BAB II
TINJAUAN PUSTAKA

2.1 Kajian Literatur

Alam, T. M., et.al., (2019) dalam penelitiannya yang berjudul *A model for early prediction of diabetes*, penelitian ini bertujuan untuk menganalisis penyakit diabetes menggunakan beberapa metode data mining diantaranya Artificial neural network (ANN), random forest (RF), dan K-means clustering, dataset yang digunakan berasal dari UCI ML Repository menggunakan 9 atribut yaitu *Pregnancies, Glucose (mg/dl), Blood Pressure (mmHg), Skin Thickness (mm), Insulin (mu U/mL), BMI (kg/m2), Diabetes Pedigree Function, Age, dan outcome* sebagai atribut *class* . Tahap awal dilakukan *Preprocessing* data menggunakan teknik Normalisasi. Kemudian hasil akurasi yang diperoleh yaitu menggunakan Artificial neural network (ANN) sebesar 75,7%, *random forest (RF)* sebesar 74,7% dan *K-means clustering* sebesar 73,6%[9].

P. B. K.Chowdary., and R. U. Kumar., (2021) dalam penelitiannya yang berjudul *An Enhanced Naïve bayes Classification Algorithm to Predict Type II Diabetes*, meggunakan dataset yang berasal dari Pima Indian Diabetes (PID) dengan jumlah atribut sebanyak 9, diantaranya yaitu *Pregnancy, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Pedigree function, Age, dan outcome* sebagai atribut *class* . Tahap awal dilakukan *Preprocessing* data menggunakan teknik *Multivariate imputation*, kemudian dilakukan split data untuk membagi data training dan data testing. Metode yang digunakan dalam penelitian ini menggunakan algoritma *naïve bayes*. Hasil akurasi yang didapat yaitu sebesar 83,33%, *Precision* sebesar 0,86 dan *Recall* sebesar 0,86[10].

Anitha, j., and Pethalakshmi, A., (2017) dalam penelitiannya yang berjudul *Comparison of Classification Algorithms in Diabetic Dataset,* penelitian ini bertujuan untuk membandingkan dua algoritma yaitu *Naïve bayes* dan *Decision Tree* (J48). Dataset yang digunakan pada penelitian ini berasal dari *UCI ML Repository* menggunakan 8 atribut yaitu *Patient_nbr, gender, Age, number_diagNoses, max_glu_serum, A1Cresult, insulin.* Eksperimen menggunakan tool WEKA. Hasi akurasi yang didapat yaitu algoritma *naïve bayes* sebesar 77,2% dan *Decision Tree* (J48) sebesar 79,6%[11].

Westari, D., and Halim, A., (2021) dalam penelitiannya yang berjudul *Performa Comparison of the K-Means Method for Classification in Diabetes Patients Using Two Normalization Methods*, tujuan dari penelitian ini yaitu membandingkan metode k-mean menggunakan dua metode Normalisasi. Pertama menggunakan metode min-max Normalisasi dan kedua menggunakan Z-score Normalisasi. Hasil menunjukkan metode min-max lebih tinggi akurasinya yaitu sebesar 79% dibandingkan metode Z-score yang memperoleh akurasi sebesar 67%[12].

Sisodia, D., and Sisodia, D.,S., (2018) dalam penelitiannya yang berjudul *Prediction of Diabetes using Classification Algorithms*, tujuan dari penelitian ini yaitu membandingkan algoritma klasifikasi *Naive bayes*, SVM, dan *Decision Tree*. Dataset diperoleh dari *Pima Indians Diabetes Database* dengan jumlah data pasien diabetes sebanyak 768 pasien dan atribut sebanyak 9 yaitu *pregnant, Plasma glucose concentration, Diastolic blood pressure (mm Hg), Skin fold thickness (mm), 2-Hour serum insulin (mu U/ml), BMI (weightinkg/(heightinm), Diabetes pedigree function, Age in years, Class* . Pada penelitian ini eksperimen dilakukan menggunakan tool WEKA. Dari hasil eksperimen yang dilakukan akurasi yang diperoleh algoritma *naïve bayes* sebesar 76,30% dengan nilai *Precision* 0,759 dan *Recall* 0,763. Untuk algoritma SVM diperoleh akurasi sebesar 65,10% dengan nilai *Precision* 0,424 dan *Recall* 0,651. Sedangkan untuk algoritma *Decision Tree* diperoleh akurasi sebesar 73,82 dengan nilai *Precision* 0,735 dan *Recall* 0,738[13].

