Volume 1 No 1 (2022) 41-50
http://jurnal.fmipa.unmul.ac.id/index.php/basis

Klasifikasi Penyakit Diabetes Melitus Menggunakan Algoritma *Naïve* Bayes *Classifier*

Latifah Uswatun Khasanah[1,*], Yuki Novia Nasution[1], Fidia Deny Tisna Amijaya[1]

[1] *Laboratorium Matematika Komputasi Program Studi Matematika Jurusan Matematika FMIPA Universitas Mulawarman*

Dikirim: Mei 2022; Diterima: Mei 2022; Dipublikasi: September 2022

Alamat Email Korespondesi: latifah.khasanah7@gmail.com

Abstrak

Klasifikasi merupakan sebuah teknik analisis data yang mengekstrak model untuk mendeskripsikannya ke dalam kelas tertentu. Salah satu algoritma yang dapat digunakan untuk klasifikasi adalah algoritma *Naïve* Bayes *Classifier.* Algoritma *Naïve* Bayes *Classifier* merupakan salah sebuah metode klasifikasi yang memprediksi peluang di masa depan berdasarkan pengalaman di masa sebelumnya dengan menggunakan Teorema Bayes. Tujuan penelitian ini adalah untuk mengetahui hasil klasifikasi pasien ke dalam dua kategori diagnosis diabetes melitus yaitu 'Ya' dan 'Tidak' menggunakan algoritma *Naïve* Bayes *Classifier* dan mengetahui tingkat akurasi dari empat proporsi data yaitu 60:40, 70:30, 80:20 dan 90:10. Sampel pada penelitian ini adalah data pasien di RS Dirgahayu Samarinda Tahun 2018 s/d 2021 sebanyak 130 data pasien. Berdasarkan hasil analisis, pada proporsi data *testing* 40% dan 30% masing-masing terdapat 4 pasien hasil klasifikasinya tidak tepat. Pada proporsi data *testing* 20% dan 10% masing-masing terdapat 2 data pasien hasil klasifikasinya tidak tepat. Adapun tingkat akurasi pada proporsi data *testing* 40%, 30%, 20% dan 10% secara berurutan adalah sebesar 92,31%; 89,74%; 92,31%; dan 84,62%. Berdasarkan tingkat akurasi yang telah diketahui, nilai akurasi terbaik adalah pada proporsi data *testing* 40% dan 20% dengan nilai akurasi sebesar 92,31%.

Kata Kunci:
Akurasi, diabetes melitus, klasifikasi, Naïve Bayes Classifier

PENDAHULUAN

Penyakit tidak menular (PTM) merupakan penyebab kematian yang tinggi di dunia. Kasus PTM paling banyak terjadi di negara yang berpenghasilan rendah. Pada tahun 2016 terdapat sekitar 36 juta jiwa atau sekitar 71% kematian yang disebabkan oleh PTM. Diabetes melitus merupakan salah satu bagian dari penyakit tidak menular (PTM) yang kasusnya cukup tinggi di Indonesia. Penyakit diabetes melitus adalah suatu penyakit metabolik dimana pankreas tidak cukup memproduksi insulin atau sel-sel dalam tubuh tidak dapat menggunakan insulin yang diproduksi dengan efektif [1].

World Health Organization (WHO) memprediksi kenaikan jumlah penderita diabetes di Indonesia dari 8,4 juta jiwa pada tahun 2000 menjadi sekitar 21,3 juta jiwa pada tahun 2030. Hasil Riset Kesehatan Dasar (Riskesdas) yang dilaksanakan pada tahun 2018 melakukan pengumpulan data penderita diabetes melitus pada penduduk berumur ≥ 15 tahun menunjukkan bahwa hasil prevalensi sebesar 2%.

Angka ini menunjukkan peningkatan prevalensi diabetes melitus pada tahun 2013 sebesar 1,5%. Peningkatan prevalensi diabetes melitus ditunjukkan oleh hampir semua provinsi di Indonesia. Empat provinsi dengan prevalensi tertinggi pada tahun 2013 dan 2018 yaitu DI Yogyakarta, DKI Jakarta, Sulawesi Utara dan Kalimantan Timur [2].

