BAB I
PENDAHULUAN

1.1.Latar Belakang

Diabetes adalah peningkatan kadar gula dalam darah yang berlebihan dan terjadi ketika pankreas tidak memproduksi insulin, atau meskipun pankreas memproduksi insulin, tubuh tidak dapat menggunakannya secara efektif.[1] Gula darah sangatlah vital bagi kesehatan karena merupakan sumber energi yang penting bagi sel-sel dan jaringan.[2] Gula darah yang tinggi berdampak pada berbagai organ tubuh manusia dan terkadang menimbulkan komplikasi pada banyak fungsi tubuh, khususnya pembuluh darah dan saraf.[3]

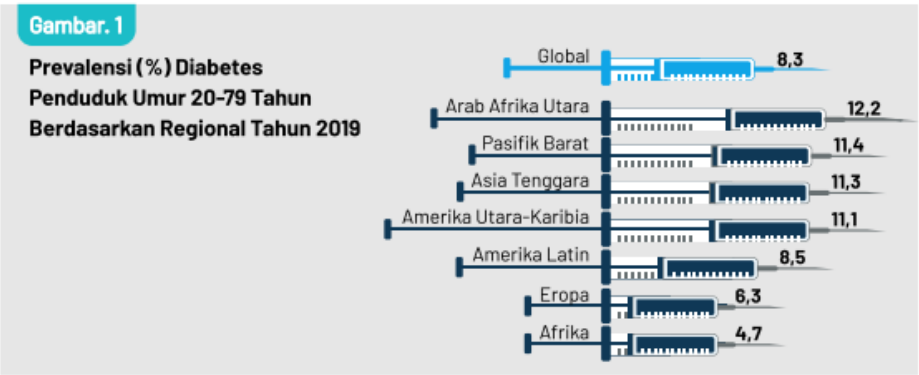
Diabetes menjadi penyakit mematikan jika tidak ditangani dengan baik. Berbagai penelitian mengenai Kesehatan atau *healthcare* ada banyak sekali dengan berbagai studi kasus yang berbeda sehingga data penelitian sangat diperlukan.[4] Contohnya penelitian prediksi penyakit diabetes yang mengambil data dari Bangladesh dengan penerapan model algoritma machine learning Decision Tree sebagai model terbaik yang memiliki akurasi sebesar 73,5%. Organisasi medis, di seluruh dunia, mengumpulkan data tentang berbagai masalah terkait kesehatan.[5] Salah satunya adalah data para penderita penyakit diabetes di dunia dan penerapan model algoritma *machine learning*.

Proses *screening* awal penyakit diabetes bisa dilakukan menggunakan *machine learning. Screening* awal penyakit diabetes sangatlah penting karena penerapan serangkaian prosedur atau tes yang dilakukan untuk mendeteksi potensi gangguan kesehatan atau penyakit tertentu pada seseorang. Tujuannya adalah deteksi dini untuk mengurangi resiko penyakit atau memutuskan metode pengobatan yang paling efektif. *Screening* awal

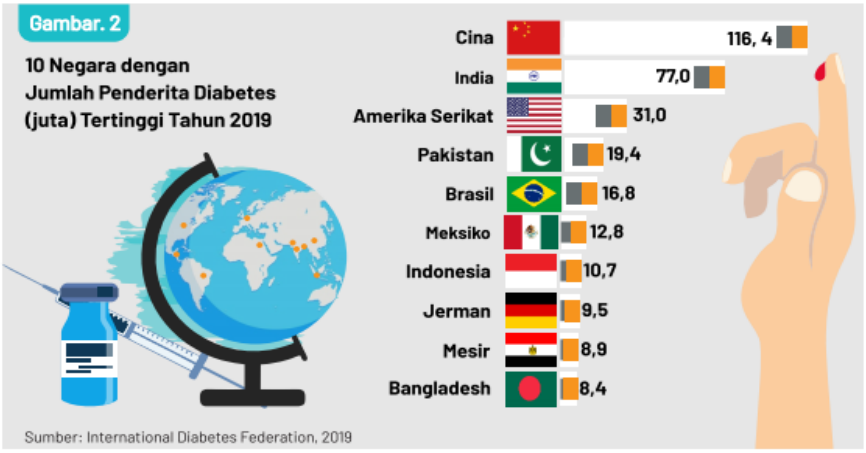
penyakit dengan penerapan *machine learning* didukung hasil akurasi model algoritma yang diterapkan secara *real time* kepada pengguna.