Sunge, A. S., et.al., (2019) dalam penelitiannya yang berjudul *Prediction Diabetes Mellitus Using Decision Tree Models*, pada penelitian ini dilakukan klasifikasi data mining menggunakan algoritma C4.5 untuk memprediksi penyakit diabetes. Dataset yang digunakan berasal dari Pima Indians Diabetes Dataset dengan jumlah record 768 data dengan jumlah atribut sebanyak 9 atribut : *Pregnancy, Glucose, Blood Pressure, Skin thickness, Insulin, Body Mass Index, Descent, Ages*, dan Diagnosis sebagai atribut *class* . Analisis meliputi tahap data *collection*, data *preprocessing, proposed method, testing algorithm C4.5* dan *test result validation. Preprocessing* data dilakukan dengan menangani nilai yang hilang dengan mengambil nilai rata-rata setiap atrbut. Pembagian dataset menggunakan teknik split data dengan pembagian 80% data training dan 20% data testing. Dari hasi eksperimen menggunakan tool rapidminer diperoleh hasil akurasi

sebesar 72, 08%, dengan nilai *precision* 0,70, nilai *recall* sebesar 0,97, dan niali AUC sebesar 0,718[6].

Faruque, F., et.al., (2019) dalam penelitiannya yang berjudul *Performance Analysis of Machine Learning Techniques to Predict Diabetes Mellitus*, tujuan dari penelitian ini yaitu membadingkan 4 algoritma data mining SVM, *Naïve bayes*, KNN, C4.5. Dataset yang digunakan berasal dari *diagnostic of Medical Centre Chittagong (MCC)*, Bangladesh dengan jumlah record 200 pasien diabetes dengan jumlah atribut sebanyak 17 yaitu *Age (Years), Sex, Weight, Diet, Polyuria , Water Consumption, Excessive Thirst, Blood Pressure (mmHg), Hyper Tension, Tiredness, Problem in Vision, Kidney Problem, Hearing Loss, Itchy Skin, Genetic, Diabetic*, dan *class* sebagai atribut target. *Preprocessing* data menggunakan tranformasi data dengan megklasifikasikan menjadi kategori sesuai atribut. Pembagian dataset menggunakan teknik cross validasi sebanyak 10-fold. Dari hasil eksperimen yang dilakukan algoritma C4.5 memiliki tingkat akurasi paling tinggi diantara algoritma yang lainnya yaitu memperoleh akurasi sebesar 73,5%[14].

Posonia, A. M., Vigneshwari, S., & Rani, D. J.,(2020) dalam penelitiannya yang berjudul *Machine Learning based Diabetes Prediction using Decision Tree J48*, tujuan dari penelitian ini yaitu mengklasifikasikan algoritma *Decision Tree* (J48) menggunakan dataset dari *Pima Indians Diabetes Database* dengan jumlah record 768 data dan menggunakan 9 atribut yaitu Preg, Plas, Pres, Skin, Insu, Mass, Pedi, *Age*, dan *Class* sebagai atribut target. Hasil eksperimen yang dilakukan dalam penelitian ini diperoleh hasil akurasi sebesar 91,2% menggnakan algortima *Decision Tree* J48[15].

Noviandi (2018) melakukan penelitian yang berjudul Implementasi Algoritma Decision Tree C4.5 untuk Prediksi Penyakit Diabetes. Tujuan dari penelitian ini adalah menggunakan algoritma C4.5 untuk membangun model prediksi kemungkinan penderita diabetes dan mengkonfirmasi keakuratan model yang dihasilkan. Model prediktif dibangun menggunakan data dari *Pima Indians Diabetes Databases* (PPID) di *UCI Machine Learning Repository*. Model prediksi menggunakan algoritma Decision Tree C4.5 mencapai akurasi 70,32% dengan menghasilkan 9 aturan dengan 4 aturan dan bukan 5 kelas aturan. Hasil rule yang

dihasilkan dapat digunakan untuk merancang aplikasi pendeteksi diabetes berbasis android[16].

Putri, S.U., Irawan, E., dan Rizky, F. (2021) melakukan penelitian yang berjudul "Implementasi Data Mining Untuk Prediksi Penyakit Diabetes Menggunakan Algoritma C4.5". Tujuan dari penelitian ini adalah membangun model prediksi menggunakan algoritma data mining C4.5. Algoritma ini menghasilkan pohon keputusan, sehingga pencegahan diabetes dapat diterapkan secepat mungkin. Penelitian ini memiliki beberapa atribut klasifikasi seperti berat badan, usia, tekanan darah, detak jantung, dan gula darah. Hasil penelitian ini digunakan sebagai acuan untuk menentukan apakah seseorang berisiko terkena diabetes berdasarkan karakteristik tertentu. Menerapkan algoritma C4.5 ke perangkat lunak Rapidminner menghasilkan akurasi sistem sebesar 90,00%. Ini berarti bahwa aturan yang dihasilkan mendekati 100% benar.[17].

