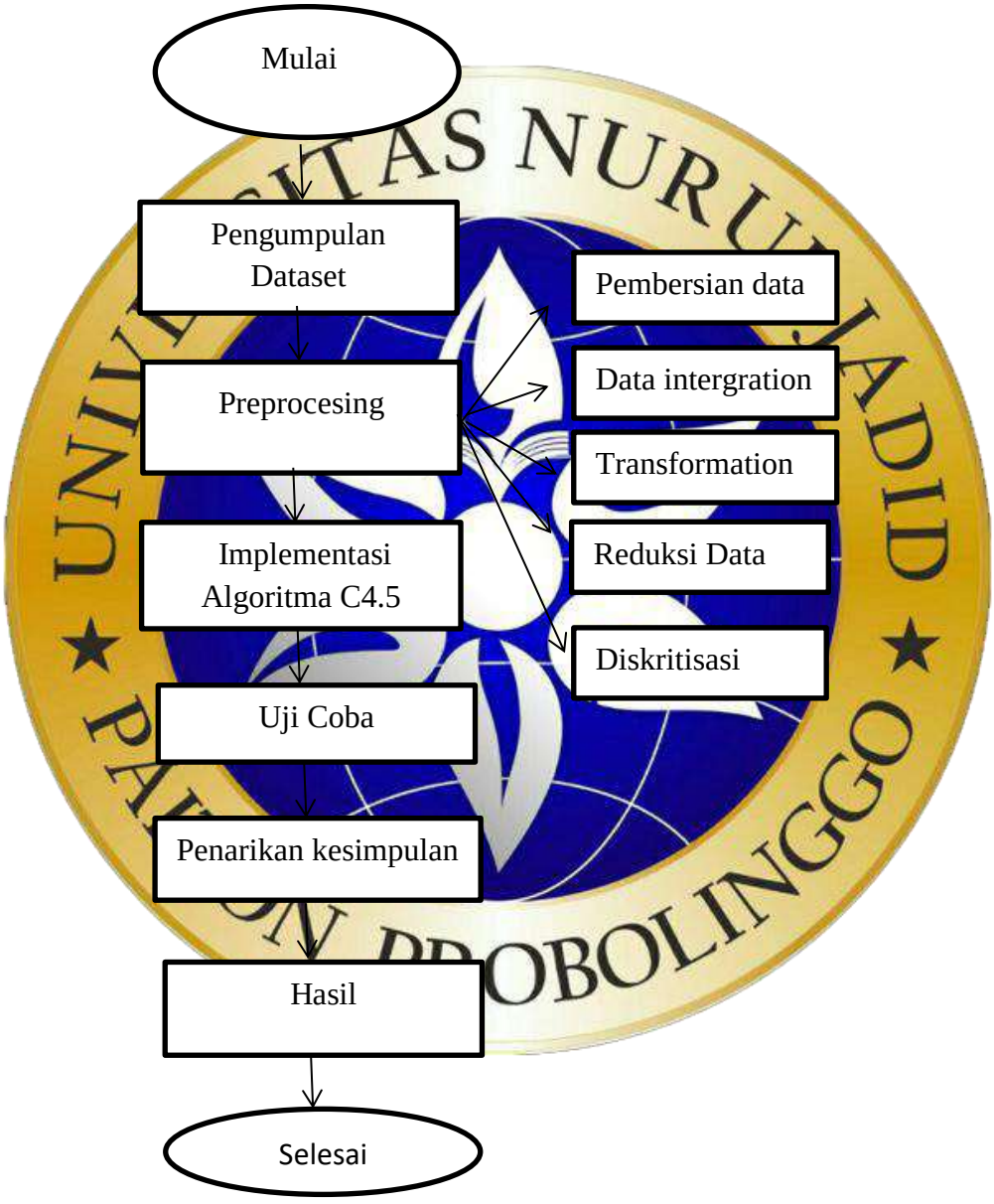
BAB III

METODE PENELITIAN

3.1 Kerangka penelitian

Kerangka penelitian ini dibuat agar memudahkan peneliti menyelesaikan proses penelitian. Terdapat beberapa tahapan yang akan di lalui.berikut alur tahapan kerangka tersebut :



