BAB I
PENDAHULUAN

1.1 Latar Belakang
1.1.1 Identifikasi Masalah

a. General

Diabetes Mellitus merupakan penyakit yang terjadi akibat kadar glukosa di dalam darah tinggi karena tubuh tidak dapat melepaskan atau menggunakan insulin secara normal. Kadar glukosa darah sepanjang hari bervariasi, meningkat setelah makan dan kembali normal dalam waktu dua jam [1] . Glukosa darah normal cenderung meningkat secara ringan tetapi progresif setelah usia 50 tahun, terutama pada orang-orang yang tidak aktif beraktifitas. Insulin adalah hormon yang dilepaskan oleh pankreas, merupakan zat utama yang bertanggungjawab dalam mempertahankan kadar glukosa darah yang tepat. Insulin menyebabkan glukosa berpindah ke dalam sel sehingga bisa menghasilkan energi. DM terjadi jika tubuh tidak menghasilkan insulin yang cukup untuk mempertahankan glukosa darah normal atau jika sel tidak memberikan respon yang tepat terhadap insulin.

Diabetes adalah salah satu penyakit kronis dan mematikan serta penyakit mahal yang banyak diamati dibanyak negara saat ini, penyakit ini terus dan menjadi semakin meningkat pada tingkat yang sangat mengkhawatirkan [2].

Menurut Report WHO [3] saat ini ada 246 juta penderita diabetes diseluruh dunia, dan jumlah ini diperkirakan akan meningkat menjadi 380 juta pada tahun 2025, selanjutnya 3,8 juta kematian disebabkan komplikasi diabetes setiap tahunya. Dibetes menyebabkan penyakit lain/komplikasi. Komplikasi yang lebih sering terjadi dan mematikan adalah serangan jantung dan stroke, hal ini karena kadar gula mengalami kenaikan terus menerus sehingga berakibat rusaknya pembuluh darah, saraf dan srtuktur internal lainya. Indonesia menempati urutan ke enam di dunia sebagai Negara dengan jumlah penderita *Diabetes Mellitus* terbanyak setelah India, Cina, Unisoviet, Jepang dan Brasil. Pada tahun 2006 jumlah penderita *Diabetes Mellitus* di Indonesia menjadi 14 juta orang, jika peningkatan penderita *Diabetes Mellitus* (DM) pertahunnya 230.000 orang, maka bisa kita bayangkan berapa banyak jumlah penderita *Diabetes Mellitus* pada tahun 2009.

b. Spesifik

Banyak penyandang *Diabetes Mellitus* (DM) yang terdiagnosis setelah mengalami komplikasi. Padahal, apabila dilakukan diagnosis secara dini, maka penanganan bisa dilakukan lebih cepat dan komplikasi yang membahayakan dapat dihindari. Dalam perkembangan di dunia kedokteran saat ini, para peneliti dan praktisi memusatkan perhatiannya untuk mendeteksi kondisi DM dan mencegah atau menghambat berkembangnya komplikasi. Untuk mendukung hal ini dapat digunakan teknik *data mining* untuk menggali informasi yang berharga dari kumpulan informasi atau histori data diabetes.

Penelitian tentang prediksi penyakit *Diabetes Mellitus* dengan menggunakan klasifikasi *data mining* sudah pernah dilakukan, baik komparasi beberapa klasifikasi *data mining models* ataupun *improvement* terhadap klasifikasi *data mining.*

