# BAB IV

# HASIL DAN PEMBAHASAN

## 4.1 Deskripsi Data, *Prepocessing*, Partisi Data

Deskripsi variabel penelitian dapat dilihat pada tabel 4.1 dibawah ini.

Tabel 4. 1 Deskripsi variabel

| No | Variabel | Data | Frekuensi |
|---|---|---|---|
| 1 | $X_1$(*Age*) | 1 : <20 tahun | 1 |
| | | 2 : 20-40 tahun | 167 |
| | | 3 : >40 tahun | 352 |
| 2 | $X_2$(*Gender)* | 1: Male | 328 |
| | | 0: Female | 192 |
| 3 | $X_3$(*Polyuria*) | 1: Yes | 258 |
| | | 0: No | 262 |
| 4 | $X_4$(*Polydipsia)* | 1: Yes | 233 |
| | | 0: No | 287 |
| 5 | $X_5$(*sudden_weight_loss)* | 1: Yes | 217 |
| | | 0: No | 303 |
| 6 | $X_6$(*weakness*) | 1: Yes | 305 |
| | | 0: No | 215 |
| 7 | $X_7$(*Polyphagia)* | 1: Yes | 237 |
| | | 0: No | 283 |
| 8 | $X_8$(*Genital_thrush*) | 1: Yes | 116 |
| | | 0: No | 404 |
| 9 | $X_9$(*visual_blurring)* | 1: Yes | 233 |
| | | 0: No | 287 |
| 10 | $X_{10}$(*Itching)* | 1: Yes | 253 |
| | | 0: No | 267 |
| 11 | $X_{11}$(*Irritability)* | 1: Yes | 126 |
| | | 0: No | 394 |
| 12 | $X_{12}$(*delayed_healing)* | 1: Yes | 239 |
| | | 0: No | 281 |
| 13 | $X_{13}$(*partial_paresis)* | 1: Yes | 224 |
| | | 0: No | 296 |
| 14 | $X_{14}$(*muscle_stiffness)* | 1: Yes | 195 |
| | | 0: No | 325 |
| 15 | $X_{15}$(*Alopecia)* | 1: Yes | 179 |
| | | 0: No | 341 |

| | | | |
|---|---|---|---|
| 16 | $X_{16}$(*Obesity)* | 1: Yes<br>0: No | 88<br>432 |
| 17 | *Y* (*class)* | 1: Positive<br>0: Negative | 320<br>200 |

Preprosessing data dilakukan dengan mendiskritisasi variabel kontinu menjadi varibel katagorik. Pada data diabetes diatas terdapat variabel kontinu yaitu $X_1$ yang merupakan variabel umur dengan interval 16-90 tahun. Jika umur pasien kurang dari 20 tahun makan dikatagorikan sebagai *young* (1), jika umur pasien berada pada interval 20 tahun sampai 40 tahun dikatagorikan sebagai *medium/*dewasa (2), dan jika umur pasien berada diatas 40 tahun maka dikatagorikan *old* (3) (Anggraeni & Ramadhani, 2018).

Partisi data dilakukan untuk membagi data ke dalam dua bagian secara acak, dimana dalam penelitian ini yang digunakan adalah *train/test* split. Data dipartisi menjadi 80% data *train* (416 pengamatan) dan 20% data *test* (104 pengamatan).

Tabel 4.2 Data *train* 80 %

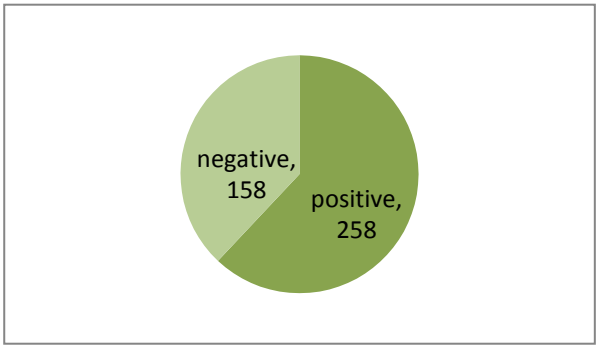
| No | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $X_6$ | $X_7$ | $X_8$ | $X_9$ | $X_{10}$ | $X_{11}$ | $X_{12}$ | $X_{13}$ | $X_{14}$ | $X_{15}$ | $X_{16}$ | $Y$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 1 |
| 2 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 2 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 1 |
| 4 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | 2 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 0 |
| . | | | | | | | | | | | | | | | | | |
| . | | | | | | | | | | | | | | | | | |
| . | | | | | | | | | | | | | | | | | |
| 416 | 2 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 1 |

Tabel 4.3 Data *test* 20%

| No | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $X_6$ | $X_7$ | $X_8$ | $X_9$ | $X_{10}$ | $X_{11}$ | $X_{12}$ | $X_{13}$ | $X_{14}$ | $X_{15}$ | $X_{16}$ | $Y$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 3 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 1 |
| 2 | 3 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 3 | 3 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 1 |
| 4 | 3 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 |
| 5 | 3 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 1 |
| . | | | | | | | | | | | | | | | | | |
| . | | | | | | | | | | | | | | | | | |
| . | | | | | | | | | | | | | | | | | |
| 104 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |

## *4.2* Mengklasfikasi Status Penyakit Diabetes dengan Metode *Decision Tree*

Langkah pertama yang dilakukan adalah dengan mencari nilai *entropy* dan juga nilai *gain*. Nilai *gain* tertinggi pertama akan dijadikan sebagai akar pohon. Banyaknya sampel yang digunakan sama dengan jumlah data *train* yaitu sebanyak 416. Berdasarkan banyaknya data sampel yang digunakan terdapat 258 orang yang positif diabetes dan 158 orang yang negative diabetes.




Gambar 4.1 Grafik jumlah data *train* positif dan negaitif

Maka peluang dari variabel respon untuk katagori positif dan negatif yaitu :

$$P(Y=0) = \frac{n(Y=0)}{n_{total}} = \frac{158}{416} = 0.3798$$

$$P(Y=1) = \frac{n(Y=1)}{n_{total}} = \frac{258}{416} = 0.6202$$

Selanjutnya akan dibentuk node ke-1 dengan mencari nilai e*ntropy* setiap variabel, yaitu sebagi berikut:

