BAB III
METODE PENELITIAN

Penelitian ini adalah penelitian eksperimen dengan langkah-langkah atau metode penelitian sebagai berikut:

1. Penentuan Masalah

Penentuan masalah ini diperoleh dari studi literature dari kasus penanganan data yang hilang atau *missing data* pada penyakit *Diabetes Mellitus*.

2. Penentuan *approach* atau pendekatan

Pendekatan ini dipilih berdasarkan studi literature tentang diabetes, *C4.5, Naïve Bayes* dan *handling missing value* untuk menentukan prediksi penyakit *Diabetes Mellitus* dengan menerapkan *sequential methods* untuk menangani *missing data.* Penerapan *sequential methods* ini untuk menangani masalah *missing data* dengan menerapkan semua teknik yang ada pada *sequential methodos.*

3. Penerapan *software Rapidminer* pada obyek penelitian

Tool RapidMiner digunakan untuk mengolah obyek penelitian yaitu dataset dari *Pima Indian Diabetes Data (PIDD).*

4. Evaluasi dan validasi penelitian

Pada penelitian ini akan dilakukan evaluasi dengan cara menghitung jumlah prediksi positif dan prediksi negatif dengan hasil jumlah observasi positif dan observasi negatif pada algoritma *C4.5* dan *Naïve Bayes*. Untuk selanjutnya dilakukan proses analisis data, akan dihitung tingkat akurasi menggunkan rumus *ConfusionMatrix.*

3.1 Teknik Pengumpulan Data

Teknik pengumpulan data pada penelitian ini adalah mengambil dataset dari *Pima Indian Diabetes Data* (PIDD) dari *Uci Machine Learning Repository.* Data sebanyak 768 *record* yang terdiri dari 8 atribut 1 label atau kelas, atribut dataset dan diskripsinya seperti tabel dibawah ini:

Tabel 3.1 Atribut Dataset dan diskripsinya

Atribut	Singkatan	Deskripsi	Satuan	Tipe Data
Pregnant	Pregnant	Banyaknya kehamilan	-	Numerik
Plasma-Glucose	Glucose	Kadar glukosa dua jam setelah makan	Mg/dL	Numerik
Diastolic Blood- Pressure	DBP	Tekanan darah	Mm Hg	Numerik
Tricepts Skin Fold Thickness	TSFT	Ketebalan kulit	mm	Numerik
Insulin	INS	Insulin	mu U/ml	Numerik
Body Mass Index	BMI	Berat Tubuh	Kg/m2	Numerik
Diabetes pedigree function	DPF	Riwayat Keturunan yang terkena diabetes	-	Numerik
Age	Age	Umur	Years	Numerik
Class variable	Class	Positif diabetes (1) dan negatif diabetes (0)	-	Nominal

Dataset original dari *Uci Machine Learning Repository* dapat dilihat pada Lampiran 1.

3.2 Teknik Diskritisasi Data

Diskritisasi atribut ini merubah data *numeric* menjadi *nominal* yang bertujuan untuk mempermudah pengelompokan nilai berdasarkan kriteria yang telah ditetapkan dan untuk menyederhanakan permasalahan serta meningkatkan akurasi dalam proses *learning.* Parameter diskritisasi atribut dapat terlihat pada tabel di bawah ini.

Tabel 3.2 Diskritisasi Atribut Pima Indian Diabetes Datasets

No	Atribut	Diskritisasi	Source
1	*Pregnant*	low (0,1), medium (2, 3, 4, 5), high (> 6)	[12]
2	*Glucose*	Normal (< 95), Medium (> 95-140), High (> 140)	[21]
3	*DBP*	Normal (< 80), Normal to high (80- 90), High (> 90)	[22]
4	*TSFT*	Yes (0), No(1)	[22]
5	*INS*	Normal (<140), Impaired Glucose Tolerance (140-200), Diabetes (>=200)	[22]
6	*BMI*	Normal (18-25), Slightly Overweight (25-30) Obese (30-35)	[22]
7	*DPF*	low (< 0.5275), high (> 0.5275)	[22]
8	*Age*	Young (< 25), Middle Aged (25-40), Aged (> 40)	[22]

Dataset dari original setelah dilakukan diskritisasi dari *numeric* ke *nominal* akan berubah seperti pada Lampiran 2.

