BAB I

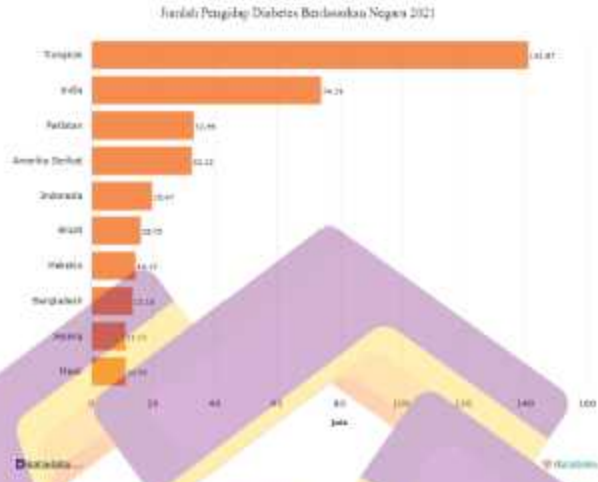
PENDAHULUAN

1.1 Latar Belakang

Diabetes merupakan penyakit kronis yang ditandai dengan tingginya kadar gula (glukosa) dalam darah yang melebihi batas normal [1].

Menurut situs resmi *Word Health Organization (WHO)* sekitar 422 juta orang di seluruh dunia menderita diabetes, mayoritas yang tinggal di negara berpenghasilan rendah dan menengah dan 1,5 juta kematian secara langsung dikaitkan dengan dengan diabetes setiap tahun [2]. Baik jumlah kasus maupun prevalensi diabetes terus meningkat selama beberapa dekade terakhir. Selama kurun waktu dua dekade terakhir diabetes masuk ke dalam 10 penyebab kematian terbanyak di dunia dan sejak tahun 2000 mengalami kenaikan sebesar 70%. WHO juga memprediksikan bahwa pada tahun 2030 penyakit diabetes akan menjadi 7 penyakit yang menyebabkan kematian di dunia [4]. Keterlambatan dalam diagnosis penyakit diabetes adalah satu penyebab terjadinya lonjakan jumlah kematian. Pasien yang mengidap penyakit diabetes dan meninggal kebanyakan dikarenakan diagnosis yang terlambat sehingga komplikasi semakin parah [4]. Dengan banyaknya kematian yang disebabkan oleh penyakit diabetes, maka diperlukan tindakan awal yaitu dengan melakukan deteksi dini [4].

Di Indonesia sendiri kasus diabetes tidak kalah banyaknya. Dari situs data databox kita dapat lihat Indonesia berada di posisi ke-5 negara dengan kasus diabetes tertinggi setelah Amerika Serikat.



Gambar 1.1 *Jumlah Kasus Diabetes Di Indonesia*

Para peneliti memusatkan perhatian untuk deteksi dini penyakit diabetes atau menghambat komplikasi yang berkelanjutan. Dengan banyaknya kasus diabetes di Indonesia dari grafik data yang sudah dipaparkan pada gambar 1.1, bisa disimpulkan bahwa dari banyaknya kasus tersebut dihasilkan data dari pasien diabetes dan dari data pasien diabetes tersebut bisa diolah menggunakan teknik data mining untuk melakukan deteksi dini penyakit diabetes. Untuk melakukan deteksi tersebut digunakan teknik *data mining* untuk mencari informasi dari sebuah histori data diabetes . *Data Mining* adalah proses untuk menemukan pola yang menarik dan pengetahuan dari data yang besar. Sumber data dapat mencakup database, data warehouses, web, atau tempat penyimpanan informasi lainnya [5]. Metode yang digunakan untuk pengekstrakan pengetahuan dalam data mining salah satunya adalah klasifikasi [6]. Klasifikasi adalah sebuah proses untuk menemukan model yang membedakan kelas data untuk dapat memprediksi kelas dari suatu objek yang