Tabel 4.4 perhitungan *entropy* dan *gain* node 1

| Variabel | Label | Jumlah | 0 | 1 | $p_0$ | $p_1$ | Entropy | Gain |
|---|---|---|---|---|---|---|---|---|
| Total | | 416 | 258 | 158 | 0.6201 | 0.3798 | 0.9579 | |
| $X_1$ | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 0.1271 |
| | 2 | 348 | 191 | 157 | 0.5488 | 0.4511 | 0.9931 | |
| | 3 | 67 | 67 | 0 | 1 | 0 | 0 | |
| $X_2$ | 0 | 152 | 136 | 16 | 0.8947 | 0.1052 | 0.4854 | 0.1485 |
| | 1 | 264 | 122 | 142 | 0.4621 | 0.5378 | 0.9958 | |
| $X_3$ | 0 | 216 | 69 | 147 | 0.3194 | 0.6805 | 0.9037 | 0.3409 |
| | 1 | 200 | 189 | 11 | 0.945 | 0.055 | 0.3072 | |
| $X_4$ | 0 | 226 | 75 | 151 | 0.3318 | 0.6681 | 0.9168 | 0.3558 |
| | 1 | 190 | 183 | 7 | 0.9631 | 0.0368 | 0.2276 | |
| $X_5$ | 0 | 249 | 111 | 138 | 0.4457 | 0.5542 | 0.9915 | 0.1522 |
| | 1 | 167 | 147 | 20 | 0.8802 | 0.1197 | 0.5286 | |
| $X_6$ | 0 | 176 | 85 | 91 | 0.4829 | 0.5170 | 0.9991 | 0.0423 |
| | 1 | 240 | 173 | 67 | 0.7208 | 0.2791 | 0.8543 | |
| $X_7$ | 0 | 226 | 104 | 122 | 0.4601 | 0.5398 | 0.9954 | 0.0972 |
| | 1 | 190 | 154 | 36 | 0.8105 | 0.1894 | 0.7003 | |
| $X_8$ | 0 | 327 | 197 | 130 | 0.6024 | 0.3975 | 0.9695 | 0.0036 |
| | 1 | 89 | 61 | 28 | 0.6853 | 0.3146 | 0.8984 | |
| $X_9$ | 0 | 228 | 115 | 113 | 0.5043 | 0.4956 | 0.9999 | 0.0510 |
| | 1 | 188 | 143 | 45 | 0.7606 | 0.2393 | 0.7939 | |
| $X_{10}$ | 0 | 220 | 136 | 84 | 0.6181 | 0.3818 | 0.9593 | 1.38951E-05 |
| | 1 | 196 | 122 | 74 | 0.6224 | 0.3775 | 0.9562 | |
| $X_{11}$ | 0 | 318 | 174 | 144 | 0.5471 | 0.4528 | 0.9935 | 0.0590 |
| | 1 | 98 | 84 | 14 | 0.8571 | 0.1428 | 0.5916 | |
| $X_{12}$ | 0 | 234 | 138 | 96 | 0.5897 | 0.4102 | 0.9766 | 0.0036 |
| | 1 | 182 | 120 | 62 | 0.6593 | 0.3406 | 0.9254 | |
| $X_{13}$ | 0 | 237 | 101 | 136 | 0.4261 | 0.5738 | 0.9842 | 0.1658 |
| | 1 | 179 | 157 | 22 | 0.8770 | 0.1229 | 0.5376 | |
| $X_{14}$ | 0 | 264 | 152 | 112 | 0.5757 | 0.4242 | 0.9833 | 0.0106 |
| | 1 | 152 | 106 | 46 | 0.6973 | 0.3026 | 0.8844 | |
| $X_{15}$ | 0 | 283 | 199 | 84 | 0.7031 | 0.2968 | 0.8773 | 0.0442 |
| | 1 | 133 | 59 | 74 | 0.4436 | 0.5563 | 0.9908 | |
| $X_{16}$ | 0 | 345 | 208 | 137 | 0.6028 | 0.3971 | 0.9692 | 0.0045 |
| | 1 | 71 | 50 | 21 | 0.7042 | 0.2957 | 0.8760 | |

$$Entropi(Y) = -\sum_{i=1}^{k} P_i log_2 P_i$$

$$= -(P(Y=0)log_2 P(Y=0) + (P(Y=1)log_2 P(Y=1)$$

$$= -((0.3798)log_2(0.3798) + (0.6202)log_2(0.6202))$$

$$= 0.9579$$

$$Entropi(X_1 = 1) = -(P(Y=0|X_1=1)log_2 P(Y=0|X_1=1) +$$

$$P(Y=1|X_1=1)log_2 P(Y=1|X_1=1))$$

$$= -((1)log_2(1) + (0)log_2(0))$$

$$= 0$$

$$Entropi(X_1 = 2) = -(P(Y=0|X_1=2)log_2 P(Y=0|X_1=2) +$$

$$P(Y=1|X_1=2)log_2 P(Y=1|X_1=2))$$

$$= -((0.4511)log_2(0.4511) + (0.5488)log_2(0.5488))$$

$$= 0.9931$$

$$Entropi(X_1 = 3) = -(P(Y=0|X_1=3)log_2 P(Y=0|X_1=3) +$$

$$P(Y=1|X_1=3)log_2 P(Y=1|X_1=3))$$

$$= -((0)log_2(0) + (1)log_2(1))$$

$$= 0$$

$$Entropi(X_2 = 0) = -(P(Y=0|X_2=0)log_2 P(Y=0|X_2=0) +$$

$$P(Y=1|X_2=0)log_2 P(Y=1|X_2=0))$$

$$= -((0.1053)log_2(0.1053) + (0.8947)log_2(0.8947))$$

$$= 0.4854$$

$$Entropi(X_2 = 1) = -(P(Y = 0|X_2 = 1)log_2P(Y = 0|X_2 = 1) + P(Y = 1|X_2 = 1)log_2P(Y = 1|X_2 = 1))$$
$$= -((0.5379)log_2(0.5379) + (0.4621)log_2(0.4621))$$
$$= 0.9958$$
$$Entropi(X_3 = 0) = -(P(Y = 0|X_3 = 0)log_2P(Y = 0|X_3 = 0) + P(Y = 1|X_3 = 0)log_2P(Y = 1|X_3 = 0))$$
$$= -((0.6805)log_2(0.6805) + (0.3194)log_2(0.3194))$$
$$= 0.9037$$

$$Entropi(X_3 = 1) = -(P(Y = 0|X_3 = 1)log_2P(Y = 0|X_3 = 1) + P(Y = 1|X_3 = 1)log_2P(Y = 1|X_3 = 1))$$
$$= -((0.055)log_2(0.055) + (0.945)log_2(0.945))$$
$$= 0.3072$$
$$Entropi(X_4 = 0) = -(P(Y = 0|X_4 = 0)log_2P(Y = 0|X_4 = 0) + P(Y = 1|X_4 = 0)log_2P(Y = 1|X_4 = 0))$$
$$= -((0.6681)log_2(0.6681) + (0.3318)log_2(0.3318))$$
$$= 0.9168$$

$$Entropi(X_4 = 1) = -(P(Y = 0|X_4 = 1)log_2P(Y = 0|X_4 = 1) + P(Y = 1|X_4 = 1)log_2P(Y = 1|X_4 = 1))$$
$$= -((0.0368)log_2(0.0368) + (0.9632)log_2(0.9632))$$

$$= 0.2276$$

Perhitungan *entropy* dilanjutkan sampai semua variabel terhitung *entropy* nya, dengan cara yang sama seperti yang dilakukan di atas. Setelah semua variabel terhitung *entropy* nya, langkah selanjutnya yaitu menghitung nilai *gain.*

$$Gain(y,x_1) = Entropi(y) - \sum_{i=1}^{3} - \frac{x_{1i}}{n_{total}} * Entropi\,(x_{1i})$$

$$= 0.9579 - ((\frac{1}{416} * 0) + (\frac{348}{416} * 0.9931) + (\frac{67}{416} * 0))$$

$$= 0.1271$$

$$Gain(y, x_2) = Entropi(y) - \sum_{i=1}^{2} - \frac{x_{2i}}{n_{total}} * Entropi\,(x_{2i})$$

$$= 0.9579 - ((\frac{264}{416} * 0.9958) + (\frac{152}{416} * 0.4854))$$

$$= 0.1485$$

$$Gain(y, x_3) = Entropi(y) - \sum_{i=1}^{2} - \frac{x_{3i}}{n_{total}} * Entropi\,(x_{3i})$$

$$= 0.9579 - ((\frac{200}{416} * 0.3072) + (\frac{216}{416} * 0.9037))$$

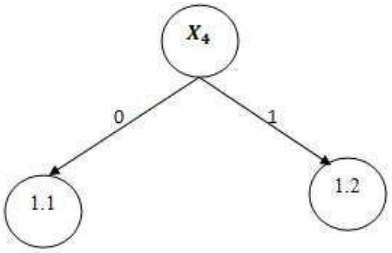
$$= 0.3409$$

$$Gain(y, x_4) = Entropi(y) - \sum_{i=1}^{2} - \frac{x_{4i}}{n_{total}} * Entropi\,(x_{4i})$$

$$= 0.9579 - ((\frac{190}{416} * 0.2276) + (\frac{226}{416} * 0.9168))$$

$$= 0.3558$$

Perhitungan *gain* dilakukan sampai semua variabel terhitung. Dari perhitungan tersebut hasil *entropy* dan juga *gain* diatas terlihat bahwa nilai *gain* tertinggi terletak pada $X_4$ yaitu sebesar 0.3558, ini berarti $X_4$ menjadi *root*/ akar dari node ke-1. Pohon keputusan untuk node 1 dapat dilihat pada gambar berikut.