Fatmawati., (2016) dalam penelitiannya yang berjudul Perbandingan Algoritma Klasifikasi Data Mining Model C4.5 dan *Naïve bayes* untuk Prediksi Penyakit Diabetes. Penelitian ini bertujuan membandingkan algoritma C4.5 dan *naïve bayes* untuk mengetahui algoritma yang memiliki akurasi yang lebih tinggi dalam mendeteksi penyakit diabetes. Dataset yang digunakan berasal dari *UCI Machine Learning Repossitory* dengan jumlah record sebanyak 768 data dan terdiri dari 9 atribut yaitu Jumlah Hamil, Konsentrasi Glukosa, Tekanan Darah, Lipatan Kulit, Serum Insulin, IMB, Riwayat Diabetes, Umur, dan Hasil sebagai atribut target. Analisis yang dilakukan dalam penelitian ini yaitu pengolahan awal data, pengukuran penelitian, analisis hasil komparasi. Pembagian data training dan data testing menggunakan teknik 10-fold cross validasi. Dari hasil komparasi yang dilakukan menunjukkan hasil akurasi yang diperoleh algortima C4.5 sebesar 73,30% dengan nilai AUC 0,733, sedangkan algoritma *naïve bayes* memperoleh hasil akurasi sebesar 75,13% dengan nilai AUC 0,810[18].

Andriani,. A (2013) dalam penelitiannya yang berjudul Sistem Prediksi Diabetes Berbasis Pohon Keputusan. Tujuan dari penelitian ini adalah untuk mengklasifikasikan data diabetes dan menerapkannya pada pengembangan sistem

prediksi diabetes. Hasil klasifikasi data diabetes dievaluasi dengan menggunakan confusion matrix dan kurva receiver operating character (ROC) untuk mengetahui keakuratan hasil klasifikasi. Hasil pengujian sistem menggunakan *confusion matrix* adalah mendapatkan nilai akurasi sebesar 73,33% dan nilai kurva ROC adalah sebesar 0,815, sehingga dikategorikan klasifikasi yang baik.[19].

Ridwan (2020) dalam penelitiannya yang berjudul Penerapan Algoritma *Naïve bayes* Untuk Klasifikasi Penyakit Diabetes Mellitus. Penelitian ini bertujuan untuk menganalisis penyakit diabetes dengan mengklasifikasikan gejala awal penyakit menggunakan metode *naïve bayes*. Sehingga dari hasil evaluasi menggunakan metode *naïve bayes* didapatkan nilai akurasinya. Data yang digunakan dalam analisis ini yaitu berasal dari dataset *machine learning UCI*, dengan judul dataset risiko diabetes tahap awal dan disumbangkan pada tahun 2020, dataset terdiri dari 17 atribut. Analisis yang dilakukan meliputi preprocessing data, model, dan evaluasi. Metode klasifikasi menggunakan algoritma *naive Bayes*. Hasil klasifikasi menunjukkan akurasi 90,20% dan nilai AUC 0,95[5].

Tabel 2.1 penelitian terkait tentang prediksi diabetes menggunakan data mining

Judul, Penulis, Tahun	Jumlah & Atribut	Algoritma	*Preprocessing*	*Feature selection*	Validasi	Open Source dataset	Akurasi
A model for early prediction of diabetes (Alam, T. M., et.al, 2019)	ada 9 atribut : *Pregnancies, Glucose (mg/dl), Blood Pressure (mmHg), Skin Thickness (mm), Insulin (mu U/mL), BMI (kg/m2), Diabetes Pedigree Function, Age,*	Artificial neural network (ANN), random forest (RF), and K-means clustering	Normalisasi	-	-	*UCI ML Repository*	Artificial neural network (ANN) = 75.7%

Judul, Penulis, Tahun	Jumlah & Atribut	Algoritma	*Preprocessing*	*Feature selection*	Validasi	Open Source dataset	Akurasi
	outcome						
An Enhanced Naïve bayes Classification Algorithm to Predict Type II Diabetes (P. B. K.Chowdary and R. U. Kumar, 2021)	ada 9 atribut : *Pregnancy, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Pedigree function, Age, outcome*	*Naïve bayes classifier*	Teknik Multivariate imputation	-	-	PID dataset	83.33%
Comparison of Classification Algorithms in Diabetic Dataset (Anitha, j. and Pethalakshmi, A, 2017)	ada 8 atribut : *Patient_nbr, gender, Age, number_diagnoses, max_glu_serum, A1Cresult, insulin*	*Naïve bayes*ian dan *Decision Tree* (J48)	-	-	-	*UCI ML Repository*	NB = 77.2 %, J48=79.6 %
Perform a Comparison of the K-Means Method for Classification in	ada 9 atribut : *Pregnant, Plasma Glucose Concentration, Diastolic Blood Pressure , Triceps*	K-Means Clustering	Min-max Normalisasi dan Z-score Normalisasi	-	-	PID dataset	Min-max 79%

