitian tesis yang berjudul “Perbandingan Data Prediksi Penyakit Diabetes Menggunakan *Machine Learning*”.

1.2 RUMUSAN MASALAH

Berdasarkan latar belakang diatas, maka dapat dirumuskan permasalahan dalam penelitian ini yaitu:

1. Bagaimana performa keseluruhan, sensitivitas, spesifisitas, dan akurasi model *Random Forest*, *Logistic Regression*, *SVC*, dan *Gradient Boosting* dalam memprediksi penyakit diabetes?

2. Bagaimana visualisasi hasil prediksi dari setiap model berbeda, terutama dalam hal membantu memahami pola data dan membuat keputusan?
3. Apakah model yang memiliki akurasi terbaik juga memiliki kompleksitas komputasi yang paling ideal untuk digunakan dalam sistem prediksi kehidupan nyata?
4. Sejauh mana setiap model dapat mengidentifikasi pola data pada kumpulan data diabetes, terutama dengan data yang tidak seimbang atau bervariasi?

1.3 BATASAN MASALAH

Untuk menghindari berbicara tentang hal-hal di luar tema dan judul penelitian, masalah yang akan dijadikan acuan akan dibatasi oleh batas-batas berikut:

1. Dataset yang digunakan dalam penelitian ini dibatasi pada dataset diabetes yang mencakup variabel-variabel standar seperti kadar glukosa, BMI, tekanan darah, usia, dan variabel relevan lainnya. Berikut Dataset yang digunakan: “<https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset/data>”
2. Model machine learning yang dibandingkan terbatas pada empat algoritma, yaitu:
 - a) *Random Forest*
 - b) *Logistic Regression*
 - c) *Support Vector Classifier* (SVC)
 - d) *Gradient Boosting*

Tidak mencakup model-model machine learning lainnya yang mungkin juga dapat digunakan untuk prediksi diabetes.

3. Aspek visualisasi data yang akan dibandingkan meliputi:
 - a) *Confusion matrix* dari masing-masing model
 - b) *ROC Curve* dan *AUC Score*
 - c) *Performance metrics visualization* (accuracy, precision, recall, F1-score)
4. Penelitian ini berfokus pada perbandingan aspek visualisasi dan interpretabilitas model, bukan pada optimasi parameter atau peningkatan performa model secara mendalam.
5. Evaluasi performa model hanya menggunakan metrics standar (*accuracy*, *precision*, *recall*, *F1-score*, dan *AUC-ROC*), tidak mencakup metrics khusus lainnya.

1.4 TUJUAN PENELITIAN

Adapun tujuan penelitian ini disusun berdasarkan perumusan masalah yang telah diidentifikasi sebelumnya sebagai berikut:

1. Menganalisis perbandingan efektivitas visualisasi data dari empat model machine learning (*Random Forest*, *Logistic Regression*, *SVC*, dan *Gradient Boosting*) dalam memprediksi penyakit diabetes.
2. Mengidentifikasi model *machine learning* yang memberikan representasi visual paling informatif dan mudah diinterpretasi untuk prediksi diabetes.

3. Mengevaluasi kekuatan dan kelemahan masing-masing model dalam hal kemampuan visualisasi performa dan interpretabilitas hasil prediksi.
4. Menganalisis bagaimana berbagai teknik visualisasi dapat membantu pemahaman tentang proses prediksi diabetes untuk setiap model yang digunakan.
5. Memberikan rekomendasi teknik visualisasi yang paling efektif untuk menjelaskan hasil prediksi diabetes kepada berbagai pemangku kepentingan (*stakeholders*).

1.5 MANFAAT PENELITIAN

Adapun manfaat pada penelitian ini yaitu :

1. Membantu menemukan model yang paling efektif dalam memprediksi diabetes berdasarkan performa visualisasi data.
2. Memberikan wawasan tentang bagaimana visualisasi hasil prediksi dapat membantu dalam interpretasi dan pengambilan keputusan medis.
3. Mengembangkan metode visualisasi yang lebih efektif dalam menunjukkan pola dan tren terkait faktor risiko diabetes.
4. Mengurangi "*black box*" effect dalam *machine learning* dengan menghadirkan hasil prediksi dalam bentuk visual yang lebih transparan dan mudah diakses.
5. Menyediakan pemahaman mendalam tentang kekuatan dan keterbatasan tiap model dalam menampilkan performa dan interpretasi hasil prediksi
6. Mempermudah analisis faktor risiko dan pola yang berkontribusi terhadap

prediksi diabetes dengan tampilan visual yang lebih intuitif.

7. Memudahkan tenaga medis, peneliti, dan pengambil Keputusan dalam mengambil keputusan yang lebih akurat dan cepat berdasarkan hasil prediksi yang divisualisasikan dengan efektif.

1.6 SISTEMATIKA PENULISAN

Supaya penyusunan laporan dapat mudah dan teratur, penulis telah menyusun rencana yang terbagi dari 5 bab, secara spesifik sebagai. Berikut:

BAB I : PENDAHULUAN

Bab ini membahas latar belakang masalah, perumusan masalah, pembatasan masalah, tujuan dan keuntungan dari penelitian, dan proses penulisan.

BAB II : LANDASAN TEORI

Bagian ini akan menjelaskan teori dasar yang berkaitan dengan penelitian ini serta teori pendukung. Selain itu, akan dibahas juga beberapa penelitian sejenis yang telah dilakukan sebelumnya sebagai referensi dan pembandingan dalam penelitian ini.

BAB III : METODOLOGI PENELITIAN

Bagian ini akan memaparkan struktur penelitian serta menguraikan langkah-langkah melaksanakan penyelesaian penelitian.

BAB IV :ANALISIS

Bab ini menyajikan hasil dari eksperimen yang telah dilakukan, termasuk hasil prediksi penyakit diabetes menggunakan machine learning dan analisis visualisasi data yang dihasilkan

BAB V : PENUTUP

Bab terakhir ini berisi kesimpulan dari penelitian yang telah dilakukan serta saran untuk pengembangan lebih lanjut terkait visualisasi data dalam prediksi penyakit diabetes menggunakan machine learning.