Pada zaman modern dimana pertukaran informasi dapat terjadi secara cepat menjadi salah satu penyebab meningkatnya jumlah data. Kecanggihan teknologi berperan penting dalam beberapa bidang salah satuya adalah bidang kesehatan. Pada bidang kesehatan dibutuhkan sistem atau alat yang dapat mendiagnosa atau memprediksi penyakit berdasarkan faktor-faktor pertimbangan tertentu. Dari sekian banyak data yang terkumpul di rumah sakit atau fasilitas layanan kesehatan lainnya dapat digunakan untuk memprediksi suatu penyakit menggunakan teknik data *mining.*

Data *mining* dapat diartikan sebagai proses penambangan data yang menghasilkan sebuah *output* (keluaran) berupa pengetahuan. Data *mining* terdiri dari gabungan beberapa bidang keilmuan seperti teknik pembelajaran mesin, *artificial intelligence*, statistika dan sistem basis data dan ilmu-ilmu lainnya. Terdapat beberapa pekerjaan yang berkaitan dengan data *mining*, antara lain analisis klaster, analisis asosiasi, deteksi anomali dan model prediksi. Model prediksi terdiri dari dua jenis, yaitu klasifikasi dan regresi [3].

Klasifikasi merupakan sebuah teknik analisis data yang mengekstrak model untuk mendeskripsikannya ke dalam kelas tertentu. Terdapat beberapa algoritma klasifikasi yang dapat digunakan, salah satunya adalah algoritma *Naïve* Bayes *Classifier. Naïve* Bayes *Classifier* merupakan salah satu algoritma klasifikasi yang menggunakan memprediksi peluang di masa depan berdasarkan pengalaman di masa sebelumnya, atau yang lebih dikenal dengan teorema Bayes [4].

Penelitian terkait algoritma *Naïve* Bayes *Classifier* telah dilakukan oleh beberapa peneliti sebelumnya. [5] melakukan penelitian mengenai aplikasi metode *Naïve* Bayes dalam prediksi risiko penyakit Jantung melakukan dua kali percobaan dengan menggunakan data *testing* sebanyak 25 data dan 50 data. [6] melakukan penelitian terkait penerapan algoritma *Naïve* Bayes untuk prediksi penerimaan siswa baru. Sementara itu [7] menerapkan metode *Naïve* Bayes dalam implementasi *text mining*.

Berdasarkan latar belakang tersebut, maka penulis tertarik melakukan penelitian yaitu klasifikasi terhadap data pasien di RS Dirgahayu Samarinda dengan menggunakan Algoritma *Naïve* Bayes *Classifier*. Pada penelitian ini menggunakan sebanyak enam variabel yaitu usia, jenis kelamin, status merokok, kadar glukosa, tekanan darah, dan kelas.

METODE PENELITIAN

Penelitian ini termasuk ke dalam penelitian kuantitatif dengan populasi yang digunakan adalah seluruh pasien di RS Dirgahayu Samarinda. Adapun sampel yang digunakan adalah pasien di RS Dirgahayu Samarinda bulan September Tahun 2018 s/d 2021. Pada penelitian ini tahapan analisis data yang dilakukan adalah sebagai berikut:

1. Pembersihan data
   Pembersihan data merupakan proses persiapan data dengan cara memeriksa apakah terdapat *missing value* atau data yang tidak konsisten.

2. Randomisasi data
   Randomisasi data merupakan proses pengacakan data agar diperoleh sampel representatif yang mewakili populasi.
3. Analisis deskriptif
   Analisis deskriptif adalah salah satu teknik analisis data kuantitatif yang digunakan untuk menganalisis data dengan cara mendeskripsikan data yang telah terkumpul.
4. Klasifikasi *Naïve* Bayes
   Penelitian ini menggunakan penerapan algoritma *Naïve* Bayes dengan langkah-langkah sebagai berikut:
   - Membaca data *training*. Penelitian ini menggunakan empat proporsi data *training* dan data *testing* yaitu 60:40, 70:30, 80:20 dan 90:10.
   - Menghitung nilai *prior* data *training*.
   - Menghitung nilai probabilitas setiap variabel terhadap setiap kelasnya dengan menggunakan rumus probabilitas bersyarat sebagai berikut

$$\boldsymbol{P(H|X) = \frac{P(X \cap H)}{P(X)}} \quad (1)$$

     dengan:
     $P(H|X)$ = probabilitas terjadinya H dengan syarat X telah terjadi
     $P(X \cap H)$ = probabilitas awal H dengan petunjuk X telah terjadi secara simultan
     $P(X)$ = probabilitas terjadinya X
   - Menghitung nilai akumulasi probabilitas dari setiap kelas menggunakan persamaan berikut

$$\boldsymbol{P(C_i|X) = \prod_{k=1}^{n} P(x_k|C_i) = P(x_1|C_i) \times P(x_2|C_i) \times \ldots \times P(x_n|C_i)} \quad (2)$$