Judul, Penulis, Tahun	Jumlah & Atribut	Algoritma	*Preprocessing*	*Feature selection*	Validasi	Open Source dataset	Akurasi
Diabetes Patients Using Two Normalization Methods (Westari, D., and Halim, A., 2021)	*Skin Fold Thickness, 2-Hour Serum Insulin , Body mass index, Diabetes Pedigree Function, Age, class*						
Prediction of Diabetes using Classification Algorithms (Sisodia, D., and Sisodia, D.,S., 2018)	ada 9 atribut : Number of times pregnant, Plasma glucose concentration, Diastolic blood pressure (mm Hg), Skin fold thickness (mm), 2-Hour serum insulin (mu U/ml), BMI (weightinkg/(heightinm), Diabetes pedigree function, *Age* in years, *Class*	*Naive bayes*, SVM, *Decision Tree*	-	-	-	PIDD-Pima Indians Diabetes Dataset	*Naïve bayes*=76.30
Prediction	ada 9 atribut :	*Decision Tree*	*Cleaning Data* dengan	-	*split validatio*	PIDD-*Pima*	72.08%

Judul, Penulis, Tahun	Jumlah & Atribut	Algoritma	*Preprocessing*	*Feature selection*	Validasi	Open Source dataset	Akurasi
Diabetes Mellitus Using Decision Tree Models (Sunge, A. S., et.al., 2019)	*Pregnancy, Glucose, Blood Pressure, Skin thickness, Insulin, Body Mass Index, Descent, Ages, Diagnosis*		teknik *Replace missing value*		*n*	*Indians Diabetes Dataset*	
Performance Analysis of Machine Learning Techniques to Predict Diabetes Mellitus (Faruque, F., et.al., 2019)	ad 17 atribut : *Age (Years), Sex, Weight, Diet, Polyuria , Water Consumption, Excessive Thirst, Blood Pressure (mmHg), Hyper Tension, Tiredness, Problem in Vision, Kidney Problem, Hearing Loss, Itchy Skin, Genetic, Diabetic, class*	SVM, *Naïve bayes*, KNN, C4.5	*Transformation data*	-	10-fold cross validasi	t*he diagNostic of Medical Centre Chittagong (MCC), Bangladesh*	C4.5=73.5%
Machine Learning based Diabetes	ada 9 atribut : *Preg, Plas, Pres, Skin, Insu, Mass, Pedi, Age,*	*Decision Tree* J48	-	-	-	*Pima Diabetes Database (PIDD)*	91.20%

Judul, Penulis, Tahun	Jumlah & Atribut	Algoritma	*Preprocessing*	*Feature selection*	Validasi	Open Source dataset	Akurasi
Prediction using Decision Tree J48 (Posonia, A. M., Vigneshwari, S., & Rani, D. J., 2020)	*Class*						
Implementasi Algoritma *Decision Tree* C4.5 untuk Prediksi Penyakit Diabetes (*No*viandi, 2018)	ada 7 atribut : Jumlah Wanita Melahirkan, Kadar Gula Darah, Tekanan Darah, Insulin, Body Mash Index, Usia, Outcome	Algoritma *Decision Tree* C4.5	-	-	-	*Pima Indians Diabetes Dataset (PPID)*	70.32%
Implementasi Data Mining Untuk Prediksi Penyakit Diabetes Dengan Algoritma C4.5 (Putri, S. U., Irawan, E., dan Rizky,	ada 6 atribut : Usia (tahun), Tekanan darah (NmHg), Berat badan (Kg), Kadar gula darah (mg/dl), variabel	Algoritma C4.5	-	-	-	RSUD. Dr. Djasamen Saragih Pematangsiantar	90,00%,