Gambar 4.2 pohon keputusan *root node*.

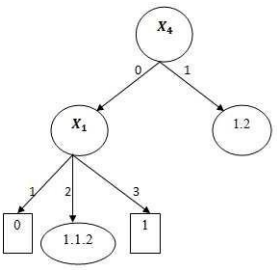
Selanjutnya akan dilakukan perhitungan *entropy* dan juga *gain* untuk semua cabang dari variabel $X_4$ sebagai akar pohon yang mana memiliki dua katagori yaitu katagori *Yes* (1) memiliki 190 kasus dan katagori *No* (0) memiliki 226 kasus. Berikut ini tabel perhitungan *entropy* dan *gain* untuk node 1.1 katagori *No* (0).

Tabel 4.5 perhitungan *entropy* dan *gain* node 1.1

| Variabel | Label | Jumlah | 0 | 1 | $p_0$ | $p_1$ | Entropy | Gain |
|---|---|---|---|---|---|---|---|---|
| $X_1$ | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 0.1474 |
| | 2 | 206 | 56 | 150 | 0.2718 | 0.7281 | 0.8440 | |
| | 3 | 19 | 19 | 0 | 1 | 0 | 0 | |
| $X_2$ | 0 | 55 | 39 | 16 | 0.7090 | 0.2909 | 0.8698 | 0.1433 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 1 | 171 | 36 | 135 | 0.2105 | 0.7894 | 0.7424 | |
| $X_3$ | 0 | 181 | 41 | 140 | 0.2265 | 0.7734 | 0.7718 | 0.1388 |
| | 1 | 45 | 34 | 11 | 0.7555 | 0.2444 | 0.8023 | |
| $X_5$ | 0 | 177 | 44 | 133 | 0.2485 | 0.7514 | 0.8090 | 0.0775 |
| | 1 | 49 | 31 | 18 | 0.6326 | 0.3673 | 0.9486 | |
| $X_6$ | 0 | 131 | 44 | 87 | 0.3358 | 0.6641 | 0.9208 | 7.25357E-05 |
| | 1 | 95 | 31 | 64 | 0.3263 | 0.6736 | 0.9111 | |
| $X_7$ | 0 | 158 | 42 | 116 | 0.2658 | 0.7341 | 0.8354 | 0.0320 |
| | 1 | 68 | 33 | 35 | 0.4852 | 0.5147 | 0.9993 | |
| $X_8$ | 0 | 180 | 57 | 123 | 0.3166 | 0.6833 | 0.9007 | 0.0028 |
| | 1 | 46 | 18 | 28 | 0.3913 | 0.6086 | 0.9656 | |
| $X_9$ | 0 | 162 | 50 | 112 | 0.3086 | 0.6913 | 0.8915 | 0.0043 |
| | 1 | 64 | 25 | 39 | 0.3906 | 0.6093 | 0.9652 | |
| $X_{10}$ | 0 | 138 | 54 | 84 | 0.3913 | 0.6086 | 0.9656 | 0.0184 |
| | 1 | 88 | 21 | 67 | 0.2386 | 0.7613 | 0.7927 | |
| $X_{11}$ | 0 | 190 | 53 | 137 | 0.2789 | 0.7210 | 0.8540 | 0.0452 |
| | 1 | 36 | 22 | 14 | 0.6111 | 0.3888 | 0.9640 | |
| $X_{12}$ | 0 | 142 | 52 | 90 | 0.3661 | 0.6338 | 0.9477 | 0.0065 |
| | 1 | 84 | 23 | 61 | 0.2738 | 0.7261 | 0.8468 | |
| $X_{13}$ | 0 | 176 | 44 | 132 | 0.25 | 0.75 | 0.8112 | 0.0730 |
| | 1 | 50 | 31 | 19 | 0.62 | 0.38 | 0.9580 | |
| $X_{14}$ | 0 | 161 | 50 | 111 | 0.3105 | 0.6894 | 0.8938 | 0.0036 |
| | 1 | 65 | 25 | 40 | 0.3846 | 0.6153 | 0.9612 | |
| $X_{15}$ | 0 | 126 | 48 | 78 | 0.3809 | 0.73 | 0.9587 | 0.0099 |
| | 1 | 100 | 27 | 73 | 0.27 | 0.6190 | 0.8414 | |
| $X_{16}$ | 0 | 196 | 65 | 131 | 0.3316 | 0.6683 | 0.9165 | 1.08257E-06 |
| | 1 | 30 | 10 | 20 | 0.333 | 0.6666 | 0.9182 | |

Dari tabel perhitungan diatas terlihat bahwa yang memiliki nilai *gain* tertinggi adalah $X_1$ dengan nilai *gain* sebesar 0.1474 . Maka $X_1$ menjadi cabang pohon pertama dari katagori 0 (*No*). Pada variabel $X_1$ memiliki tiga katagori dimana dari ketiga katagori tersebut dua diantaranya memiliki hasil akhir untuk klasifikasi penyakit diabetes. Dua katagori tersebut yaitu pada katagori 1 (umur kurang dari 20 tahun (*young*)) masuk dalam klasifikasi negatif diabetes dengan jumlah 1 orang dan katagori 3 (umur lebih dari 40 tahun (*old*)) masuk dalam klasifikasi positif diabetes dengan jumlah 19 orang. Pohon keputusan dengan akar $X_4$ dan katagori 0 (*No*) adalah $X_1$dapat dilihat pada gambar berikut.