   - Menghitung perkalian probabilitas *prior* dengan nilai akumulasi probabilitas variabel bebas dari masing-masing kelas.
   - Menentukan hasil klasifikasi obyek berdasarkan nilai paling maksimum dari kedua kelas.
5. Evaluasi model klasifikasi
   Evaluasi model klasifikasi bertujuan untuk mengetahui tingkat akurasi model klasifikasi yang telah dibuat dengan rumus

$$\boldsymbol{Accuracy = \frac{TP + TN}{P + N}} \quad (3)$$

   dengan keterangan sebagai berikut

   1) *TP* atau *True Positives* adalah jumlah *tuple* positif yang dilabeli dengan benar oleh *classifier*. Tuple positif adalah *tuple* aktual yang berlabel positif, seperti *tuple* dengan label Bonus = 'Ya'.
   2) *TN* atau *True Negatives* adalah jumlah *tuple* negatif yang dilabeli dengan benar oleh *classifier*. *Tuple* negatif adalah *tuple* aktual yang berlabel negatif, seperti *tuple* dengan label Bonus = 'Tidak'.
   3) *FP* atau *False Positives* adalah jumlah *tuple* negatif yang salah dilabeli oleh *classifier*. Misalnya, sebuah *tuple* pelanggan yang berlabel Bonus = 'Tidak' akan tetapi oleh *classifier* dilabeli Bonus = 'Ya'.

4) *FN* atau *False Negatives* adalah jumlah *tuple* positif yang salah dilabeli oleh *classifier*. Misalnya, sebuah *tuple* pelanggan yang berlabel Bonus = 'Ya' akan tetapi oleh *classifier* dilabeli Bonus = 'Tidak'.
Empat istilah di atas dapat digambarkan sebagai *confusion matrix* seperti yang diilustrasikan pada Tabel 1 berikut

Tabel 1. *Confusion Matrix*

No.	Kelas Aktual	Kelas Hasil Prediksi		
		Ya	Tidak	Jumlah
1	Ya	*TP*	*FN*	*P*
2	Tidak	*FP*	*TN*	*N*
3	Jumlah	P'	N'	$P + N$

HASIL DAN PEMBAHASAN

Data yang digunakan dalam penelitian ini berjumlah 130 data dengan enam variabel yaitu usia, jenis kelamin, status merokok, kadar glukosa, tekanan darah dan kelas.

Teknik analisis data yang dilakukan pertama kali adalah menggunakan teknik analisis deskriptif untuk mengetahui gambaran dari keseluruhan data sampel yang digunakan. Langkah selanjutnya adalah membagi data *training* dan data *testing* sesuai dengan masing-masing proporsi yang dapat dilihat pada Tabel 2 berikut.

Tabel 2. Jumlah Data *Training* dan Data *Testing*

No.	Proporsi Data *Training* dan Data *Testing*	Jumlah Data *Training*	Jumlah Data *Testing*	Total Data
1.	60 : 40	78	52	130
2.	70 : 30	91	39	130
3.	80 : 20	104	26	130
4.	90 : 10	117	13	130

Algoritma *Naïve* Bayes *Classifier*

Pada bagian sebelumnya telah diketahui bahwa proporsi data *training* dan data *testing* yang digunakan adalah 60:40, 70:30, 80:20 dan 90:10. Adapun data *training* proporsi data 60:40 dapat dilihat pada Tabel 3.

Tabel 3. Data *Training* Proporsi 60:40

No.	F_1	F_2	F_3	F_4	F_5	C
1.	MENENGAH	P	TIDAK	TERINDIKASI	HIPERTENSI	YA
2.	MENENGAH	P	TIDAK	TIDAK TERINDIKASI	TIDAK HIPERTENSI	TIDAK
3.	MENENGAH	L	TIDAK	TERINDIKASI	HIPERTENSI	YA
…	…	…	…	…	…	…
…	…	…	…	…	…	…
78.	MENENGAH	P	TIDAK	TERINDIKASI	TIDAK HIPERTENSI	YA

dengan:
F_1 = Usia (tahun)
F_2 = Jenis Kelamin
F_3 = Status Merokok

F_4 = Kadar Glukosa
F_5 = Tekanan Darah
C = Kelas

Berikut ini adalah contoh perhitungan manual algoritma *Naïve* Bayes *Classifier* dengan menggunakan proporsi data *training* sebesar 60% dan data *testing* sebesar 40%.