Judul, Penulis, Tahun	Jumlah & Atribut	Algoritma	*Preprocessing*	*Feature selection*	Validasi	Open Source dataset	Akurasi
F., 2021)							
Sistem Prediksi Penyakit Diabetes Berbasis *Decision Tree* (Andriani,. A., 2013)	Tidak disebutkan	*Decision Tree*	-	-	-	*UCI Machine Learning Repository*	73.33%
Perbandingan Algoritma Klasifikasi Data Mining Model C4.5 dan *Naïve bayes* untuk Prediksi Penyakit Diabetes (Fatmawati., 2016)	Ada 9 atribut : Jumlah Hamil, Konsentrasi Glukosa, Tekanan Darah, Lipatan Kulit, Serum Insulin, IMB, Riwayat Diabetes, Umur, Hasil	*Naïve bayes dan C4.5*	-	-	10-fold cross validasi	*UCI Machine Learning Repository*	NB = 75,13%, C4.5= 73,30%
Penerapan Algoritma *Naïve bayes* Untuk Klasifikasi Penyakit	ada 17 atribut : *Age, Sex, Polyuria , Polydipsia , Sudden weight loss, Weakness , Polyphagia , Genital thrush,*	*Naïve bayes*		-		UCI Machine Learning Repository	90.20%

Judul, Penulis, Tahun	Jumlah & Atribut	Algoritma	*Preprocessing*	*Feature selection*	Validasi	Open Source dataset	Akurasi
Diabetes Mellitus (Ridwan, A., 2020)	*Visual blurring, Itching , Irritability, Delayed healing, Partial paresis, Muscle stiffness, Alopecia, Obesity, Class*						
Optimasi Algoritma *Naïve bayes* Menggunakan *Feature selection* untuk Prediksi Penyakit Diabetes (Diki, 2022)	ada 17 atribut : *Age, Sex, Polyuria , Polydipsia , Sudden weight loss, Weakness , Polyphagia , Genital thrush, Visual blurring, Itching , Irritability, Delayed healing, Partial paresis, Muscle stiffness, Alopecia, Obesity, Class*	Algoritma *Naïve bayes*	Cleaning Data menggunakan teknik *filter examples*	*Forward selection*	10-fold Cross validasi	*UCI Machine Learning Repository*	*Naïve bayes* = ?

Dari review beberapa jurnal baik jurnal internasional maupun nasional pada tabel diatas belum menerapkan metode *feature selection*. Sehingga peneliti ingin menerapkan *feature selection* menggunakan metode *forward selection* pada algoritma *naïve bayes*. Merujuk pada jurnal yang berjudul "Penerapan Algoritma *Naïve bayes* Untuk Klasifikasi Penyakit Diabetes" memiliki kesamaan yaitu

mengenai dataset dan algoritma yang digunakan, sehingga dapat digunakan sebagai pembanding dengan peneliti yang menambahkan *feature selection* untuk meningkatkan nilai akurasi dari performa model.

2.2 *Preprocessing* dalam Data Mining

Data preprocessing adalah teknik data mining yang digunakan untuk mengubah data mentah dalam format yang berguna dan efisien. Berikut adalah langkah-langkah preprocessing data.

1. ***Data Cleaning***

   Data dapat memiliki banyak bagian yang tidak relevan dan hilang. Untuk menangani bagian ini dilakukan pembersihan data. Ini melibatkan penanganan data yang hilang, data yang berisik, dan lain-lain.

   A. Data Hilang

   Situasi ini muncul ketika beberapa data hilang dalam data. Itu bisa ditangani dengan berbagai cara.

   - **Abaikan tupel:**
     Pendekatan ini hanya cocok bila kumpulan data yang kita miliki cukup besar dan beberapa nilai hilang dalam sebuah tupel.
   - **Isi Nilai yang Hilang:**
     Ada berbagai cara untuk melakukan tugas ini. Anda dapat memilih untuk mengisi nilai yang hilang secara manual, menurut rata-rata atribut atau nilai yang paling mungkin.

   B. Data Bising

   Data bising adalah data tidak berarti yang tidak dapat diinterpretasikan oleh mesin. Data tersebut dapat dihasilkan karena pengumpulan data yang salah, kesalahan entri data, dan lain-lain. Hal ini dapat ditangani dengan cara berikut:

- **Metode Binning:**

  Metode ini bekerja pada data yang diurutkan untuk menghaluskannya. Seluruh data dibagi menjadi segmen dengan ukuran yang sama dan kemudian berbagai metode dilakukan untuk menyelesaikan tugas. Setiap tersegmentasi ditangani secara terpisah. Seseorang dapat mengganti semua data dalam suatu segmen dengan mean atau nilai batasnya dapat digunakan untuk menyelesaikan tugas.

- ***Regresi*:**

  Di sini data dapat diperhalus dengan menyesuaikannya dengan fungsi regresi. Regresi yang digunakan bisa linier (memiliki satu variabel bebas) atau berganda (memiliki banyak variabel bebas).

- ***Clustering*:**

  Pendekatan ini mengelompokkan data serupa dalam sebuah cluster. Pencilan mungkin tidak terdeteksi atau akan jatuh di luar cluster.