Penelitian yang dilakukan oleh [4], mengkomparasi tiga model untuk memprediksi penyakit diabetes atau prediabetes dari faktor resiko. Model yang dibandingkan adalah *logistic regression, artificial neural networks (ANNs)* dan *C4.5* untuk memprediksi diabtes atau prediabetes dengan penggunaan faktor resiko yang umum. Datasetnya berjumlah 1487 yang telah berumur 20 tahun atau lebih, pasien yang terdiri dari 2 kelompok di Guangzhou, China yaitu terdiri dari 735 pasien dikonfirmasikan telah mengidap penyakit diabetes atau prediabetes dan 752 pasien dinyatakan tidak terkena penyakit diabetes. Dataset *input* terdiri dari 1 *output variable* dan 12 atribut/variabel yaitu *gender, age, marital status, educational level, family history of diabetes, BMI, coffee drinking, physical activity, sleep duration, work stress, consumptio of fish* dan *preference for salty foods.* Pengolahan data atau statistik menggunakan *SPSS Statistical program version 13.0.* Dataset dibagi secara random menjadi 2 bagian yang terdiri dari 1031 kasus atau 70% dari dataset untuk data testing dan 456 kasus atau sekitar 30% untuk data training. Adapun hasil performa dari ketiga prediksi model dapat dilihat pada tabel di bawah ini.

Tabel 1.1 Performa tiga prediksi model terhadap penyakit diabetes

No	Predictive models	Logistic Regression	Artificial Neural Network	Decision Tree *C4.5*
1	Accuracy (%)	76.13	73.23	77.87
2	Sensitivity (%)	79.59	82.18	80.68

3	Specificity (%)	72.74	64.49	75.13

Dari perbandingan ketiga *prediction models* bahwa model *decision tree (C4.5)* adalah menghasilkan akurasi terbaik yang diikuti oleh *logistic regression model* dan *ANNs* menghasilkan akurasi yang paling rendah.

Menurut [5] tentang pemanfaatan pendekatan data mining untuk prediksi penyakit dengan membandingkan algoritma decision trees, bayesian clasifier, back propagation, neurol network, multivariate adaptive regresion splines, adaptive-network-based fuzzzy inference system, genetic algorithm, fuzzy rulebase, association rule dan k means. Penyakit yang diprediksi diantaranya adalah penyakit jantung, kanker, stroke, gangguan paru-paru, dan diabetes. Dari penelitian ini kinerja Decision Tree dan Bayesian Classification memiliki akurasi yang sama dalam prediksi penyakit jantung dengan menghasilkan 2 label yaitu sakit jantung atau tidak sakit jantung, dengan maksimum akurasi mencapai 81%. Keuntungan utama dari menggunakan classifier *Naive Bayes*ian kecepatan penggunanya dan kesederhanaan untuk menangani dataset yang berisi banyak atribut dengan cara mudah dan sederhana. Sebagaimana diketahui bahwa penyakit jantung dan diabetes banyak banyak diderita karena komplikasi dari penyakit diabetes. Literatur ini digunakan karena atribut untuk penderita penyakit diabetes dan penyakit jantung tidak jauh berbeda.

Pada penelitian [6] Pada Decision Tree untuk diagnosis Diabetes Type II dengan akurasi 78, 176% . Disini variabel yang digunakan ada 8 plus kelas yaitu *Number of times pregnant* (banyaknya kehamilan), *Plasma glucose concentration a 2 hours in an oral glucose tolerance test* (kadar glukosa dua jam setelah makan), *Diastolic blood pressure* (tekanan darah), Triceps skin fold thickness (ketebalan kulit), *2-Hour serum insulin* (insulin), *Body mass index* (berat tubuh), *Diabetes pedigree function* (riwayat diabetes dalam keluarga), *Age* (umur), dan *Class variable* (positive diabetes (1) dan negative diabetes (0). Pada penelitian ini dilakukan penanganan *handling missing value* agar mendapat akurasi yang lebih baik, dimana atribut *number of times pregnant, record* yang yang bernilai 0 (*misssing value)* tidak dihapus dengan asumsi bahwa pasien belum pernah melahirkan. *Missing value* perlu dilakukan karena penanganan *missing value* akan memperhatikan keterkaitan dengan atribut lain, sehingga dari penanganan *missing value* tersebut tidak akan kehilangan informasi yang

berharga dari dataset yang ada, selain itu dengan penangan *misiing data* yang benar akurasi yang dihasilkanpun tidak akan bias dan akurasinya siknifikan atau berkualitas.