Machine Learning (ML) kini menjadi semakin populer dan telah dilaporkan sebagai salah satu metode paling efektif dalam berbagai aplikasi dalam perawatan kesehatan *preventif*.[6] *Machine Learning* memiliki keuntungan terkait seperti komputasi yang relatif murah, ketahanan, kemampuan generalisasi dan kinerja tinggi.[1] Dalam berkembangnya alat, perlengkapan dan peralatan medis, pengetahuan lanjutan dapat diperoleh di bidang diagnosis penyakit. Pengambilan keputusan dengan bantuan komputer, yaitu *Machine Learning* membantu manusia dengan memproses kumpulan data medis yang kompleks dan menganalisisnya untuk memberikan wawasan klinis. [7], [8] Ekstraksi pengetahuan dari data merupakan faktor penting untuk prediksi dan diagnosis penyakit dalam industri medis.[9], [10]

Salah satu algoritma *Machine Learning* adalah K-Nearest Neighbors. Algoritma ini mencari nilai k yang optimal dalam proses evaluasi model. Nilai k yang optimal dapat mempengaruhi nilai akurasi data. Selain itu, proses pengolahan data tidak terlepas dari *overfitting* yang dapat mempengaruhi tingkat akurasi. Sehingga perlu adanya optimasi untuk meminimalkan *overfitting*. *Cross Validation* adalah salah satu cara untuk meminimalkan *overfitting* dalam data.[11] Data yang dimaksudkan adalah data penyakit diabetes.



Gambar 1. 1 Prevalensi (%) Diabetes Penduduk di Tahun 2019 [12]



Gambar 1. 2 Prevalensi (%) Diabetes Penduduk di Tahun 2019 [12]

Berdasarkan referensi dari *International Diabetes Federation* pada tahun 2019, terlihat Gambar 1.1 yang menunjukkan prevalensi diabetes dalam persentase penduduk umur 20 sampai 79 tahun. Daerah yang memiliki prevalensi paling tinggi adalah Arab Afrika utara dengan tingkat pravelensi 12,2 %. Kemudian 10 negara dengan jumlah penderita diabetes tertinggi dalam tingkatan juta yang terlihat pada Gambar 1.2. Negara yang memiliki jumlah penderita diabetes tertinggi adalah Cina dan berjumlah 116,4 juta penderita. Selain itu, Indonesia merupakan negara peringkat ketujuh yang memiliki penderita diabetes sebanyak 10,7 juta penderita.

Dengan melihat data penderita diabetes yang semakin meningkat tiap tahunnya di seluruh dunia, *prevalensi* diabetes diperkirakan akan terus meningkat secara eksplosif, dari 578,4 juta (10,2%) pada tahun 2030 menjadi 700,2 juta (10,9%) pada tahun 2045.[13] Sehingga penyakit diabetes ini menjadi rahasia umum masyarakat dan tidak sedikit yang mengalaminya, khususnya masyarakat dari negara Indonesia.

Dalam hal ini, ada aspek positif bahwa manusia mencoba mengatasi masalah Kesehatan diabetes agar tidak terkena ataupun kondisi penderita penyakit diabetes semakin parah. Diabetes adalah penyakit jangka panjang yang tidak dapat disembuhkan. Namun, gejala dan komplikasi selanjutnya dapat dikendalikan dengan perawatan yang diperlukan dan gaya hidup sehat.

Dengan demikian, penelitian ini memiliki arti penting karena memungkinkan mendeteksi diabetes pada tahap awal.[1]

Dalam hal ini, penulis menggunakan model algoritma *machine learning K-Nearest Neighbors* untuk proses perhitungan akurasi dari dataset yang sudah dipersiapkan sebelumnya. Dataset yang akan diuji, penulis ambil dari dokumen asli *National Institute of Diabetes and Digestive and Kidney Diseases* pada tahun 2021 [14]. Model *Machine Learning Supervised Learning* yang digunakan adalah *K nearest neighbors*.[15]. *K nearest neighbors* adalah salah satu algoritma klasifikasi terbaik dan banyak digunakan dalam berbagai penelitian dengan parameter k yang optimal.

Sebelumnya, ada tambahan Teknik *cross validation* untuk optimasi algoritma *K-nearest neighbors.* Teknik ini akan mengoptimalkan hasil akurasi dan konsep yang penting dalam Ilmu data dan analisis data. Itu pun digunakan untuk mencegah atau setidaknya meminimalkan *overfitting. Overfitting* berarti bahwa model yang disesuaikan terlalu banyak dengan data pelatihan. Pada umumnya, saat mengerjakan kumpulan data kecil, pilihan ideal adalah validasi silang k-fold dengan nilai k yang besar (tetapi lebih kecil dari jumlah instance) atau validasi silang tanpa-satu sedangkan saat mengerjakan kumpulan data kolosal, pikiran pertama adalah untuk gunakan validasi ketidaksepakatan.