2. Data Transformation

Langkah ini dilakukan untuk mengubah data dalam bentuk yang sesuai untuk proses penambangan. Ini melibatkan cara-cara berikut:

A. Normalisasi:

Hal ini dilakukan untuk menskalakan nilai data dalam rentang tertentu (-1.0 hingga 1.0 atau 0.0 hingga 1.0)

B. Pemilihan Atribut:

Dalam strategi ini, atribut baru dibangun dari kumpulan atribut yang diberikan untuk membantu proses penambangan.

C. Diskritisasi:

Ini dilakukan untuk mengganti nilai mentah atribut numerik dengan level interval atau level konseptual.

D. Pembuatan Hirarki Konsep:

Di sini atribut diubah dari tingkat yang lebih rendah ke tingkat yang lebih tinggi dalam hierarki. Misalnya Atribut "kota" dapat dikonversi menjadi "negara".

3. Data Reduction

Data mining digunakan untuk memproses data dalam skala yang besar. Sehingga menyulitkan dalam menganalisis kasus oleh sebab itu dapat dilakukan teknik reduksi data untuk meningkatkan efisiensi penyimpanan dan mengurangi biaya penyimpanan dan analisis data. Berbagai langkah untuk reduksi data adalah:

A. Agregasi Kubus Data:

Operasi agregasi diterapkan pada data untuk konstruksi kubus data.

B. Seleksi Subset Atribut:

Seleksi Subset Atribut diperlukan untuk memilih atribut yang relevan dan membuang atribut yang tidak relevan. Untuk melakukan pemilihan atribut, dapat menggunakan tingkat signifikansi dan nilai p dari atribut yang ada. Atribut yang memiliki nilai p lebih besar dari tingkat signifikansi dapat dibuang.

C. Pengurangan Numerositas:

Ini memungkinkan untuk menyimpan model data alih-alih seluruh data, misalnya: Model *Regresi*.

D. Pengurangan Dimensi:

Ini mengurangi ukuran data dengan mekanisme pengkodean. Ini bisa

menjadi lossy atau lossless. Jika setelah rekonstruksi dari data terkompresi, data asli dapat diambil, pengurangan tersebut disebut pengurangan lossless atau disebut pengurangan lossy. Dua metode pengurangan dimensi yang efektif adalah: Transformasi wavelet dan PCA (Analisis Komponen Utama)[20].

2.3 Algoritma *Naïve Bayes*

Algoritma *naïve bayes* adalah salah satu algoritma klasifikasi data mining yang cukup dikenal dan sering digunakan di bidang kesehatan untuk memprediksi suatu penyakit. Algoritma ini yaitu menggunakan pengelompokkan berdasarkan probabilitas. Menurut pendapat Bramer *Naive Bayes* adalah metode tanpa aturan. *Naive Bayes* menemukan potensi terbesar dari kemungkinan klasifikasi dengan memeriksa frekuensi setiap klasifikasi data pelatihan menggunakan bidang matematika yang dikenal sebagai teori probabilitas. *Naive Bayes* adalah metode klasifikasi umum dan salah satu dari 10 algoritma teratas untuk data mining. Algoritma ini juga dikenal sebagai Idiot's Bayes, Simple Bayes, dan Independence Bayes. [21].

Klasifikasi Bayes didasarkan pada teorema Bayes, diambil dari nama seorang ahli matematika yang juga menteri Presbyterian Inggris, Thomas Bayes (1702-1761), yaitu [21]:

$$P(x|y) = \frac{P(y|x)P(x)}{P(y)} \quad \text{.....(1)}$$

$$P(y) = \sum_{n=1}^{n} P(y|x)P(x) \quad \text{.....(2)}$$

Keterangan :

Y : data dengan kelas yang belum diketahui

X : hipotesis data y merupakan suatu kelas spesifik

P(x|y) : probabilitas hipotesis x berdasarkan kondisi y (posteriori probability)

P(x) : probabilitas hipotesis x (prior probability)

P(y|x) : probabilitas y berdasarkan kondisi pada hipotesis x

p(y) : probabilitas dari y

2.4 Confusion Matrix

Confusion matrix adalah tabel yang berisi rincian klasifikasi, kelas diprediksi ditampilkan di bagian atas *matrix* dan kelas yang diamati ditampilkan di bagian sebelah kiri [22]. Evaluasi model *Confusion matrix* menggunakan tabel seperti matrix di bawah ini:

Tabel 2.2. Matrik Klasifikasi untuk Model 2 *Class*

		Kelas Prediksi	
	Klasifikasi	Kelas =*Yes*	Kelas =*No*
Kelas yang diamati	Kelas =*Yes*	(*True Positive* –TP)	(*False Negative* –FN)
	Kelas =*No*	(*False Positive* – FP)	(*True Negative* –TN)