Gambar 4.3 pohon keputusan node 1.1

Selanjutnya melakukan perhitungan *entropy* dan juga *gain* dengan cara yang sama untuk seluruh cabang pohon keputusan. Proses perhitungan pohon keputusan dihentikan jika semua data sampel berada dalam kelas yang sama, tidak ada lagi atribut yang akan dilakukan partisi, atau tidak ada data sampel lagi yang akan diuji. Dari perhitungan yang dilakukan maka akan terbentuk sebuah model pengkondisian untuk menentukan klasifikasi pada penyakit diabetes. Model pengkondisian dari metode *decision tree* dapat dilihat pada gambar dibawah ini.

```
Polydipsia = 0
|   Age = 1: 0 {1=0, 0=1}
|   Age = 2
|   |   Polyuria = 0
|   |   |   Gender = 0
|   |   |   |   Alopecia = 0
|   |   |   |   |   visual_blurring = 0
|   |   |   |   |   |   muscle_stiffness = 0
|   |   |   |   |   |   |   Polyphagia = 0
|   |   |   |   |   |   |   |   Irritability = 0: 1 {1=6, 0=2}
|   |   |   |   |   |   |   |   Irritability = 1: 0 {1=0, 0=1}
|   |   |   |   |   |   |   Polyphagia = 1: 0 {1=0, 0=1}
|   |   |   |   |   |   muscle_stiffness = 1: 1 {1=4, 0=0}
|   |   |   |   |   visual_blurring = 1: 1 {1=11, 0=0}
|   |   |   |   Alopecia = 1: 0 {1=0, 0=12}
|   |   |   Gender = 1
|   |   |   |   Irritability = 0
|   |   |   |   |   Alopecia = 0: 0 {1=0, 0=68}
|   |   |   |   |   Alopecia = 1
|   |   |   |   |   |   Itching = 0
|   |   |   |   |   |   |   Polyphagia = 0
|   |   |   |   |   |   |   |   weakness = 0: 0 {1=0, 0=6}
|   |   |   |   |   |   |   |   weakness = 1: 1 {1=2, 0=2}
|   |   |   |   |   |   |   Polyphagia = 1: 1 {1=1, 0=0}
|   |   |   |   |   |   Itching = 1: 0 {1=0, 0=38}
|   |   |   |   Irritability = 1
|   |   |   |   |   Genital_thrush = 0
|   |   |   |   |   |   weakness = 0: 0 {1=0, 0=5}
|   |   |   |   |   |   weakness = 1
|   |   |   |   |   |   |   visual_blurring = 0: 1 {1=1, 0=0}
|   |   |   |   |   |   |   visual_blurring = 1: 0 {1=0, 0=4}
|   |   |   |   |   Genital_thrush = 1: 1 {1=3, 0=0}
|   |   Polyuria = 1
|   |   |   Itching = 0: 1 {1=20, 0=0}
|   |   |   Itching = 1
|   |   |   |   delayed_healing = 0: 1 {1=7, 0=0}
|   |   |   |   delayed_healing = 1
|   |   |   |   |   Gender = 0: 1 {1=1, 0=0}
|   |   |   |   |   Gender = 1: 0 {1=0, 0=11}
|   Age = 3: 1 {1=19, 0=0}
```

Gambar 4.4 Pengkondisian pohon keputusan metode *decision tree* menggunanakan software *rapid miner*

Dari gambar diatas terlihat sebagian pengkondisian dari pohon keputusan dengan metode *decision tree. Polydipsia* ( $\boldsymbol{X_4}$ ) terpilih sebagai akar pohon. Jika *polydipsia* (perasaan sangat haus) dengan katagori *No* (0) dan umur dengan katagori 1 (umur kurang dari 20 tahun ) maka terklasifikasi kedalam katagori negatif diabetes. Jika *polydipsia* (perasaan sangat haus) dengan katagori *No* (0) dan umur dengan katagori 3 (umur lebih dari 40 tahun ) maka terklasifikasi kedalam katagori positif diabetes. Sedangkan jika *polydipsia* (perasaan sangat haus) dengan katagori *No* (0) kemudian umur dengan katagori 2 (umur antara 20-

40 tahun ), *polyuria* (kelainan produksi air seni) dengan katagori *No* (0) , jenis kelamin dengan katagori perempuan (0), *alopecia* (Kerontokan Rambut/kebotakan) dengan katagori *No* (0), *visual blurring* (Penglihatan tidak jelas) dengan katagori *No* (0), *muscle stiffness* (Kekakuan otot) dengan katagori *No* (0), *polyphagia* (peningkatan nafsu makan berlebih) dengan katagori *No* (0), dan *irritability* (kepekaan terhadap rangsangan ) dengan katagori *No* (0) maka terklasifikasi kedalam katagori positif diabetes, sedangakan jika *irritability* (kepekaan terhadap rangsangan) dengan katagori *Yes* (1) maka terklasifikasi kedalam katagori negatif diabetes. Selanjutnya mencari setiap cabang lain yang belum terklasifikasi.

Model klasifikasi pada *decision tree* secara lengkap dapat dilihat pada **Lampiran**. Hasil klasifikasi dengan menggunakan metode *decision tree* dapat dilihat pada tabel 4.6 dibawah ini.

Tabel 4.6 *confusion matrix* metode *decision tree*

| | *True* 1 | *True* 0 |
|---|---|---|
| *Predict* 1 | 58 | 4 |
| *Predict* 0 | 5 | 37 |

Berdasarkan tabel yang ada diatas dapat dilihat bahwa 58 orang di prediksi benar masuk dalam klasifikasi positif diabetes, sedangkan 37 orang diprediksi benar masuk dalam klasifikasi negatif diabetes. Terdapat 5 orang di prediksi masuk dalam katagori positif diabetes, namun ternyata masuk dalam katagori negatif diabetes pada data sebenarnya. Sedangkan 4 orang yang diprediksi masuk

dalam katagori negatif diabetes ternyata masuk dalam katagori positif diabetes pada data sebenarnya. Kita juga dapat melihat tingkat akurasi, presisi, recall, specificity dan F1 score dari tabel diatas berikut ini.

$$akurasi = \frac{58 + 37}{104} = 0.9135$$

$$presisi = \frac{58}{58 + 4} = 0.9355$$

$$recall = \frac{58}{58 + 5} = 0.9206$$

$$specificity = \frac{37}{37 + 4} = 0.9024$$

$$F1\ score = 2\,(\frac{0.9206 * 0.9355}{0.9206 + 0.9355}) = 0.9280$$

Jadi, dengan menggunakan metode *decision tree* akurasi yang didapatkan sebesar 0.9135, presisi sebesar 0.9355, recall sebesar 0.9206, specificity sebesar 0.9024, dan F1 score sebesar 0.9280, hal ini menunjukkan bahwa ketepatan akurasi dalam memprediksi klasifikasi data penyakit diabetes dengan menggunakan metode *decision tree* adalah 91.35%.

## *4.3* Mengklasfikasi Status Penyakit Diabetes dengan Metode *Random Forest*

Langkah awal yang dilakukan dalam *Random Forest* yaitu menentukan banyaknya pohon yang akan di bentuk ($n_{pohon}$) dan juga banyaknya varibel yang dipilih untuk membangun pohon keputusan. Dalam penelitian akan dibangun 11 pohon dengan 4 variabel terpilih. Selanjutnya mengambil sampel data *train*,

dimana banyaknya data *train* ini yaitu sebesar 416 untuk kemudian dilakukan *bootstrap sampling* atau pengambilan sampel secara acak dengan pengembalian. Proses ini dilakukan setiap akan membangun suatu pohon keputusan, sehingga setiap pohon akan memiliki sampel yang berbeda-beda. Hasil dari proses *bootstrap sampling* untuk pohon pertama dapat dilihat pada tabel dibawah ini.