Gambar 3. 1 Kerangka Penelitian

3.2 Prosedur penelitian

Pada tahap ini di uraikan setiap langkah-langkah pada kerangka rancangan peneliti berdasar kerangka gambar 2

3.2.1 Pengumpulam Data

pengumpulan data merupakan bagian yang paling penting dalam proses penelitian. Pengambilan Data di lakukan untuk mempermudah peneliti dalam melakukan penelitian. Peneliti mengambil Dataset dari MetropolitanUniversitySylhet,Bangladeshrahmansadik004 '@' gmail.com(https://archive.ics.uci.edu/ml/machinelearningdatabases/00529/diabetes_data_upload.csv) dari semua dataset di banglades terdiri dari 17 variabel yang terdiri dari Age (usia), Gender(Jenis kelamin), Polyuria (sering buang air kecil), Polydipsia (sering haus), Sudden weight los (Penurunan Berat Badan Mendadak), Wealnes (Kelemahan), Polyphagia (banyak makan), Genital thursh (Sariawan Genital), Visual blurring (nyeri pada lutut) Itching (Gatal), Irritability (lekas marah), Delayed healing (penyembuhan Tertunda), Paresis Parsial (kondisi lumpuh karena gangguan saraf), Mucles stiffness (ketidak seimbangan Kerja Otot), Alopecia (kerontokan rambut tiba-tiba), Obesitas (kegemukan), Label (tingkatan). Jumblah keseluruhan dataset tersebut sebanyak 520 data, dengan jumlah pasien positive sebanyak 320, dan jumblah pasien yang negative sebanyak 200. Agar lebih jelas lihat pada tabel 3.1

Tabel 3. 1 Status Diabetes Pasien berdasarkan Jenis Kelamin

Jenis Kelamin	Positif	Negative	Total
Perempuan	173	19	192
Laki-laki	147	181	328
Total	320	200	520

3.2.2 Processing

Preprocessing yaitu pembersihan data agar mendapatkan data yang baik dan data yang berkualitas untuk di lakukan proses data mining, proses preprocessing di bagi menjadi 4 yaitu :

A. Pembersian data

Pembersian data di gunakan untuk menghilangkan atau melengkapi data yang tidak lengkap, menghaluskan data noisy, serta memperbaiki data yang kurang konsisten.

- Contoh Data

Tabel 3. 2 Processing

Atribut

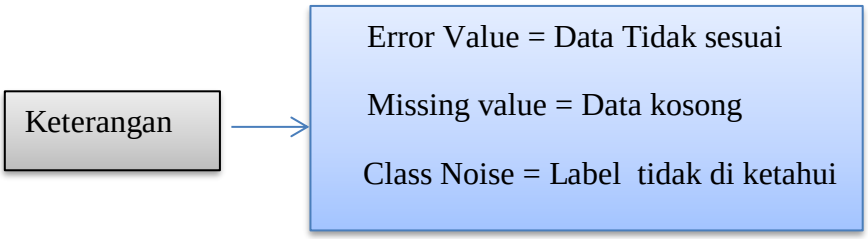
Class

Age	Gender	Polyuria	Polydipsia	Sudden Weight Loss	Label
44	Male	1	1	0	1
61	Male	1	1	1	1
43	Famale	1	1	1	1
67	Male	1	0	0	0
43	Male	0	0	1	0
30	Famale	1	0	1	1
48	Famale	0	0	1	0
40	Male	0	0	1	1
55	Famale	1	?	1	1
58	Male	=	1	0	?

Error Value

Missing value

Class



- Cara menangani data yang tidak lengkap atau kurang konsisten yaitu :
  a) Menghilangkan atau mengabaikan
  b) Mengisi dengan konstanta misalkan saja “yes” atau “no”
  c) Mengisi nilai dengan manual.
  d) Mengisi nilai yang hilang secara otomatis (baik dengan metode yang tepat).