3.3 Teknik Penanganan *Missing Values*

Dari dataset seperti pada lampiran 1, terdapat data yang hilang atau *missing attribute values,* sejumlah pada tabel berikut ini:

Tabel 3.3 Jumlah missing value pada Pima Indian Diabetes Datasets

No.	Atribut	Jumlah Missing Value
1	*Pregnant*	111
2	*Glucose*	5
3	*DBP*	35
4	*TSFT*	227
5	*INS*	374
6	*BMI*	11
7	*DBF*	Lengkap
8	*Age*	Lengkap
9	*Class*	Lengkap

Dari teknik *handling missing value* yang ada pada *sequential methods* yang telah diuraikan pada BAB II tersebut diatas, juga akan dilakukan secara khusus terhadap beberapa atribut dibawah ini dengan cara:

1. Nilai nol pada atribut pregnant dapat diasumsikan bahwa nilai tersebut menyatakan pasien belum pernah melahirkan, maka angka nol dibiarkan saja walaupun *missing data* termasuk kelompok *manageable,* sehingga hal ini dimungkinkan sesuai kondisi sebenarnya [6].
2. Data dengan nilai nol pada atribut *glucose* tetap dibiarkan dan tidak dilakukan perlakuan khusus*,* karena data yang hilang atau *missing value* kurang dari 1% yaitu hanya 5 *record* dari 768 *instance,* maka *missing data* ini tidak akan bermasalah pada proses *Knowledge Discovery in Database (KDD)* [7].
3. Karena atribut TSFT dan INS memiliki jumlah nilai yang tidak ada sangat besar, maka kedua atribut ini tidak mungkin dihilangkan dan tidak mungkin dipakai dalam pengklasifikasian. Oleh karena itu, dalam penelitian ini atribut TSFT dan INS tidak diikutkan pada proses klasifikasi atau pengujian [7].

Untuk menangani data yang hilang tersebut di atas akan dilakukan perlakuan atau *treatment* satu persatu dengan teknik yang ada pada *sequential methods,* kemudian dikomparasi. Dataset yang ada pada lampiran 2 tersebut akan dikomparasi dengan 5 teknik penanganan *missing value* yang ada pada *sequential methods.* Untuk kedua teknik yang ada pada *sequential methods* yaitu *Replacing Missing Attribute Values by the Attribute Mean* dan *Replacing Missing Attribute Values by the Attribute Mean Restricted to a Concep* tidak digunakan karena teknik ini digunakan pada dataset *numeric* karena sudah terwakili oleh teknik *The Most Common Value of an Attribute dan The most Common Value of an Attribute Restrcted to a Concept*. Di bawah ini dataset setelah dilakukan *handling missing values* dengan kelima teknik yang ada pada *sequential methods.* Yang dilakukan *handling missing value* terbatas pada atribut *DBP* dan *BMI* karena jumlah data yang hilang sesuai dengan *rule* yang ada [7].

1. *Deleting Cases with Missing Attribute Value*

Atribut *DBP* dan *BMI* dari dataset yang sudah didiskritisasi yang mempunyai *missing data, record* tersebut di hapus dari atribut.

2. *The Most Common Value of an Attribute*

Dataset dari lampiran 2 atau yang sudah didiskritisasi dilakukan *handling missing values* dengan teknik *The Most Common Value of an Attribute,* ini tekniknya *missing data* yang ada pada atribut *DBP* dan *BMI* akan diisi nilai yang paling banyak sering muncul.

3. *The Most Common Value of an Attribute Restricted to a Concept*

Teknik ini mengisi *record* yang *missing data* dengan nilai yang sering muncul tetapi khusus *record* yang mempunyai kelas yang sama. Artinya *missing data* yang mempunya kelas yang sama dikumpulkan kemudian dicari nilai yang paling banyak, setelah itu nilai yang paling banyak muncul tersebut untuk mengisi *record* yang *missing data* tadi atau berkelas sama.

4. *Assigning All Possible Attribute Values to a Missing Attribute Value*

Pada teknik ini *missing data* diisi dengan nilai yang ada pada atribut yang bersangkutan, misalnya satu atribut mempunyai 3 nilai, maka *record* yang *missing data* tadi diisi 3 nilai yang ada tadi.

5. *Assigning All Possible Attribute Values Restricted to a Concept*

Teknik ini sama dengan teknik nomor 4 di atas, tetapi bedanya diisi berdasarkan kelas yang sama dari *record* yang *missing data* tadi. Artinya jumlah *record* yang *missing data* dikumpulkan yang sasa kelasnya, kemudian yang sudah terkumpul tadi bisa dilihat ada berapa nilai yang ada, kemudian *record* yang *hilang* tadi diisi dengan semua nilai yang ada tadi khususnya untuk yang mempunyai kelas yang sama.

3.4 Pengaruh handling missing value pada algoritma *C4.5* dan *Naïve Bayes* terhadap prediksi penyakit *Diabetes Mellitus*

Setelah dilakukan teknik *handling missing value* dengan *sequential methods,* secara langsung dataset original akan berubah sesuai dengan perlakuan atau (*treatment)* dari masing-masing teknik yang ada pada *sequential methods.*

Adapun teknik perlakuan dataset setelah *handling missing value* adalah sebagai berikut:

1. Dataset original diskritisasi atribut dan setelah *Deleting Cases with Attribute Values* menjadi input dari algoritma *C4.5* dan *Naïve Bayes,* dilakukan evaluasi dengan *cross validation* kemudian akan diketahui akurasi dan AUC.
2. Dataset original setelah diskritisasi atribut dan setelah dilakukan *handling missing values* dengan teknik *The Most Common Value of an Attribute* menjadi input dari algoritma *C4.5* dan *Naïve Bayes,* dilakukan evaluasi dengan *cross validation* kemudian akan diketahui akurasi dan AUC.
3. Dataset original setelah diskritisasi atribut dan setelah dilakukan *handling missing values* dengan teknik *The Most Common Value of an Attribute Restricted to a Concept* menjadi input dari algoritma *C4.5* dan *Naïve Bayes,* dilakukan evaluasi dengan *cross validation* kemudian akan diketahui akurasi dan AUC.
4. Dataset original setelah diskritisasi atribut dan setelah dilakukan *handling missing values* dengan teknik *Assigning All Possible Attribute Values to a Missing Attribute Value* menjadi input dari algoritma *C4.5* dan *Naïve Bayes,* dilakukan evaluasi dengan *cross validation* kemudian akan diketahui akurasi dan AUC.
5. Dataset original setelah diskritisasi atribut dan setelah dilakukan *handling missing values* dengan teknik *Assigning All Possible Attribute Values Restricted to a Concept* menjadi input dari algoritma *C4.5* dan *Naïve Bayes,* dilakukan evaluasi dengan *cross validation* kemudian akan diketahui akurasi dan AUC.

Dataset yang telah dilakukan perlakuan *handling missing value* akan menjadi input pada algoritma *C4.5* dan *Naïve Bayes* untuk diolah dengan *RapidMiner* untuk menghasilkan akurasi. Oleh karena perubahan dataset setelah mendapat perlakuan *handling missing value* secara langsung akan berpengaruh pada hasil akurasi dari algoritma *C4.5* dan *Naïve Bayes* sesuai dengan input datasetnya.

3.4.1 Algoritma *C4.5*

Rumus algoritma *C4.5* untuk menghitung *GainRatio* sebagai penentu akar pohon.

$$GainRatio\ (A) = \frac{Gain\ (A)}{SplitEntropy\ (A)}$$

GainRatio digunakan untuk pengukuran seleksi atribut. Artinya *GainRatio* ini digunakan untuk menentukan simpul akar. *GainRatio* tertinggi dari atribut akan menjadi akar dari pohon keputusan atau *decision tree.*

Contoh penerapan atau perhitungan manual telah dijelaskan di BAB II.

3.4.2 Algoritma Naïve Bayes

Rumus algoritma *Naïve Bayes* sebagai penentu propability class X

$$P(X|H) = P(H|X)P(X)$$

Rumus ini artinya bahwa probabilitas *X* berdasarkan kondisi pada hipotesis *H* untuk menghitung probabilitas hipotesis *H* berdasarkan kondisi *X* (*posteriori probability*) dengan probabilitas dari *X*

Contoh penerapan atau perhitungan studi kasus telah dijelaskan pada BAB II.

3.5 Teknik Pengujian atau Evaluasi dan Validasi Hasil Penelitian

Teknik evaluasi dan validasi dengan *Cross Validation.* Sedangkan *Cross Validation* sendiri dilakukan dengan cara atau istilah *10 FoldCross Validation,* ini artinya validasi akan mengulang sebanyak 10 kali dan hasil pengukuran adalah nilai rata-rata dari 10 kali pengujian tersebut.

Hasil dari berbagai percobaan dan pembuktian teoristik, menunjukkan bahwa *10 FoldCross Validation* adalah pilihan terbaik untuk mendapatkan hasil validasi yang akurat.

Pada tahap ini akan dilakukan evaluasi *cross validation* terhadap dataset dari *Pima Indian Diabetes Data (PIDD)* yang telah mendapat perlakuan *handling missing value* yang menghasilkan *ConfusionMatrix* dari *PerformanceVector,* adapun untuk menghitung validasi hasil penelitian dengan cara menghitung jumlah prediksi positif dan prediksi negatif dengan hasil jumlah observasi positif dan observasi negatif, dengan rumus sebagai berikut:

1. *Accuracy* adalah proporsi jumlah prediksi yang benar, seperti pada rumus (8).
2. *Sensitivity* juga dapat dikatakan *true positive rate* (TP *rate*) atau *recall*. Sebuah *sensitivity* 100% berarti bahwa pengklasifikasian mengakui sebuah kasus yang diamati positif, seperti pada rumus (9).

3. *Specificity* adalah mengukur atau mengamati kasus bahwa yang diamati teridentifasi negatif, seperti pada rumus (10).
4. *PPV (*nilai prediktif positif) adalah tingkat positif salah (FP) adalah proporsi kasus negatif yang salah diklasifikasikan sebagai positif, seperti pada rumus (11).
5. *NPV (*Nilai prediktif Negatif) adalah tingkat negatif sejati (TN) didefinisikan sebagai proporsi kasus negatif yang diklasifikasikan dengan benar, seperti pada rumus (12).

Dari hasil evaluasi tersebut kemudian dilakukan validasi hasil penelitian atau analisis data. Sehingga akan diketahui akurasi tertinggi dan masuk kategori apa, seperti kelompok hasil analisis menurut [17].