Tabel 4.7 *bootstrap sampling* pohon pertama

| No | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $X_6$ | $X_7$ | $X_8$ | $X_9$ | $X_{10}$ | $X_{11}$ | $X_{12}$ | $X_{13}$ | $X_{14}$ | $X_{15}$ | $X_{16}$ | $Y$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 72 | 2 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 0 |
| 180 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 162 | 2 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 0 | 0 | 1 |
| 159 | 2 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| 92 | 2 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 |
| . | | | | | | | | | | | | | | | | | |
| . | | | | | | | | | | | | | | | | | |
| . | | | | | | | | | | | | | | | | | |
| 251 | 2 | 0 | 1 | 1 | 0 | `1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 |

Berdasarkan banyaknya data sampel yang digunakan pada tabel diatas terdapat 250 orang yang positif diabetes dan 166 orang yang negatif diabetes. Maka peluang dari variabel respon untuk katagori positif dan negatif yaitu :

$$P(Y = 0) = \frac{n(Y = 0)}{n_{total}} = \frac{166}{416} = 0.3991$$

$$P(Y = 1) = \frac{n(Y = 1)}{n_{total}} = \frac{250}{416} = 0.6009$$

Selanjutnya akan dibentuk node ke-1 untuk pohon pertama dengan mencari nilai *entropy* dan juga g*ain* untuk variabel yang ditentukan secara acak. Pada node selanjutnya juga akan dilakukan pengacakan variabel. Pada node pertama ini didapatkan variabel $X_2$, $X_3$, $X_4$, $X_5$ yang akan dihitung nilai *entropy* dan juga g*ain*, yaitu sebagai berikut:

Tabel 4.8 Perhitungan *entropy* dan juga *gain* untuk variabel $X_2$,$X_3$,$X_4$,$X_5$

| Atribut | Label | Jumlah | 1 | 0 | $p_1$ | $p_0$ | Entropy | Gain |
|---|---|---|---|---|---|---|---|---|
| Total | | 416 | 250 | 166 | 0.6009 | 0.3991 | 0.9704 | |
| $\boldsymbol{X_2}$ | 0 | 150 | 132 | 18 | 0.88 | 0.12 | 0.5293 | 0.1459 |
| | 1 | 266 | 118 | 148 | 0.4436 | 0.5563 | 0.9908 | |
| $\boldsymbol{X_3}$ | 0 | 216 | 59 | 157 | 0.2731 | 0.7268 | 0.8459 | 0.4038 |
| | 1 | 200 | 191 | 9 | 0.955 | 0.045 | 0.2648 | |
| $\boldsymbol{X_4}$ | 0 | 228 | 72 | 156 | 0.3158 | 0.6842 | 0.8997 | 0.3418 |
| | 1 | 188 | 178 | 10 | 0.9468 | 0.0532 | 0.2998 | |
| $\boldsymbol{X_5}$ | 0 | 250 | 105 | 145 | 0.42 | 0.58 | 0.9814 | 0.1619 |
| | 1 | 166 | 145 | 21 | 0.8735 | 0.1265 | 0.5477 | |

$$Entropi(Y) = -\sum_{i=1}^{2} P_i log_2 P_i$$

$$= -(P(Y=0)log_2 P(Y=0) + (P(Y=1)log_2 P(Y=1)$$

$$= -((0.3991)log_2(0.3991) + (0.6009)log_2(0.6009))$$

$$= 0.9703$$

$$Entropi(X_2 = 0) = -(P(Y=0|X_2=0)log_2 P(Y=0|X_2=0) +$$

$$P(Y=1|X_2=0)log_2 P(Y=1|X_2=0))$$

$$= -((0.12)log_2(0.12) + (0.88)log_2(0.88))$$

$$= 0.5293$$

$$Entropi(X_2 = 1) = -(P(Y=0|X_2=1)log_2 P(Y=0|X_2=1) +$$

$$P(Y=1|X_2=1)log_2 P(Y=1|X_2=1))$$

$$= -((0.5563)log_2(0.5563) + (0.4436)log_2(0.4436))$$

$$= 0.9908$$

$$Entropi(X_3 = 0) = -(P(Y=0|X_3=0)log_2 P(Y=0|X_3=0) +$$

$$P(Y = 1|X_3 = 0)log_2P(Y = 1|X_3 = 0))$$

$$= -((0.7268)log_2(0.7268) + (0.2731)log_2(0.2731))$$

$$= 0.8459$$

$$Entropi(X_3 = 1) = -(P(Y = 0|X_3 = 1)log_2P(Y = 0|X_3 = 1) +$$

$$P(Y = 1|X_3 = 1)log_2P(Y = 1|X_3 = 1))$$

$$= -((0.054)log_2(0.054) + (0.955)log_2(0.955))$$

$$= 0.2647$$

$$Entropi(X_4 = 0) = -(P(Y = 0|X_4 = 0)log_2P(Y = 0|X_4 = 0) +$$

$$P(Y = 1|X_4 = 0)log_2P(Y = 1|X_4 = 0))$$

$$= -((0.6842)log_2(0.6842) + (0.3157)log_2(0.3157))$$

$$= 0.8997$$

$$Entropi(X_4 = 1) = -(P(Y = 0|X_4 = 1)log_2P(Y = 0|X_4 = 1) +$$

$$P(Y = 1|X_4 = 1)log_2P(Y = 1|X_4 = 1))$$

$$= -((0.0532)log_2(0.0532) + (0.9468)log_2(0.9468))$$

$$= 0.2998$$

$$Entropi(X_5 = 0) = -(P(Y = 0|X_5 = 0)log_2P(Y = 0|X_5 = 0) +$$

$$P(Y = 1|X_5 = 0)log_2P(Y = 1|X_5 = 0))$$

$$= -((0.58)log_2(0.58) + (0.42)log_2(0.42))$$

$$= 0.9814$$

$$Entropi(X_5 = 1) = -(P(Y = 0|X_5 = 1)log_2P(Y = 0|X_5 = 1) +$$

$$P(Y = 1|X_5 = 1)log_2P(Y = 1|X_5 = 1))$$

$$= -((0.1265)log_2(0.1265) + (0.8734)log_2(0.8734))$$

$$= 0.5477$$

$$Gain(y, x_2) = Entropi(y) - \sum_{i=1}^{2} - \frac{x_{2i}}{n_{total}} * Entropi\ (x_{2i})$$

$$= 0.9703 - ((\frac{150}{416} * 0.5293) + (\frac{266}{416} * 0.9990))$$

$$= 0.1459$$

$$Gain(y, x_3) = Entropi(y) - \sum_{i=1}^{2} - \frac{x_{3i}}{n_{total}} * Entropi\ (x_{3i})$$

$$= 0.9703 - ((\frac{216}{416} * 0.8459) + (\frac{200}{416} * 0.2647))$$

$$= 0.4038$$

$$Gain(y, x_4) = Entropi(y) - \sum_{i=1}^{2} - \frac{x_{4i}}{n_{total}} * Entropi\ (x_{4i})$$

$$= 0.9703 - ((\frac{228}{416} * 0.8997) + (\frac{188}{416} * 0.2998))$$

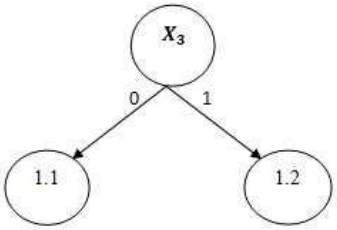
$$= 0.3417$$

$$Gain(y, x_5) = Entropi(y) - \sum_{i=1}^{2} - \frac{x_{5i}}{n_{total}} * Entropi\ (x_{5i})$$

$$= 0.9703 - ((\frac{250}{416} * 0.9814) + (\frac{166}{416} * 0.5477))$$

$= 0.1619$

Berdasarkan hasil diatas, maka didapatkan nilai gain tertinggi yaitu pada $X_3$ (kelainan produksi air seni atau *polyuria)* sebesar 0.4038. Dengan hasil ini, maka $X_3$ menjadi akar pohon pertama. Selanjutnya akan di cari cabang dari pohon $X_3$ yang memiliki dua katagori. Cara perhitungan sama seperti yang dilakukan di atas. Pohon untuk *root node* dapat dilihat pada gambar di bawah ini.