  Contoh data yang tidak lengkap atau kurang konsisten Yaitu dengan cara memberi nilai dominan yang di miliki atribut (angka yang sering muncul).

  Tabel 3. 3 Data yang tidak lengkap

Age	Gender	Polyuria	Polydipsia	Sudden Weight Loss	Label
44	Male	1	1	0	1
61	Male	?	1	1	1
43	Famale	1	1	1	1
67	Male	1	0	0	0
43	Male	0	0	1	0
30	Famale	1	?	1	1
48	Famale	0	0	1	0
40	Male	0	0	?	1
55	Famale	1	0	1	1
58	Male	1	1	0	1

Age	Gender	Polyuria	Polydipsia	Sudden Weight Loss	Label
44	Male	1	1	0	1
61	Male	1	1	1	1
43	Famale	1	1	1	1
67	Male	1	0	0	0
43	Male	0	0	1	0
30	Famale	1	0	1	1
48	Famale	0	0	1	0
40	Male	0	0	1	1
55	Famale	1	0	1	1
58	Male	1	1	0	1

B. Data intergration

Data intergration di gunakan untuk menggabungkan data dari berbagai database ke dalam satu database baru.

C. Data Transformation

Data Transformation di gunakan untuk merubah data menjadi bentuk yang sesuai untuk proses data mining. Proses transformasi yaitu menyatukan atribut/variable yang nantinya di gunakan untuk prediksi. diskritisasi sendiri memiliki konsep hierarki yaitu mengurangi data melalui pengumpulan dan penggantian konsep level rendah (seperti nilai numerik untuk usia) dengan konsep level lebih tinggi seperti muda, middle-aged, manual dan lain-lain.

D. Reduksi Data

Reduksi data adalah untuk mendapatkan representasi yang di reduksi dalam volume tetapi menghasilkan hasil analitika yang sama atau mirip.

E. Diskritisasi Data

Diskritisasi data adalah bagian dari reduksi data tetapi dengan kepentingan khusus terutama data numeric.

3.2.3 Implementasi Algoritma Decision tree C4.5

Pada tahap ini sistem akan memproses dataset kemudian di implementasikan dengan menggukan algoritma decision tree c4.5 adalah sala satu program untuk melakukan klasifikasi. Decision tree adalah sebuah sturktur yang di gunakan untuk membagi kumpulan data yang besar menjadi himpunan record yang lebih kecil dengan menerapkan aturan pohon keputusan. Dengan mangsing - masing rangkaian pembagian anggota himpunan hasil menjadi mirip satu dengan yang lain. Simpul pohon keputusan di bagi menjadi 3 yaitu akar simpul, simpul percabangan, dan simpul akhir. Pohon keputusan dengan algoritma C4.5 sangat berhubungan karena pada dasarnya algoritma adalah pohon keputusan. Algoritma ini mempunyai inputan berupa data atribut, data training dan data testing. ada beberapa tahapan dalam membuat sebuahalgoritma decision C4.5 adalah :

1. Mempersiapkan data training atau data asli mengambil Dataset di ambil dari Metropolitan University Sylhet, Bangladesh rahmansadik004 '@' gmail.com
2. Menghitung akar dari pohon.akar di ambil dari atribut yang di pilih dengan cara menghitung dengan nilai gain masing-masing atribut, nilai atribut yang paling tinggi menjadi akar pertama, sebelum menghitung nilai gain dari atribut, hitung terlebih dahulu nilai entropy Untuk menghitung nilai entropy di gunakan rumus :
    - Tahapan Algoritma Decision Tree
    - Menyiapakan Data Training
    - Menentukan Akar Dari Pohon
    - Menghitung Nilai Gain

$$Entropy\ (S) = \sum_{i=1}^{n} - pi * log2\ pi$$

S : Himpunan Kasus

A : Fitur

N : Jumlah Partisi S

Pi : Proporsi dari Si terhadap S

$$\text{Gain}\ (S,A) = Entropy\ (S) - \sum_{i=1}^{n} - pi * log2\ pi$$

S : Himpunan Kasus

A : Atribut

N : Jumblah Partisi atribut A

|Si| : jumlah kasus pada partisi ke-i

|S| : Jumlah kasus dalam S

Dari rumus di atas berikut penjelasan beserta contoh hitungan manual data, untuk data hanya menggambil 14 stempel data untuk menghitung nilai entropy dan gain. Berikut penjelasan pada gambar berikut :