Menurut [7] *missing data* adalah masalah yang umum dan sering terjadi di dalam dataset pada analisis statistik. Besarnya *missing data* dikelompokkan menjadi 3 kelompok yaitu: pertama, apabila data yang hilang kurang dari 1% dari jumlah dataset yang ada, maka *missing data* ini dikategorikan *considered trivial* (dianggap sepele) karena tidak akan membuat masalah pada proses *Knowledge Discovery in Database (KDD);* kedua apabila data yang hilang antara 1-5% dari total jumlah dataset, maka dikategorikan *manageable;* ketiga: apabila data yang hilang antara 5-15%, maka memerlukan metode khusus untuk menanganinya; keempat apabila data yang hilang lebih dari 15%, maka sangat besar dampaknya pada penafsiran atau interprestasi sekaligus pada hasil akurasinya. Metode penangan *missing value* juga hampir sama dengan metode peneliti yang lain yaitu, *Case Deletion (CD), Mean Imputation (MI), Median Imputation (MDI), KNN Imputation (KNNI)*. Dataset yang digunakan adalah 11 dataset dari the Machine Learning Database Repository at the University of California, Irvine. Metode *data mining* untuk *treatment* adalah *LDA* dan *KNN*. Pengaruh penanganan *missing* data lebih berpengaruh pada kualitas akurasi yang dihasilkan.

Untuk itu penelitian terkait diatas juga dapat diambil kesimpulan mengapa perlu dilakukan setrategi untuk menangani *missing value,* karena penanganan *missing value* selain berdampak pada kualitas akurasi juga akan memperhatikan keterkaitan dataset yang ada, dengan kata lain tidak akan kehilangan informasi berharga dari data set tersebut karena akan diperhatikan penanganannya dengan teknik *handling missing value.*

Selain itu pengaruh penanganan *missing value* lebih pada kualitas akurasi, karena hasil akurasi yang bias dan tidak memperlihatkan kenyataan yang sesungguhnya akan bisa dihindari. Pembiaran terhadap *missing data* bisa saja menghasilkan akurasi yang tinggi karena diperoleh dari dataset yang tidak mencerminkan kondisi yang sebenarnya.

Alasan penanganan *missing value* menggunakan *sequential methods* karena sebagaimana diketahui bahwa *sequential methods* juga disebut *preprocessing.* Pada *sequential methods* di dalamnya akan dilakukan penanganan *missing value* secara

lengkap baik dengan cara *deleting cases* dan *replacing a missing attribute value* dengan *modus, assigning* maupun *concept* [8].

1.2 Rumusan Masalah

Dari latar belakang di atas dapat diketahui rumusan maslahnya , yaitu belum diketahuinya teknik terbaik untuk penanganan *missing value* yang ada pada *sequential methods* dalam memprediksi penyakit *Diabetes Mellitus* dengan menggunakan algoritma *C4.5* dan *Naïve Bayes.*

1.3 Tujuan Penelitian

Untuk mengetahui teknik *sequential methods* terbaik untuk menangani *missing value* yang diterapkan dengan algoritma *C4.5* dengan algoritma *Naïve Bayes* untuk mendapatkan akurasi terbaik dan signifikan dengan memperhatikan keterkaitan antara atribut yang satu dengan yang lain tanpa harus kehilangan informasi yang berharga dari dataset yang ada sehingga dataset tetap mencerminkan kondisi yang sebenarnya.

1.4 Manfaat Penelitian

Manfaat teoritis untuk ilmu pengetahuan dan teknologi dari penelitian ini yaitu dapat memberikan kontribusi dalam keilmuan terutama dalam bidang *data mining* khususnya untuk penanganan *missing value* dengan melakukan komparasi *sequential methods* pada *preprocessing data* sehingga akan dihasilkan akurasi yang baik dan signifikan dan tetap mencerminkan kondisi yang sebenarnya.