Gambar 4.5 Pohon keputusan *root node*

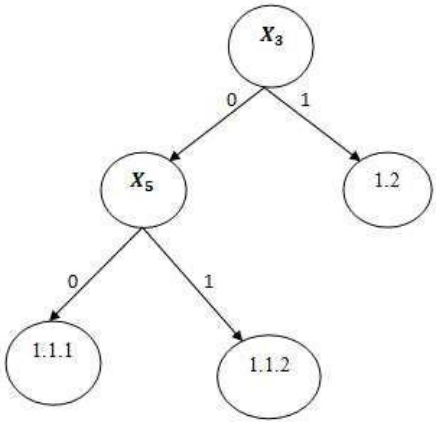
Selanjutnya melakukan perhitungan untuk cabang $X_3$. Dimana perhitungan ini untuk katagori *No* (0) terlebih dahulu. Pada variabel $X_3$ memiliki 216 kasus dengan 59 orang termasuk penderita diabetes dan 157 orang termasuk yang tidak menderita diabetes. Sebelum melakukan perhitungan *entropy* dan *gain*, dilakukan pengacakan variabel terlebih dahulu. Pilih 4 variabel untuk dilakukan pengacakan, kecuali $X_3$ karena variebl ini sudah menjadi akar pohon atau *root node.* Variabel yang terpilih yaitu $X_5$,$X_6$,$X_{10}$,$X_{14}$. Perhitungan untuk *entropy* dan *gain* pada variabel tersebut dapat dilihat pada tabel berikut ini.

Tabel 4.9 Perhitungan *entropy* dan juga *gain* untuk variabel $X_5$,$X_6$,$X_{10}$,$X_{14}$

| Atribut | Label | Jumlah | 1 | 0 | $p_1$ | $p_0$ | Entropy | Gain |
|---|---|---|---|---|---|---|---|---|
| Total | | 216 | 59 | 127 | 0.2731 | 0.7268 | 0.8459 | |
| $X_5$ | | | | | | | | 0.0512 |
| | 0 | 177 | 38 | 139 | 0.2146 | 0.7853 | 0.7503 | |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | 1 | 39 | 21 | 18 | 0.5384 | 0.4615 | 0.9957 | |
| $X_6$ | 0 | 116 | 33 | 83 | 0.2844 | 0.7155 | 0.8267 | 0.0005 |
| | 1 | 100 | 26 | 74 | 0.26 | 0.74 | 0.8614 | |
| $X_{10}$ | 0 | 120 | 34 | 86 | 0.2833 | 0.7166 | 0.8273 | 0.0004 |
| | 1 | 96 | 25 | 71 | 0.2604 | 0.7395 | 0.8599 | |
| $X_{14}$ | 0 | 151 | 39 | 112 | 0.2582 | 0.7417 | 0.8241 | 0.0018 |
| | 1 | 65 | 20 | 45 | 0.3076 | 0.6923 | 0.8904 | |

Berdasarkan tabel diatas variabel yang memiliki nilai *gain* tertinggi yaitu $X_5$ (kehilangan berat badan secara drastis/*sudden weight lost*) dengan nilai *gain* sebesar 0.0512. Variabel $X_5$ memiliki 2 katagori yaitu katagori *No* (0) dan *Yes* (1). Pohon untuk cabang pertama pada variabel $X_3$ dapat dilihat berikut ini.



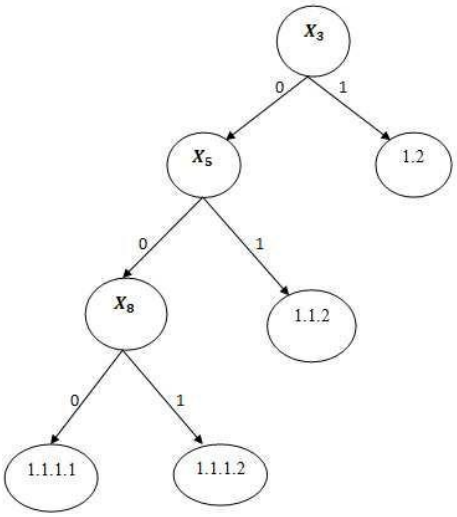

Gambar 4.6 Pohon keputusan *node* 1.1

Selanjutnya mencari cabang dari variabel $X_5$ dengan katagori *No* (0) yaitu dengan mengambil 4 variabel secara acak dari variabel yang tersisa selain dari variabel yang telah digunakan sebagai *node*. Pada node 1.1.1 variabel yang terpilih untuk digunakan yaitu $X_2$, $X_8$ , $X_{11}$, $X_{16}$ . Perhitungan *entropy* dan juga *gain* dapat dilihat pada tabel 4.8 berikut ini.

Tabel 4.10 Perhitungan *entropy* dan juga *gain* untuk variabel $X_2$,$X_8$,$X_{11}$,$X_{16}$

| Atribut | Label | Jumlah | 1 | 0 | $p_1$ | $p_0$ | Entropy | Gain |
|---|---|---|---|---|---|---|---|---|
| Total | | 177 | 38 | 139 | 0.2146 | 0.7853 | 0.7503 | |
| $X_2$ | 0 | 38 | 22 | 16 | 0.5789 | 0.4210 | 0.9819 | 0.1349 |
| | 1 | 139 | 16 | 123 | 0.1151 | 0.8848 | 0.5151 | |
| $X_8$ | 0 | 142 | 30 | 112 | 0.2112 | 0.7887 | 0.7438 | 1.1938 |
| | 1 | 35 | 8 | 27 | 0.2285 | 0.7714 | 0.7755 | |
| $X_{11}$ | 0 | 157 | 28 | 129 | 0.1783 | 0.8216 | 0.6764 | 0.0373 |
| | 1 | 20 | 10 | 10 | 0.5 | 0.5 | 1 | |
| $X_{14}$ | 0 | 162 | 32 | 130 | 0.1975 | 0.8024 | 0.7169 | 0.0118 |
| | 1 | 15 | 6 | 9 | 0.4 | 0.6 | 0.9709 | |

Berdasarkan tabel diatas variabel yang memiliki nilai *gain* tertinggi yaitu $\boldsymbol{X_8}$ (infeksi jamur/*Genital trush*) dengan nilai *gain* sebesar 1.1938. Variabel $\boldsymbol{X_8}$ memiliki 2 katagori yaitu katagori *No* (0) dengan 142 orang dan *Yes* (1) dengan 35 orang. Pohon untuk cabang node 1.1.1 pada variabel $\boldsymbol{X_8}$ dapat dilihat berikut ini.




Gambar 4.7 Pohon keputusan *node* 1.1.1

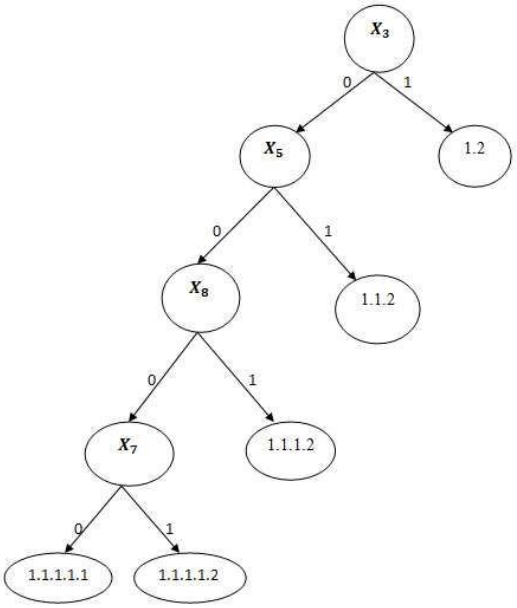
Selanjutnya mencari cabang dari variabel $\boldsymbol{X_8}$ dengan katagori *No* (0) yaitu dengan mengambil 4 variabel secara acak dari variabel yang tersisa selain dari

variabel yang telah digunakan sebagai *node*. Pada node 1.1.1.1 variabel yang terpilih untuk digunakan dalam perhitungan cabang selanjutnya yaitu $\boldsymbol{X_7}$, $\boldsymbol{X_9}$ , $\boldsymbol{X_{11}}$, $\boldsymbol{X_{15}}$ **.** Perhitungan *entropy* dan juga *gain* dapat dilihat pada tabel 4.9 berikut ini.