no	Age	Gender	Polyuria	Polydipsia	Sudden weight loss	Label
1	44	Male	1	1	0	1
2	61	Male	1	1	1	1
3	43	Famale	1	1	1	1
4	67	Male	1	0	0	0
5	43	Male	0	0	1	0
6	30	Famale	1	0	1	1
7	48	Famale	0	0	1	0
8	40	Male	0	1	1	1
9	55	Famale	1	1	1	1
10	58	Male	0	1	0	0
11	35	Famale	1	1	0	1
12	40	Famale	1	0	1	1
13	56	Male	0	0	1	0
14	43	Male	1	1	1	1

Gambar 3. 2 contoh dataset

Setelah di hitung terdapat 14 studi kasus penyaki diabetes. Terdapat 9 orang yang positive dan 5 orang negative . lebih jelas liat pada **gambar 3.3**

		JTT	P(Si)	N (Si)
Total		14	9	5
Umur				
	Umur(30-50)	9	7	2
	Umur(50-60)	3	1	2
	Umur(60-70)	2	1	1
Jenis				
Kelamin	Laki-Laki	8	4	4
	Perempuan	6	5	1
Polyuria	Terkena	9	8	1
	T. Terkena	5	1	4
Polydipsia	Terkena	8	7	1
	T. Terkena	6	2	4
sudden weight loss	Terkena	10	7	3
	T. Terkena	4	2	2

Gambar 3. 3 Perhitugan dataset

a. Cara menghitung entropy sebagai berikut :

Entropy (Total)

1. ((-9/14) *imlog2 (9/14)+(-5/14) *imlog2 (5/14))
   = 0.940285959

2. ((-7/9) *imlog2 (7/9)+(-2/9) *imlog2 (2/9))
   = 0.764204507

3. ((-1/3) *imlog2 (1/3)+(-2/3) *imlog2 (2/3))
   = 0.918295834

4. ((-1/2) *imlog2 (1/2)+(-1/2) *imlog2 (1/2))
   = 1

5. ((-4/8) *imlog2 (4/8)+(-4/8) *imlog2 (4/8))