kelasnya tidak diketahui sebelumnya [6]. Dalam penelitian kali ini kita akan menggunakan metode klasifikasi. Pada metode klasifikasi terdapat beberapa algoritma yang dapat digunakan yaitu antara lain *C4.5, Naïve Bayes (NB), K-Nearest Neighbor (kNN), Regresi Linear, Artificial Neural Network* dan lain-lain [7] . Pada penelitian ini algoritma yang akan digunakan adalah *C4.5, Naïve Bayes (NB), dan K-Nearest Neighbor (kNN).* Metode-metode tersebut digunakan pada penelitian ini dikarenakan mempunyai akurasi yang baik pada penelitian sebelumnya. Pada penelitian yang pernah dilakukan oleh Erni Ermawati [8], menggunakan algoritma *C4.5* mendapatkan akurasi sebesar 98.56 % dan nilai AUC sebesar 0.979 yang berarti termasuk ke dalam kategori *Excellent Classification* tanpa dilakukan optimasi. Algoritma *C4.5* memiliki dasar ide yaitu melalui *Decision Tree* atau pembentukan pohon keputusan. Yang dimana pohon keputusan merupakan metode klasifikasi yang sangat mudah dipahami dan dapat diinterpresentasikan dengan cepat. Sehingga dapat membantu dalam memecahkan masalah dalam pengambilan keputusan [9]. Pada penelitian lainnya yang dilakukan oleh Rahayu Febryani dan Toni Arifin [10], menggunakan algoritma *Naïve Bayes* mendapatkan hasil akurasi sebesar 97.22 % dan nilai AUC sebesar 0.991 dengan kombinasi pembobotan *Particle Swarm Optimization* yang termasuk kategori *Excellent Classification.* Algoritma *Naïve Bayes* juga dapat diaplikasikan pada data yang jumlahnya besar dan data yang mempunyai *missing value* serta dapat menangani data dengan atribut yang tidak relevan [11]. Sementara itu kNN juga termasuk algoritma yang termasuk ke dalam kategori baik, pada penelitian yang dilakukan oleh Nia Nuraeni [12], memperoleh hasil tingkat akurasi sebesar 90,39%.

Dalam beberapa tahun terakhir, penelitian terhadap penyakit diabetes ini sudah dilakukan dengan berbagai macam algoritma klasifikasi untuk deteksi dini. Berikut adalah beberapa percobaan yang sudah dilakukan, antara lain penelitian [13], pada penelitian yang dilakukan yaitu melakukan komparasi algoritma *Decision Tree* dan *Naïve Bayes* untuk prediksi dini penyakit diabetes. Pada penelitian yang dilakukan menggunakan dataset primer yang terdiri dari 520 data pasien dan 17 atribut. Pada metode penelitian yang dilakukan, *cross validation* dilakukan untuk melakukan pengujian model, selanjutnya evaluasi model dilakukan menggunakan *confusion matrix* untuk mendapatkan algoritma terbaik dalam prediksi dini penyakit diabetes. Hasil penelitian yang diperoleh adalah algoritma *decision tree* lebih baik dalam memprediksi penyakit diabetes ditandai dengan nilai akurasi 95,58% dan nilai AUC 0,981 lebih tinggi dibandingkan dengan naïve bayes yaitu akurasi 87,69% dan nilai AUC 0,947. Peneliti juga menyarankan untuk penelitian selanjutnya dilakukan optimasi sehingga nilai akurasi bisa lebih baik lagi. Penelitian lainnya yaitu penelitian [1], penelitian mengenai deteksi dini penyakit diabetes dengan mengkomparasi algoritma *Decision Tree, Naïve Bayes,* dan *Neural Network*. Menggunakan metode penelitian seperti yang dilakukan kebanyakan penelitian yaitu pengumpulan data, pengolahan data awal, metode, pengujian model, dan yang terakhir evaluasi dan validasi hasil. Hasil penelitian yang diperoleh adalah *Decision tree* dengan akurasi 96,96%, *Naïve Bayes* dengan akurasi 87,69%, dan *Neural Network* dengan akurasi 61,54%. Penelitian tersebut juga menyarankan untuk dilakukan optimasi pada penelitian selanjutnya dengan menggunakan *Feature Selection*, Algoritma Genetika, dan lain-lain. Penelitian

lainnya yang sudah dilakukan lagi adalah penelitian [6], penelitian tersebut melakukan klasifikasi penyakit diabetes menggunakan metode *CFS* dan *ROS* dengan menggunakan algoritma J48 dan berbasis Adaboost. Penelitian menggunakan *Feature Selection* yaitu metode *Correlation Feature Selection* (CFS) untuk menghilangkan fitur yang kurang relevan yang diharapkan dapat mengoptimalkan tingkat akurasi dan peneliti juga menggunakan metode *Random Over Sampling* (ROS) untuk meningkatkan kelas minoritas sehingga kelas minoritas dapat dikenali, hal tersebut dilakukan dengan tujuan agar algoritma dapat bekerja dengan baik [6]. Adaboost atau *Adaptive Boosting* juga digunakan oleh peneliti untuk meningkatkan ketelitian pada klasifikasi dengan membangkitkan kombinasi dari suatu model. Hasil penelitian diperoleh hasil akurasi 92,3%.