Tabel 4.11 Perhitungan *entropy* dan juga *gain* untuk variabel $X_7$,$X_9$,$X_{11}$,$X_{15}$

| Atribut | Label | Jumlah | 1 | 0 | $p_1$ | $p_0$ | Entropy | Gain |
|---|---|---|---|---|---|---|---|---|
| Total | | 142 | 30 | 112 | 0.2112 | 0.7887 | 0.7438 | |
| $\boldsymbol{X_7}$ | 0 | 109 | 18 | 91 | 0.1651 | 0.8348 | 0.6464 | 0.0279 |
| | 1 | 33 | 12 | 21 | 0.3636 | 0.63636 | 0.9456 | |
| $\boldsymbol{X_9}$ | 0 | 88 | 13 | 75 | 0.1477 | 0.8522 | 0.6041 | 0.0277 |
| | 1 | 54 | 17 | 37 | 0.3148 | 0.6851 | 0.8986 | |
| $\boldsymbol{X_{11}}$ | 0 | 127 | 25 | 102 | 0.1968 | 0.8031 | 0.7155 | 0.0069 |
| | 1 | 15 | 5 | 10 | 0.3333 | 0.6667 | 0.9182 | |
| $\boldsymbol{X_{15}}$ | 0 | 89 | 24 | 65 | 0.2696 | 0.7303 | 0.8409 | 0.0266 |
| | 1 | 53 | 6 | 47 | 0.1132 | 0.8867 | 0.5095 | |

Berdasarkan tabel diatas variabel yang memiliki nilai *gain* tertinggi yaitu $\boldsymbol{X_7}$ (peningkatan nafsu makan berlebih/*Polyphagia*) dengan nilai *gain* sebesar 0.0279. Variabel $\boldsymbol{X_7}$ memiliki 2 katagori yaitu katagori *No* (0) dengan 109 orang dan *Yes* (1) dengan 33 orang. Pohon untuk cabang node 1.1.1.1 pada variabel $\boldsymbol{X_7}$ dapat dilihat berikut ini.




Gambar 4.8 Pohon keputusan *node* 1.1.1.1

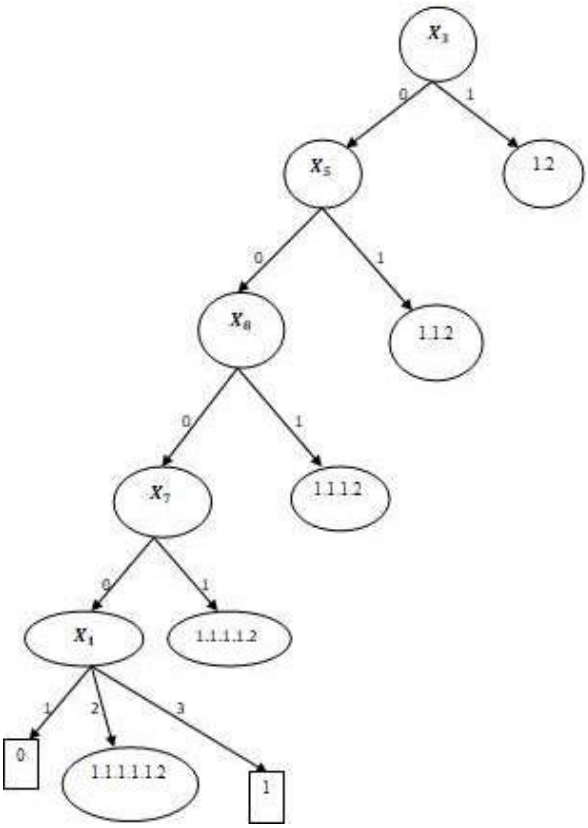
Selanjutnya mencari cabang dari variabel $\boldsymbol{X_7}$ dengan katagori *No* (0) yaitu dengan mengambil 4 variabel secara acak dari variabel yang tersisa selain dari variabel yang telah digunakan sebagai *node*. Pada node 1.1.1.1.1 variabel yang terpilih untuk digunakan dalam perhitungan cabang selanjutnya yaitu $\boldsymbol{X_1}$, $\boldsymbol{X_6}$, $\boldsymbol{X_{12}}$, $\boldsymbol{X_{13}}$. Perhitungan *entropy* dan juga *gain* dapat dilihat pada tabel 4.10 berikut ini.

Tabel 4.12 Perhitungan *entropy* dan juga *gain* untuk variabel $X_1$,$X_6$,$X_{12}$,$X_{13}$

| Atribut | Label | Jumlah | 1 | 0 | $p_1$ | $p_0$ | Entropy | Gain |
|---|---|---|---|---|---|---|---|---|
| Total | | 109 | 18 | 91 | 0.1651 | 0.8348 | 0.6464 | |
| $\boldsymbol{X_1}$ | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 0.1574 |
| | 2 | 102 | 12 | 90 | 0.1176 | 0.8823 | 0.5225 | |
| | 3 | 6 | 6 | 0 | 1 | 0 | 0 | |
| $\boldsymbol{X_6}$ | 0 | 72 | 17 | 55 | 0.2361 | 0.7638 | 0.7885 | 0.0647 |
| | 1 | 37 | 1 | 36 | 0.0270 | 0.9729 | 0.1792 | |
| $\boldsymbol{X_{12}}$ | 0 | 82 | 16 | 66 | 0.1951 | 0.8048 | 0.7120 | 0.0164 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | 1 | 27 | 2 | 25 | 0.0740 | 0.9259 | 0.3809 | |
| $X_{13}$ | 0 | 89 | 10 | 79 | 0.1123 | 0.8876 | 0.5069 | 0.0543 |
| | 1 | 20 | 8 | 12 | 0.4 | 0.6 | 0.9709 | |

Berdasarkan tabel diatas variabel yang memiliki nilai *gain* tertinggi yaitu $X_1$ (umur) dengan nilai *gain* sebesar 0.1574. Variabel $X_1$ memiliki 3 katagori yaitu katagori 1 (umur kurang dari 20 tahun) dengan jumlah 1 orang yang telah masuk dalam klasifikasi tidak menderita diabetes atau negatif, katagori 2 ( umur 20 tahun sampai dengan 40 tahun) dengan jumlah 102 orang dan katagori 3 (umur diatas 40 tahun) dengan jumlah 6 orang yang masuk dalam katagori penderita diabetes atau positif diabetes. Katagori 2 pada variabel $X_1$ belum memberikan hasil seperti 2 katagori lainya, maka akan dilakukan perhitungan untuk cabang selanjutnya. Pohon untuk cabang node 1.1.1.1.1 pada variabel $X_1$ dapat dilihat berikut ini.