1. Perhitungan Nilai Probabilitas *Prior* dari Setiap Kelas

Pada proporsi 60% data *training* dan 40% data *testing* diperoleh jumlah data *training* sebanyak 78 orang dan data *testing* sebanyak 52 orang. Langkah awal klasifikasi menggunakan algoritma *Naïve* Bayes *Classifier* adalah menghitung nilai probabilitas *prior* pada kedua kelas dalam data *training* menggunakan rumus

$$P(C_i) = \frac{n_i}{N} \quad (4)$$

a) kelas pertama (pasien RS Dirgahayu dengan diagnosa penyakit diabetes "Ya")
Pada 78 data *training* yang ada, terdapat 28 pasien memiliki diagnosa penyakit diabetes melitus sehingga nilai probabilitas *prior* kelas pertama adalah

$$P(C_1) = \frac{28}{78} = 0{,}359$$

b) kelas kedua (pasien RS Dirgahayu dengan diagnosa penyakit diabetes "Tidak")
Pada 78 data *training* yang ada, terdapat 50 pasien tidak memiliki diagnosa penyakit diabetes melitus sehingga nilai probabilitas *prior* kelas kedua adalah

$$P(C_2) = \frac{50}{78} = 0{,}641$$

2. Perhitungan Nilai Probabilitas Setiap Variabel Terhadap Setiap Kelas

Perhitungan nilai probabilitas setiap variabel pada kedua kelas berdasarkan masing-masing probabilitas dari data *training* menggunakan Persamaan (1). Data *testing* yang digunakan adalah data pasien ke-1 yang dapat dilihat pada Tabel 4.

Tabel 4. Data *Testing* Pertama Proporsi 60:40

No.	F_1	F_2	F_3	F_4	F_5	C
1.	Menengah	Perempuan	Tidak	Tidak Terindikasi	Hipertensi	?

- **Usia (F_1)**

Pada variabel usia terdapat 1 pasien kategori usia "Muda", 66 pasien kategori usia "Menengah", dan 11 pasien kategori usia "Tua". Data *testing* pertama menunjukkan kategori usia "Menengah", sehingga kategori usia inilah yang digunakan dalam perhitungan probabilitas bersyarat. Diketahui bahwa dari 66 pasien kategori usia "Menengah", terdapat 22 pasien berada di dalam kelas "Ya" dan 44 pasien berada di dalam kelas "Tidak".

Nilai probabilitas variabel usia (F_1) dengan kategori "Menengah" terhadap kelas "Ya" melalui perhitungan sebagai berikut:

$$P(F_1 = \text{"Menengah"} \mid C = \text{"Ya"}) = \frac{22}{66} = 0{,}333$$

sehingga nilai probabilitas variabel usia (F_1) dengan kategori "Menengah" pada kelas "Ya" adalah 0,333.

Nilai probabilitas variabel usia (F_1) dengan kategori "Menengah" pada kelas "Tidak" melalui perhitungan sebagai berikut:

$$P(F_1 = \text{"Menengah"} \mid C = \text{"Tidak"}) = \frac{44}{66} = 0{,}667$$

sehingga nilai probabilitas variabel usia (F_1) dengan kategori "Menengah" pada kelas "Tidak" adalah 0,667.

- **Jenis Kelamin** ($\boldsymbol{F_2}$)

Pada variabel jenis kelamin terdapat 46 pasien dengan jenis kelamin "Laki-laki" dan 32 pasien dengan jenis kelamin "Perempuan". Data *testing* pertama menunjukkan kategori jenis kelamin "Perempuan", sehingga kategori jenis kelamin inilah yang digunakan dalam perhitungan probabilitas bersyarat. Diketahui bahwa dari 46 pasien kategori jenis kelamin "Perempuan" terdapat 15 pasien berada di dalam kelas "Ya" dan 31 pasien berada di dalam kelas "Tidak".