Dari beberapa penelitian yang sudah dilakukan, beberapa memberikan saran untuk melakukan optimasi pada klasifikasi yang dilakukan pada penelitian [1] d, maka dari itu pada penelitian ini akan melakukan optimasi pada algoritma klasifikasi data mining antara lain adalah algoritma *C4.5, Naïve Bayes,* dan *K-Nearest Neighbor.* Optimasi yang dilakukan yaitu menggunakan salah satu metode Feature Selection yaitu *Backward Elimination.* Tujuan menggunakan *Feature Selection* adalah untuk mengidentifikasi atribut data yang sama pentingnya dan menghilangkan atribut yang tidak relevan [14]. *Backward Elimination* adalah salah satu metode *Feature Selection* yang dilakukan dengan cara pemilihan kedepan yaitu menguji atribut yang ada lalu menghapus atribut yang tidak relevan [15]. Pada penelitian ini akan dilakukan optimasi menggunakan metode *Backward Elimination* pada algoritma *C4.5, Naïve Bayes,* dan *juga K- Nearest Neighbor.*

Hasil penelitian ini nantinya akan membandingkan algoritma tersebut dengan menggunakan tingkat akurasi dan nilai AUC jika atau tidak dilakukan optimasi menggunakan *Backward Elimination* dan memilih algoritma terbaik untuk deteksi dini atau klasifikasi penyakit diabetes.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah dipaparkan, terdapat beberapa rumusan masalah yang dapat diangkat adalah sebagai berikut:

1. Berapa tingkat akurasi dan nilai AUC yang didapatkan algoritma *C4.5, Naïve Bayes,* dan *K-Nearest Neighbor* dengan dikombinasikan atau tidak dengan *backward elimination* ?
2. Apakah ada perbedaan yang signifikan akurasi antara sebelum dan setelah algoritma dilakukan optimasi menggunakan *backward elimination* ?
3. Apa algoritma terbaik untuk klasifikasi penyakit diabetes berdasarkan akurasi dan nilai auc yang dihasilkan ?

1.3 Batasan Masalah

Adapun batasan masalah dalam penelitian ini adalah sebagai berikut:

1. Algoritma klasifikasi yang digunakan adalah *C4.5, Naïve Bayes,* dan *K-Nearest Neighbor.*
2. *Dataset* yang digunakan adalah data pasien penyakit diabetes yang didapatkan dari situs *UCI Machine Learning Repository* dengan 520 data pasien dengan 17 atribut.

3. Metode *Feature Selection* yang digunakan adalah metode *Backward Elimination.*
4. Proses validasi dilakukan menggunakan metode *10-fold cross validation.*
5. Pengujian akan dilakukan dengan *Google Colab.*

1.4 Tujuan Penelitian

Tujuan penelitian ini adalah untuk mengetahui algoritma terbaik dari segi nilai akurasi dan nilai AUC tertinggi untuk melakukan klasifikasi atau deteksi dini penyakit diabetes dan algoritma yang dibandingkan adalah *C4.5, Naïve Bayes,* dan *K-Nearest Neighbor.*

1.5 Manfaat Penelitian

Penelitian ini bermanfaat sebagai acuan bagi peneliti atau pihak yang memiliki topik yang sama untuk mengetahui algoritma terbaik dalam hal akurasi dan nilai AUC setelah dilakukan optimasi untuk deteksi dini penyakit diabetes. Optimasi dilakukan untuk meningkatkan atau mengoptimalkan nilai akurasi dan nilai AUC dari beberapa algoritma klasifikasi. Sehingga jika nantinya terdapat peneliti ataupun pihak yang ingin membuat aplikasi deteksi dini penyakit diabetes maka diharapkan penelitian ini dapat menjadi referensi sehingga tidak perlu melakukan pemilihan atau mencari algoritma terbaik, hal tersebut dapat dilewati sehingga implementasi algoritma ke dalam aplikasi tidak memerlukan banyak waktu dan jika akurasi algoritma yang digunakan tinggi maka deteksi akan lebih akurat mengingat deteksi dini ini berkaitan dengan keselamatan atau nyawa seseorang sehingga deteksi harus mendekati akurasi yang sempurna agar tidak

terjadi kesalahan diagnosis. Hal tersebut manfaat utama dari penelitian ini, namun ada beberapa manfaat bagi pihak lainnya yaitu sebagai berikut:

1.5.1 Bagi Objek

Hasil penelitian ini diharapkan mampu bermanfaat untuk:

a. Mengatasi keterlambatan dalam deteksi dini penyakit diabetes menggunakan algoritma klasifikasi yang memiliki akurasi tinggi sehingga kesalahan dapat diminimalisir.
b. Mencegah kematian yang lebih banyak lagi dikarenakan keterlambatan diagnosis.

1.5.2 Bagi Peneliti

Hasil penelitian ini diharapkan mampu bermanfaat untuk:

a. Berpartisipasi dalam penelitian untuk mengembangkan ilmu pengetahuan khusunya dalam data mining atau lebih luas lagi yaitu ilmu IT.
b. Menambah wawasan penulis mengenai metode-metode yang belum pernah didengar sebelumnya.
c. Dapat mengimplementasikan ilmu yang sudah didapatkan di saat perkuliahan.

1.5.3 Bagi Universitas Amikom Yogyakarta

Hasil penelitian ini diharapkan mampu bermanfaat untuk:

a. Laporan karya ilmiah mahasiswa.
b. Referensi untuk mahasiswa Universitas Amikom Yogyakarta lainnya.

1.5.4 Bagi Masyarakat

Hasil penelitian ini diharapkan mampu bermanfaat untuk:

a. Sebagai referensi dalam klasifikasi penyakit diabetes di rumah sakit manapun.
b. Sebagai bahan pertimbangan dalam klasifikasi penyakit diabetes.

1.5.5 Bagi Peneliti Selanjutnya

Hasil penelitian ini diharapkan mampu bermanfaat sebagai referensi untuk penelitian dengan topik yang sama dan diharapkan menjadi acuan yang baik untuk penelitian selanjutnya.

1.6 Metodologi Penelitian

Metode penelitian yang digunakan adalah metode analisis deskriptif kuantitatif yaitu penelitian yang dilakukan akan menekankan pada data numeric atau angka yang dimana mempunyai tujuan yaitu untuk memperjelas gambaran tentang suatu keadaan berdasarkan data yang sudah ada dengan penyajian, pengumpulan, dan analisis data yang nantinya menjadi sebuah informasi yang dapat untuk menganalisis masalah.

1.6.1 Metode Pengumpulan Data

Metode pengumpulan data yang digunakan pada penelitian ini adalah menggunakan *dataset* statistik. Penggunakan dataset statistik adalah penggunaan data yang sudah tersedia. *Dataset* yang digunakan dikumpulkan oleh pihak ketiga yang mempunyai otoritas. *Dataset* untuk penelitian ini didapatkan dari situs *UCI Machine Learning Repository*.

1.6.2 Metode Analisis

Adapun untuk menganalisa data untuk pengimplementasian *data mining* ini menggunakan metodologi CRISP-DM (*Cross Industry Standart Process for Data Mining)* dengan 6 tahapan [5]:

1. *Business Understanding*

   Menentukan latar belakang masalah tujuan penelitian pada penelitian ini yaitu mengenai penyakit diabetes.

2. *Data Understanding*

   Dari dataset diabetes yang sudah dikumpulkan dari *UCI Machine Learning Repository* yaitu dataset pasien diabetes yang memiliki 520 instance dan 17 atribut. Dari atribut akan dianalisis kualitas datanya untuk lebih mengetahui bagaimana kualitas dari dataset tersebut yaitu apakah terdapat nilai *missing*. Nilai *missing* dapat dicari menggunakan *Google Colab* dengan cara *check missing value*. Hal ini dilakukan sebagai langkah awal menangani *missing value*. Setelah *check missing value* dilakukan langkah selanjutnya adalah penjabaran atribut. Penjabaran atribut dilakukan untuk mengetahui secara pasti atribut apa yang akan digunakan dan definisinya, sehingga penelitian dapat lebih terarah dan tepat tujuan.