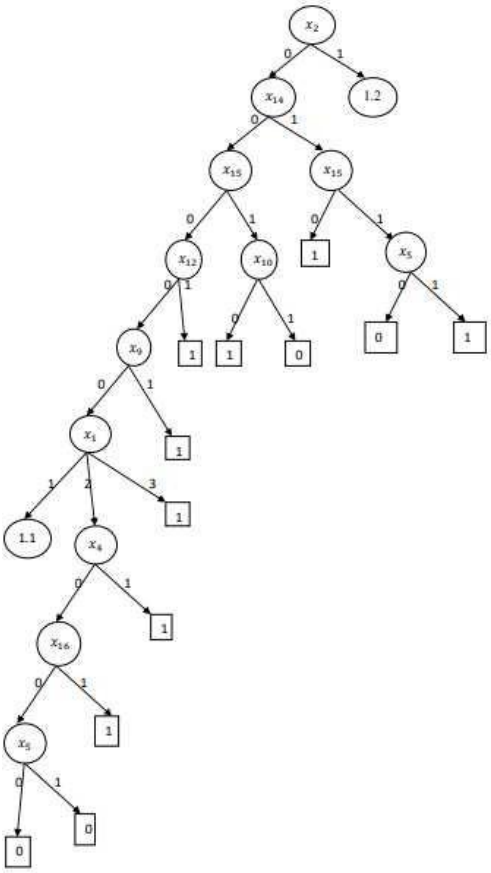



Gambar 4.9 Pohon keputusan pertama

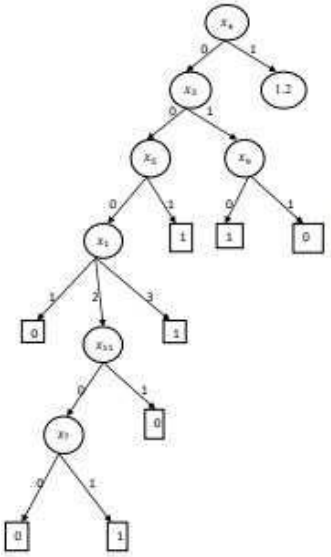
Proses perhitungan dilakukan sampai semua cabang sudah pada *terminal node* atau tidak ada lagi cabang yang tesisa. Setelah membentuk 1 pohon maka akan terlihat model pengkondisian yang akan digunakan sebagai penentu prediksi klasifikasi pada data *test.* Model pengkondisian dapat dilihat pada **Lampiran.**

Langkah selanjutnya yaitu mencari pohon keputusan dari pohon ke 2 sampai dengan pohon ke 1820 dengan cara perhitungan yang sama. Pertama, melakukan *bootstrap* pada data *training* untuk setiap pohon. kemudian melakukan pengambilan 4 variabel secara acak untuk menentukan node. Dari ke 4 variabel tersebut akan dilakukan perhitungan *entropy* dan juga *gain*. Variabel yang

memiliki nilai *gain* tertinggi akan menjadi root node/ akar pohon. Ulangi proses perhitungan sampai semua cabang pohon selesai dicari atau telah mencapai *terminal node.* Beberapa pohon keputusan dari *random forest* dapat dilihat berikut ini :




Gambar 4.10 Pohon keputusan kedua



Gambar 4.11 Pohon keputusan ketiga

Semua pohon yang telah dibangun akan dibuat pengkodisian sehingga dapat dilakukan proses klasifikasi dengan cara *majorty voting*. Setiap data *test* memiliki 1820 klasifikasi berdasarkan pengkondisian dari 1820 pohon. Penggabungan dari masing-masing pohon keputusan pada setiap data *test* merupkan hasil klasifikasi akhir. Hasil klasifikasi akhir dari 1820 pohon dengan menggunakan *Random Forest* dapat dilihat pada berikut ini.

Tabel 4.13 *confusion matrix* metode *Random Forest*

| | *True* 1 | *True* 0 |
|---|---|---|
| *Predict* 1 | 62 | 0 |
| *Predict* 0 | 2 | 40 |

Berdasarkan tabel yang ada diatas dapat dilihat bahwa 62 orang di prediksi benar masuk dalam klasifikasi positif diabetes, sedangkan 40 orang diprediksi

benar masuk dalam klasifikasi negatif diabetes. Terdapat 2 orang di prediksi masuk dalam katagori negatif diabetes, namun ternyata masuk dalam katogori positif diabetes pada data sebenarnya. Sedangkan tidak ada orang yang diprediksi masuk dalam katagori positif diabetes dan negatif diabetes pada data sebenarnya. Kita juga dapat melihat tingkat akurasi, presisi, recall, specificity, dan F1 score dari tabel diatas berikut ini.

$$akurasi = \frac{62 + 40}{104} = 0.9808$$

$$presisi = \frac{62}{62 + 0} = 1$$

$$recall = \frac{62}{62 + 2} = 0.9688$$

$$specificity = \frac{40}{40 + 0} = 1$$

$$F1\ score = 2\left(\frac{0.9687 * 1}{0.9687 + 1}\right) = 0.9841$$

Jadi, dengan menggunakan metode *random forest* akurasi yang didapatkan sebesar 0.9808, presisi sebesar 1. Recall sebesar 0.9687, specificity sebesar 1 dan F1 score sebesar 0.9841, hal ini menunjukkan bahwa ketepatan akurasi dalam memprediksi klasifikasi data penyakit diabetes dengan menggunakan metode *random forest* adalah 98.08%.

## 4.4 Perbandingan Tingkat Ketepatan Klasifikasi

Berdasarkan perhitungan yang telah dilakukan maka didapatkan hasil perbandingan akurasi dari kedua metode yang dapat dilihat pada tabel berikut ini.

Tabel 4.14 perbandingan akurasi 2 metode

| Tingkat ketepatan | Metode Decision Tree C4.5 | Metode Random Forest |
|---|---|---|
| Akurasi | 91.35% | 98.08%. |
| Presisi | 93.55% | 100% |
| Recall | 92.06% | 96.88% |
| Specificity | 90.24% | 100% |
| F1 score | 92.80% | 98.41% |

Dari perbandingan tingkat akurasi tersebut, terlihat bahwa metode *decision tree* C4.5 memiliki akurasi sebesar 91.35% sedangkan pada metode *random forest* akurasi yang didapatkan lebih tinggi yaitu sebesar 98.08%. Begitupun dengan nilai dari presisi, recall, specificity dan juga Fi score pada *metode random forest* lebih tinggi dibandingkan dengan nilai dari metode *decision tree* C4.5. Jika variabel umur pada data penyakit diabetes tidak di diskritisasi pada metode *decision tree* maka nilai akurasi, presisi, recall, specificity dan F1 score yang didapat secara berturut-turut sebesar 85.58%, 86.18%, 85.71%, 85.58%, dan 85.77%. Sedangkan pada metode *random forest* nilai akurasi, presisi, recall, specificity, dan F1 score untuk variabel umur tanpa diskritisasi secara berturut-turut sebesar 75.56%, 76.55%, 78.92%, 77.65%, dan 76.11%. Untuk data

prediksi penyakit diabetes dari pasien *Diabetes Sylhet* Rumah sakit di Sylhet, Bangladesh metode *random forest* dinilai lebih baik dibandingkan dengan metode *decision tree*. Metode *Random Forest* memiliki tingkat akurasi yang lebih baik karena menggabungkan hasil dari klasifikasi masing-masing pohon kemudian dilakukan *majorty voting* sehingga dapat mengurangi tingkat kesalahan pada proses prediksi klasifikasi.