Nilai probabilitas variabel jenis kelamin (F_2) dengan kategori "Perempuan" pada kelas "Ya" melalui perhitungan sebagai berikut:

$$P(F_2 = \text{"Perempuan"} \mid C = \text{"Ya"}) = \frac{15}{46} = 0{,}326$$

sehingga nilai probabilitas variabel jenis kelamin (F_2) dengan kategori "Perempuan" pada kelas "Ya" adalah 0,326.

Nilai probabilitas variabel jenis kelamin (F_2) dengan kategori "Perempuan" pada kelas "Tidak" melalui perhitungan sebagai berikut:

$$P(F_2 = \text{"Perempuan"} \mid C = \text{"Tidak"}) = \frac{31}{46} = 0{,}674$$

sehingga nilai probabilitas variabel jenis kelamin (F_2) dengan kategori "Perempuan" pada kelas "Tidak" adalah 0,674.

- **Status Merokok** ($\boldsymbol{F_3}$)

Pada variabel status merokok terdapat 77 pasien dengan status merokok "Tidak" dan 1 pasien dengan status merokok "Ya". Data *testing* pertama menunjukkan kategori status merokok "Tidak", sehingga kategori status merokok inilah yang digunakan dalam perhitungan probabilitas bersyarat. Diketahui bahwa dari 77 pasien kategori status merokok "Tidak" terdapat 28 pasien berada di dalam kelas "Ya" dan 49 pasien berada di dalam kelas "Tidak".

Nilai probabilitas variabel status merokok (F_3) dengan kategori "Tidak" pada kelas "Ya" melalui perhitungan sebagai berikut:

$$P(F_3 = \text{"Tidak"} \mid C = \text{"Ya"}) = \frac{28}{77} = 0{,}364$$

sehingga nilai probabilitas variabel status merokok (F_3) dengan kategori "Tidak" pada kelas "Ya" adalah 0,364.

Nilai probabilitas variabel status merokok (F_3) dengan kategori "Tidak" pada kelas "Tidak" melalui perhitungan sebagai berikut:

$$P(F_3 = \text{"Tidak"} \mid C = \text{"Tidak"}) = \frac{49}{77} = 0{,}636$$

sehingga nilai probabilitas variabel status merokok (F_3) dengan kategori "Tidak" pada kelas "Tidak" adalah 0,636.

- **Kadar Glukosa** ($\boldsymbol{F_4}$)

Pada variabel kadar glukosa terdapat 56 pasien dengan kategori "Tidak terindikasi" dan 22 pasien dengan kategori "Terindikasi". Data *testing* pertama menunjukkan kategori "Tidak terindikasi", sehingga kategori kadar glukosa inilah yang digunakan dalam perhitungan probabilitas bersyarat. Diketahui bahwa dari 56 pasien kategori "Tidak terindikasi" terdapat 7 pasien berada di dalam kelas "Ya" dan 49 pasien berada di dalam kelas "Tidak".

Nilai probabilitas variabel kadar glukosa (F_4) dengan kategori "Tidak Terindikasi" pada kelas "Ya" melalui perhitungan sebagai berikut:

$$P(F_4 = \text{"Tidak Terindikasi"} \mid C = \text{"Ya"}) = \frac{7}{56} = 0{,}125$$

sehingga nilai probabilitas variabel kadar glukosa (F_4) dengan kategori "Tidak Terindikasi" pada kelas "Ya" adalah 0,125.

Nilai probabilitas variabel kadar glukosa (F_4) dengan kategori "Tidak Terindikasi" pada kelas "Tidak" melalui perhitungan sebagai berikut:

$$P(F_4 = \text{"Tidak Terindikasi"} \mid C = \text{"Tidak"}) = \frac{49}{56} = 0{,}875$$

sehingga nilai probabilitas variabel kadar glukosa (F_4) dengan kategori "Tidak Terindikasi" pada kelas "Tidak" adalah 0,875.

- **Tekanan Darah (F_5)**

Pada variabel tekanan darah terdapat 23 pasien dengan kategori "Hipertensi" dan 55 pasien dengan kategori "Tidak hipertensi". Data *testing* pertama menunjukkan kategori tekanan darah "Hipertensi", sehingga kategori tekanan darah inilah yang digunakan dalam perhitungan probabilitas bersyarat. Diketahui bahwa dari 23 pasien kategori tekanan darah "Hipertensi" terdapat 14 pasien berada di dalam kelas "Ya" dan 9 pasien berada di dalam kelas "Tidak".