Model algoritma yang dikembangkan pada akhirnya di-deploy agar layak digunakan secara *real time. K-Nearest Neighbors* (KNN) dan *grid search cross validation* (CV) telah digunakan untuk melatih dan mengoptimalkan model untuk memberikan hasil terbaik. Keuntungannya adalah akurasi dalam prediksi yang telah terlihat sebesar 80%. [16] Kemudian membuat *interactive web application* menggunakan streamlit. Streamlit adalah *library python* untuk membuat model yang dibangun menjadi berbasis web dan mudah digunakan.[17]. Streamlit akan digunakan sebagai alat pengujian secara *real time* kepada user berdasarkan hasil akurasi yang diperoleh dari penelitian ini.

Berdasarkan latar belakang diatas, metode yang digunakan dalam menghitung akurasi dan prediksi atau *screening* awal penyakit diabetes adalah *K-Nearest Neighbors* dengan *Cross Validation* untuk meminimalkan *overfitting.* Model yang dibangun menggunakan library streamlit berbasis *interactive web application* untuk pengujian *screening* awal penyakit diabetes terhadap pengguna berdasarkan hasil akurasi model algoritma yang digunakan.Sehingga penelitian ini akan dituangkan ke dalam tugas akhir dengan judul **"OPTIMASI ALGORITMA K-NEAREST NEIGHBORS DENGAN TEKNIK CROSS VALIDATION DENGAN STREAMLIT (Studi Data: Penyakit Diabetes)".**

1.2.Perumusan Masalah

Berdasarkan latar belakang yang telah diuraikan diatas, maka dapat diketahui permasalahan bahwa penerapan *K-Nearest Neighbors* sebagai algoritma klasifikasi harus memperhatikan Teknik validasi silang dalam evaluasi algoritma. Proses evaluasi data tanpa validasi silang bisa mengalami kesulitan dikarenakan memakan waktu komputasi yang lama pada jumlah attributes dataset yang banyak. Selain itu, Teknik validasi silang digunakan untuk meminimalkan *overfitting.* Kemudian belum diketahuinya performasi akurasi mengenai prediksi penyakit diabetes menggunakan algoritma *K-Nearest Neighbors* dengan *Teknik cross validation.*

1.3.Pertanyaan Penelitian

Berdasarkan rumusan masalah yang telah disebutkan diatas, dapat menghadirkan pertanyaan sebagai berikut:

1. Bagaimana cara meningkatkan proses evaluasi data pada algoritma *K-Nearest Neighbors* menggunakan Teknik *cross validation* dalam prediksi penyakit diabetes dan meminimalkan *overfitting.*
2. Berapa nilai akurasi berdasarkan perhitungan *confusion matrix* dari proses algoritma *K-Nearest Neighbors* dengan Teknik *cross validation* dalam prediksi penyakit diabetes.

1.4.Batasan Masalah

Berdasarkan rumusan masalah dan tujuan penelitian, maka untuk mewujudkan penelitian yang sesuai dengan masalah yang ada diperoleh batasan-batasan masalah penelitian sebagai berikut:

1. Dataset yang digunakan adalah dataset diabetes berupa data csv yang berasal dari *National Institute of Diabetes and Digestive and Kidney Diseases.*
2. Difokuskan pada cara untuk mengatasi kesulitan proses evaluasi data pada algoritma *K-Nearest Neighbors* dengan Teknik *cross validation* dalam penelitian prediksi penyakit diabetes.
3. Difokuskan untuk mendapatkan kinerja hasil akurasi dan hasil klasifikasi dari algoritma *K-Nearest Neighbors* dengan Teknik *cross validation* dalam proses penelitian prediksi penyakit diabetes.

1.5.Tujuan Penelitian

Berdasarkan rumusan masalah yang ada maka dapat diketahui tujuan dari penelitian ini adalah:

1. Mengoptimalisasi proses komputasi yang lama dan meminimalkan *overfitting* dalam memberikan estimasi akurasi pada algoritma *K-Nearest Neighbors* dengan menggunakan teknik *cross validation* pada prediksi penyakit diabetes.
2. Mengetahui hasil akurasi melalui perhitungan *confusion matrix* dari proses algoritma *K-Nearest Neighbors* dengan teknik *cross validation* dalam prediksi penyakit diabetes.

1.6.Manfaat Penelitian

Berdasarkan uraian diatas maka penelitian ini mempunyai manfaat teoritis dan praktis sebagai berikut:

1. Manfaat teoritis,
   - Bagi Penulis dapat menjadi tambahan Ilmu pengetahuan dalam penerapan algoritma K-Nearest Neighbors dengan Teknik cross validation
   - Bagi Institusi dapat menjadi referensi bagi yang hendak melakukan penelitian mengenai optimasi algoritma untuk prediksi penyakit.
2. Manfaat praktis,
   - Bagi Penulis dapat menjadi tambahan Ilmu dalam bidang machine learning serta menjadi referensi untuk penelitian selanjutnya.
   - Bagi Institusi dapat menjadi tambahan referensi untuk penelitian selanjutnya.