Sumber: Gorunescu (2011)

Keterangan :

TP : Total kasus *positive* yang dikategorikan sebagai *positive*

FP : Total kasus *Negative* yang dikategorikan sebagai *positive*

TN : Total kasus *Negative* yang dikategorikan sebagai *Negative*

FN : Total kasus *positive* yang dikategorikan sebagai *Negative*

Berdasarkan nilai True Negatif (TN), False *Positive* (FP), False Negatif (FN), dan True *Positive* (TP) dapat diperoleh nilai akurasi, presisi, dan recall[23].

a. Akurasi

   Akurasi adalah tingkat kedekatan antara nilai prediksi dengan nilai aktual. Dengan rumus sebagai berikut :

   $$\text{Akurasi} = \frac{(\text{TP}+\text{TN})}{(\text{TP}+\text{TN}+\text{FP}+\text{FN})}$$

b. presisi (Precision)

Presisi adalah tingkat ketepatan antara informasi yang diminta oleh pengguna dengan jawaban yang diberikan oleh sistem. Dengan rumus sebagai berikut :

$$\text{Presisi} = \frac{\text{TP}}{(\text{TP}+\text{FP})}$$

c. Recall

Recall adalah tingkat keberhasilan sistem dalam menemukan kembali sebuah informasi. Dengan rumus sebagai berikut :

$$\text{Recall} = \frac{\text{TP}}{(\text{TP}+\text{FN})}$$

2.5 Kurva ROC (*Receiver Operating Characteristics*)

Kurva ROC adalah grafik digunakan untuk mengevaluasi hasil prediksi, kurva ROC adalah teknik untuk memvisualisasikan, mengatur, dan memilih pengklasifikasian berdasarkan kinerja mereka [22].

Kurva ROC adalah alat dua dimensi yang digunakan untuk mengevaluasi kinerja klasifikasi menggunakan dua kelas keputusan, di mana setiap objek ditugaskan ke salah satu anggota himpunan pasangan positif atau negatif. Pada kurva ROC, laju TP diplot pada sumbu Y dan laju FP diplot pada sumbu X..

Untuk klasifikasi data mining, nilai AUC dapat dibagi menjadi beberapa kelompok.

a. 0.90-1.00 = *Excellent Classification*
b. 0.80-0.90 = *Good Classification*
c. 0.70-0.80 = *Fair Classification*
d. 0.60-0.70 = *Poor Classification*
e. 0.50-0.60=*Failure*

2.6 *Cross validation*

Cross validation adalah teknik data mining yang bertujuan untuk membagi data latih dan data uji dengan kelas yang seimbang untuk memperoleh hasil akurasi yang maksimal. Metode ini sering juga disebut dengan 10-fold *cross validation* dimana percobaan sebanyak k kali untuk satu model dengan parameter yang sama [24]. Fungsinya dari penggunaan *cross validation* adalah :

a. Mengetahui performa dari suatu model algoritma dengan melakukan percobaan sebanyak k kali
b. Meningkatkan tingkat performansi dari model tersebut
c. Mengolah dataset menjadi kelas yang seimbang

Percobaan di bawah adalah contoh ilustrasi dari 5-fold *cross validation* yang artinya adalah melakukan percobaan sebanyak 5 kali tahapan.

Percobaan 1	Data Uji	Data Latih	Data Latih	Data Latih	Data Latih
Percobaan 2	Data Latih	Data Uji	Data Latih	Data Latih	Data Latih
Percobaan 3	Data Latih	Data Latih	Data Uji	Data Latih	Data Latih
Percobaan 4	Data Latih	Data Latih	Data Latih	Data Uji	Data Latih
Percobaan 5	Data Latih	Data Latih	Data Latih	Training	Data Uji

- Percobaan 1: pada bagian partisi pertama menjadi data Data Uji dan partisi lainnya menjadi data Data Latih.
- Percobaan 2: pada bagian partisi kedua menjadi data Data Uji dan partisi lainnya menjadi data Data Latih.
- Percobaan 3: pada bagian partisi ketiga menjadi data Data Uji dan partisi lainnya menjadi data Data Latih dan begitu seterusnya.

Dari hasil kelima percobaan tersebut, digunakan untuk mendapatkan nilai evaluasi kinerja model dan untuk menentukan nilai rata-rata dari setiap percobaan. Kemudian dari hasil evaluasi dapat dilihat eksperimen mana yang dipilih yang dapat digunakan sebagai referensi. Dalam beberapa penelitian yang dilakukan

oleh pakar data mining, uji model, atau validasi model algoritma klasifikasi, model validasi yang menerapkan 10 cross-validation sudah merupakan metode validasi standar dan canggih yang dapat meningkatkan nilai performa model. [25].