Nilai probabilitas variabel tekanan darah (F_5) dengan kategori "Hipertensi" pada kelas "Ya" melalui perhitungan sebagai berikut:

$$P(F_5 = \text{"Hipertensi"} \mid C = \text{"Ya"}) = \frac{14}{23} = 0{,}609$$

sehingga nilai probabilitas variabel tekanan darah (F_5) dengan kategori "Hipertensi" pada kelas "Ya" adalah 0,609.

Nilai probabilitas variabel tekanan darah (F_5) dengan kategori "Hipertensi" pada kelas "Tidak" melalui perhitungan sebagai berikut:

$$P(F_5 = \text{"Hipertensi"} \mid C = \text{"Tidak"}) = \frac{9}{23} = 0{,}391$$

sehingga nilai probabilitas variabel tekanan darah (F_5) dengan kategori "Hipertensi" pada kelas "Tidak" adalah 0,391.

3. Perhitungan Nilai Akumulasi Probabilitas dari Setiap Kelas

Jika bagian sebelumnya telah diketahui probabilitas masing-masing variabel pada setiap kelas, maka langkah selanjutnya adalah menghitung nilai akumulasi dari setiap kelas dengan cara mengalikan semua probabilitas pada kelas. Adapun perhitungan nilai akumulasi probabilitas berdasarkan Persamaan (2) dapat dilihat sebagai berikut:

- kelas pertama (pasien RS Dirgahayu dengan diagnosa penyakit diabetes "Ya")

$$\prod_{k=1}^{k=5} P(F_k|C_1) = P(F_1|\,C_1) \times P(F_2|\,C_1) \times P(F_3|\,C_1) \times P(F_4|\,C_1) \times P(F_5|\,C_1)$$
$$= 0{,}333 \times 0{,}326 \times 0{,}364 \times 0{,}125 \times 0{,}609$$
$$= 0{,}003$$

sehingga diperoleh nilai akumulasi probabilitas semua variabel ada kelas "Ya" adalah 0,003.

- kelas kedua (pasien RS Dirgahayu dengan diagnosa penyakit diabetes "Tidak")

$$\prod_{k=1}^{k=5} P(F_k|C_2) = P(F_1|\,C_2) \times P(F_2|\,C_2) \times P(F_3|\,C_2) \times P(F_4|\,C_2) \times P(F_5|\,C_2)$$

$$= 0{,}667 \times 0{,}674 \times 0{,}636 \times 0{,}875 \times 0{,}391$$
$$= 0{,}098$$

sehingga diperoleh nilai akumulasi probabilitas semua variabel ada kelas "Tidak" adalah 0,098.

4. Perhitungan Perkalian Probabilitas Prior dengan Nilai Akumulasi Probabilitas dari Masing-masing Kelas

Nilai akumulasi probabilitas dari masing-masing kelas selanjutnya perlu dikalikan dengan nilai probabilitas *prior* setiap kelas yang telah dihitung pada langkah awal.

- kelas pertama (pasien RS Dirgahayu dengan diagnosa penyakit diabetes "Ya"). Diketahui perhitungan nilai probabilitas *prior* dan nilai akumulasi probabilitas dari kelas pertama pada proporsi data 60:40 adalah sebagai berikut:

  $P(C_1) = 0{,}359$

  $\left(\prod_{k=1}^{k=5} P(F_k|C_1)\right) = 0{,}003$

  Hasil perhitungan perkalian probabilitas *prior* dengan nilai akumulasi probabilitas pada kelas pertama adalah sebagai berikut:

$$P(C_1|F) = P(C_1) \times \left(\prod_{k=1}^{k=5} P(F_k|C_1)\right)$$
$$= 0{,}359 \times 0{,}003$$
$$= 0{,}00108$$

- kelas kedua (pasien RS Dirgahayu dengan diagnosa penyakit diabetes "Tidak"). Diketahui perhitungan nilai probabilitas *prior* dan nilai akumulasi probabilitas dari kelas kedua pada proporsi data 60:40 adalah sebagai berikut:

  $P(C_2) = 0{,}641$

  $\left(\prod_{k=1}^{k=5} P(F_k|C_2)\right) = 0{,}098$

  Hasil perhitungan perkalian probabilitas *prior* dengan nilai akumulasi probabilitas pada kelas kedua adalah sebagai berikut:

$$P(C_2|F) = P(C_2) \times \left(\prod_{k=1}^{k=5} P(F_k|C_2)\right)$$
$$= 0{,}641 \times 0{,}098$$
$$= 0{,}06282$$