3. *Data Preparation*

   Setelah dilakukan *data understanding* atau pemahaman data, selanjutnya dilakukan fase pengolahan data atau *data preparation,* yaitu dengan menyiapkan data awal yaitu *dataset Early stage diabetes risk*

prediction dataset yang didapatkan dari situs *UCI Machine Learning Repository*. Dari fase pemahaman data kita dapat mengetahui apakah terdapat data dengan nilai missing atau tidak. Data *missing* akan diolah menggunakan metode *data cleaning* dengan menggunakan Google Colab. *Data cleaning* dilakukan untuk menghapus baris data yang mempunyai *missing value*, sehingga *missing value* dapat teratasi. Langkah selanjutnya setelah *data cleaning* adalah transformasi data. *Transform data* dilakukan untuk mengubah bentuk data sesuai yang dibutuhkan. Dikarenakan pada *dataset* yang akan digunakan berbentuk data categorical yaitu berbentuk huruf, maka data akan diubah bentuknya menjadi data numerical, hal ini dilakukan karena untuk proses selanjutnya yaitu *modelling* dibutuhkan data numerical untuk dapat diolah. Setelah proses *data cleaning* dan *transform data* sudah dilakukan maka data siap untuk dilakukan *modelling*.

4. *Modeling*

Setelah data siap, maka penelitian dilanjutkan ke tahap *modelling*. Terdapat beberapa model yang akan dilakukan pada penelitian ini. Untuk model 1 yaitu algoritma *Decision Tree*, model 2 yaitu algoritma *Decision Tree + Backward Elimination*, model 3 yaitu algoritma k-NN, model 4 yaitu algoritma k-NN + *Backward Elimination*, model 5 yaitu algoritma *Naïve Bayes*, dan model terakhir yaitu algoritma *Naïve Bayes + Backward Elimination*. Dari semua model tersebut akan dilakukan proses validasi menggunakan metode *k-fold cross validation* dengan nilai k

sebanyak 10. Hal tersebut dilakukan untuk mendapatkan nilai akurasi yang lebih optimal.

5. *Evaluation*

   Setelah proses *modelling* akan dilakukan evaluasi dari model. Evaluasi dilakukan terhadap nilai performa akurasi dan nilai AUC. Nilai akurasi didapatkan dari *confusion matrix* dan nilai AUC didapatkan dari kurva ROC yaitu suatu grafik yang terdiri dari garis vertical dan hotizontal. Setelah evaluasi performa akan dilakukan uji *paired t-test* untuk melihat apakah ada perbedaan sebelum dan sesudah dilakukan optimasi.

6. *Deployment*

   Setelah didapatkan model yang memiliki nilai akurasi dan nilai AUC terbaik maka model tersebut akan digunakan untuk membuat laporan sederhana mengenai hasil penelitian. Hal ini ditujukan untuk sebagai bentuk informasi atau pengetahuan yang dapat digunakan untuk pengguna atau pembaca.

1.7 Sistematika Penulisan

Sistematika dalam penulisan skripsi ini adalah sebagai berikut:

BAB I PENDAHULUAN

Bab ini berisi tentang latar belakang masalah, rumusan masalah, batasan masalah, tujuan penelitian, manfaat penelitian, metodologi penelitian, dan sistematika penulisan.

BAB II LANDASAN TEORI

Bab ini berisi tentang beberapa gagasan yang sudah dilakukan pada penelitian terdahulu dan menjadi acuan dalam skripsi ini, dapat berbentuk definisi, aturan, maupun model yang berkaitan dengan penelitian ini serta teori-teori pendukung pada penelitian ini.

BAB III METODOLOGI PENELITIAN

Bab ini berisi tentang analisis mengenai masalah penelitian dan metodologi yang akan digunakan pada penelitian. Metodologi digunakan untuk memberikan gambaran terhadap penelitian yang akan dilakukan sehingga penelitian dapat terarah sesuai rencana.

BAB IV HASIL DAN PEMBAHASAN

Bab ini berisi tentang implementasi metode CRISP-DM pada algoritma *C4.5, Naïve Bayes,* dan *kNN* dan penjelasan dari hasil metode yang dilakukan.

BAB V PENUTUP

Bab ini berisi kesimpulan dan saran dari penelitian yang sudah dilakukan, dan akan dijadikan bahan untuk penelitian selanjutnya.

DAFTAR PUSTAKA

Berisi daftar pustaka yang digunakan pada penelitian sebagai acuan atau referensi penulisan skripsi.