= 1

6. ((-5/6) *imlog2 (5/6)+(-1/6) *imlog2 (1/6))
   = 0.650022422

7. ((-8/9) *imlog2 (8/9)+(-1/9) *imlog2 (1/9))
   = 0.503258335

8. ((-1/5) *imlog2 (1/5)+(-4/1) *imlog2 (4/1))
   = 0.721928095

9. ((-7/8) *imlog2 (7/8)+(-1/8) *imlog2 (1/8))
   =0.543564443

10. ((-7/8) *imlog2 (7/8)+(-1/8) *imlog2 (1/8))
    =0.543564443

11. ((-2/6) *imlog2 (2/6)+(-4/6) *imlog2 (4/6))
    = 0.9182958335

12. ((-7/10) *imlog2 (7/10)+(-3/10) *imlog2 (3/10))
    = 0.881290899

13. ((-2/4) *imlog2 (2/4)+(-2/4) *imlog2 (2/4))
    = 1

b. Cara menghitung nilai Gain

1. 940285959((9/14*0.764)+(3/14*0.918) + (2/14*1))
   = 0.109376811

2. 0.940285959 ((8/14*1) + (6/14*0.650)
   = 0.090276349

3. 0.940285959((9/14*0.503)+ (5/14*0.721)
   = 0.874594206

4. 0.940285959((8/14*0.543)+(6/14*0.918)
   = 0.236122348

5. 0.940285959((10/14*0.881) + (4/14*1)
   = 0.025078174

Untuk lebih jelas bisa liat **gambar 3.4**

		JTT	P(Si)	N (Si)	Entrophy	Gain
Total		14	9	5	0.940285959	
Umur						0.109376811
	Umur(30-50)	9	7	2	0.764204507	
	Umur(50-60)	3	1	2	0.918295834	
	Umur(60-70)	2	1	1	1	
Jenis						0.090276349
Kelamin	Laki-Laki	8	4	4	1	
	Perempuan	6	5	1	0.650022422	
						0.874594206
Polyuria	Terkena	9	8	1	0.503258335	
	T. Terkena	5	1	4	0.721928095	
						0.236122348
Polydipsia	Terkena	8	7	1	0.543564443	
	T. Terkena	6	2	4	0.918295834	
						0.025078174
sudden weight loss	Terkena	10	7	3	0.881290899	
	T. Terkena	4	2	2	1	

Gambar 3. 4 perhitungan nilai entropy & gain

3. Dari dataset di atas peneliti telah memilah dataset menjadi data training dan data testing. Untuk data training (80%) dan data testing (20%) dan nanti akan menghasilkan prediksi dan akurasi.

Tabel 3. 4 kreteria Status Diabetes Data Training (80%)

No	Age	Gender	Polyuria	Polydipsia	Obesity	Label
1	58	Male	1	0	1	1
2	54	Male	0	1	0	0
3	67	Male	0	1	1	1
4	66	Male	1	0	0	0
5	54	Female	1	1	1	1
6	39	Male	0	1	0	0
7	48	Female	0	0	1	0
8	58	Female	1	1	0	0
….	….	….	….	….	….	….
416	40	Male	0	0	1	0

Tabel 3. 5 kreteria Status Diabetes Data Testing (20%)

No	Age	Gender	Polyuria	Polydipsia	Obesity	Label
1	39	Female	1	0	1	1
2	48	Female	1	1	0	1
3	85	Male	0	1	0	0
….	….	….	….	….	….	….
104	72	Male	1	0	0	0

3.2.4 Uji Coba

Di tahap ini akan di lakukan uji coba untuk mengetahui keberhasilan dari penerapan decision tree c4.5. proses uji coba di lakukan pada dataset yang berjumblah 50 data. Pada proses uji coba, tingkat keberhasilan di ukur dari tingkat akurasi yang di hasilkan.

Untuk mengitung akurasi (accuracy), di perlukan pemahaman tentang variabel yang akan di nilai. Terdapat 7 variabel yang di gunakan. Kemungkinan dari hasil deteksi yang harus di pahami yaitu nilai True Positive, True Negative, False Positive, Dan False Nagative. Penjelasan tentang variable tersebut akan di rangkum pada **gambar 3.5**

		Predicted Class	
		Class Yes	*Class No*
Actual Class	*Class Yes*	*True Positive* (TP)	*False Negative* (FN)
	Class No	*False Positive* (FP)	*True Negative* (TN)

Gambar 3. 5 Rumus Accuracy

a. True Positive (TP) atau Benar Positif adalah kasus dimana data diprediksi positif oleh sistem dan prediksi sistem memang benar adanya.
b. True Negative (TN) atau Benar Negatif adalah kasus dimana data diprediksi negatif oleh sistem dan prediksi sistem memang benar adanya
c. False Positive (FP) atau Salah Positif adalah kasus dimana data diprediksi positif oleh sistem tetapi kenyataannya adalah data bernilai negatif
d. False Negative (FN) atau Salah Negatif adalah kasus dimana data diprediksi negatif oleh sistem tetapi kenyataannya adalah data bernilai positive

Accuracy atau akurasi merupakan rasio prediksi benar positif dan benar negatif (TP + TN) dibandingkan dengan keseluruhan data (TP + FP + TN + FN). Rumus perhitungan akurasi dapat dilihat **gambar 3.6**

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

Gambar 3. 6 Rumus Akurasi

3.2.5 Penarikan kesimpulan

Pada hasil uji coba akan di analisis dan di bahas tentang metode *decision tree C4.5* dalam proses mengimplementasi penyakit diabetes. Tahap ini, akan menghasilkan prediksi, pohon keputusan dan akurasi dan akan mengetahui akhir nilai Positive (1) dan Negative (0).