2.7 *Feature selection*

Seleksi fitur adalah proses menghilangkan fitur yang berlebihan dan tidak relevan dari dataset yang ada. Fitur yang tidak relevan dapat menurunkan dan mempengaruhi akurasi klasifikasi. Dengan menghilangkan fitur yang tidak relevan dapat meningkatkan nilai akurasi[26]. Selain itu seleksi fitur dapat membantu mengurangi biaya pemrosesan data. [27]. Algoritma seleksi fitur dibagi menjadi tiga: filters, wrappers, dan embedded selectors. Filters menilai setiap fitur secara independen dari pengklasifikasi, memberi peringkat fitur berdasarkan skor, dan memilih yang terbaik.[28]. Wrappers mendapatkan subset dari set fungsi, mengevaluasi kinerja classifier untuk subset itu, kemudian mengevaluasi subset lainnya oleh classifier. Subset yang memiliki nilai paling tinggi dalam klasifikasi yang dipilih. Oleh karena itu, wrappers tergantung pada pengklasifikasi yang di pilih. Bahkan wrappers lebih dapat diandalkan karena algoritma klasifikasi yang mampu mempengaruhi nilai akurasi[29]. Teknik Embedded melakukan seleksi fitur selama proses mempelajari data sama seperti yang dilakukan jaringan syaraf tiruan.

2.7.1 *Forward selection*

Forward selection adalah prosedur langkah demi langkah yang bertujuan untuk menambahkan variabel kontrol ke persamaan satu per satu. *Forward selection* dimulai dengan kumpulan fitur kosong dan menambahkan fitur yang digunakan pada putaran pertama. Semua fitur dievaluasi secara individual. Fitur ditambahkan dan dinilai kembali ke kumpulan fitur yang merupakan bagian dari fitur sebelumnya dan yang baru dibuat. Hanya subset fitur terbaik yang digunakan..

Data pelatihan dimulai secara bertahap dari satu variabel sampai pada jumlah variabel yang menghasilkan kinerja terbaik atau yang memiliki tingkat kesalahan terkecil. Misalnya, pengujian data dengan dua variabel menghasilkan tingkat kesalahan lebih kecil dan ketika diujikan lagi dengan tiga variabel dan menghasilkan tingkat kesalahan lebih besar dibandingkan dengan dua variabel maka kesalahan terkecil didapatkan pada variabel ke kedua yang berarti variabel kedua signifikan, proses dihentikan bila semua variabel sudah dilakukan pengujian[30].

2.8 Rapidminer

Rapidminer adalah perangkat lunak yang dikembangkan oleh Dr. Markus Hofmann dari Institute of Technology Blanchardstown dan Raif Klinkenberg dari rapid-i.com memiliki tampilan GUI (graphical user interface) untuk membuat perangkat lunak lebih mudah digunakan oleh pengguna. Perangkat lunak ini open source, ditulis menggunakan bahasa Java di bawah Lisensi Publik GNU, dan Rapidminer berjalan di sistem operasi apapun. Tidak diperlukan pengetahuan pemrograman khusus saat mengoperasikan Rapidminer. Rapidminer dikhusukan untuk penggunaan data mining[18].

2.9 Pengertian Diabetes Melitus

Organisasi Kesehatan Dunia (WHO) menyatakan bahwa diabetes tidak dapat didefinisikan secara ringkas dan jelas, tetapi dapat digambarkan sebagai kumpulan kompleks masalah anatomi dan kimia karena beberapa faktor defisiensi insulin absolut yang meningkat. [31].

Diabetes adalah gangguan metabolisme yang ditandai dengan hiperglikemia yang disebabkan oleh defisiensi sekresi insulin, kerja insulin, atau keduanya. Hiperglikemia kronis pada diabetes menyebabkan banyak kerusakan pada organ tubuh manusia seperti ginjal, mata, saraf, jantung dan pembuluh darah. [32].

Menurut Dorland, diabetes melitus (DM) adalah suatu kondisi yang ditandai dengan buang air kecil yang berlebihan. Selain itu, diabetes mellitus, atau dikenal dengan sebutan kencing manis adalah kondisi kronis yang ditandai dengan kadar glukosa (gula) darah di atas normal, yaitu kadar glukosa darah terkait atau di atas 200 mg / dl. Diagnosis khas DM yang umum adalah poliuria (urin berat), polidipsia (banyak minum), polifagia (diet berat), dan penurunan berat badan secara signifikan [33].