5. Penentuan Hasil Klasifikasi Obyek Berdasarkan Nilai Paling Maksimum dari Kedua Kelas Kelas

Pada data *testing* pertama nilai probabilitas *posterior* pasien pada kelas "Tidak" sebesar 0,06282 lebih maksimum dibandingkan dengan pasien pada kelas "Ya" sebesar 0,00108, dengan demikian data *testing* pertama yang tertera pada Tabel 3 diklasifikasikan masuk ke dalam kelas kedua yaitu status "TIDAK" terdiagnosis diabetes melitus.

Pada data *testing* selanjutnya menggunakan cara perhitungan yang sama dengan di atas, sehingga hasil klasifikasi proporsi data 60:40 dapat dilihat pada Tabel 5.

Tabel 5. Hasil Klasifikasi Proporsi Data *Testing* 40%

No.	F_1	F_2	F_3	F_4	F_5	C
1.	Menengah	Perempuan	Tidak	Tidak Terindikasi	Hipertensi	TIDAK
2.	Menengah	Perempuan	Tidak	Tidak Terindikasi	Tidak Hipertensi	TIDAK
3.	Menengah	Laki-laki	Tidak	Terindikasi	Hipertensi	YA
…	…	…	…	…	…	…
10.	Menengah	Laki-laki	Ya	Tidak Terindikasi	Hipertensi	TIDAK
…	…	…	…	…	…	…
21.	Menengah	Laki-laki	Tidak	Tidak Terindikasi	Tidak Hipertensi	YA
…	…	…	…	…	…	…
25.	Menengah	Laki-laki	Tidak	Terindikasi	Tidak Hipertensi	YA
…	…	…	…	…	…	…
45.	Menengah	Laki-laki	Tidak	Terindikasi	Tidak Hipertensi	YA
…	…	…	…	…	…	…
52.	Menengah	Perempuan	Tidak	Tidak Terindikasi	Tidak Hipertensi	TIDAK

Pada Tabel 5 menunjukkan hasil klasifikasi proporsi data *testing* 40% terdapat 4 objek yang salah klasifikasi yaitu pada data ke-10, 21, 25 dan 45.

Pada proporsi data lainnya yaitu proporsi data 70:30, 80:20 dan 90:10 menggunakan cara perhitungan manual yang sama dengan cara perhitungan pada proporsi data 60:40.

Evaluasi Model

Model klasifikasi yang telah dibuat selanjutnya dievaluasi menggunakan nilai akurasi yang dapat dihitung dengan menggunakan rumus (3) yang direpresentasikan menggunakan *confusion matrix* sehingga diperoleh nilai akurasi pada proporsi data 60:40 adalah

$$confusion\ matrix = \begin{bmatrix} 17 & 2 \\ 2 & 31 \end{bmatrix}$$

sehingga nilai akurasinya adalah

$$\begin{aligned} akurasi &= \frac{17+31}{19+33} \\ &= \frac{48}{52} \\ &= 0{,}9231 \\ &= 92{,}31\% \end{aligned}$$

Dengan menggunakan cara yang sama, hasil nilai akurasi dari masing-masing proporsi data dapat dilihat pada Tabel 6 berikut

Tabel 6. Hasil Nilai Akurasi

No.	Proporsi Data		Nilai Akurasi
	Data *Training*	Data *Testing*	
1.	60%	40%	92,31%
2.	70%	30%	89,74%
3.	80%	20%	92,31%
4.	90%	10%	84,62%

PENUTUP

Berdasarkan hasil analisis diperoleh hasil klasifikasi penyakit diabetes melitus pada pasien RS Dirgahayu Samarinda bulan September Tahun 2018 s/d 2021 dengan

empat proporsi data. Pada proporsi data 60:40 memiliki nilai akurasi sebesar 92,31 %. Pada proporsi data 70:30 memiliki nilai akurasi sebesar 89,74%. Pada proporsi data 80:20 memiliki nilai akurasi sebesar 92,31%. Pada proporsi data 90:10 memiliki nilai akurasi sebesar 84,62%.

DAFTAR PUSTAKA

[1] Kementerian Kesehatan RI. (2019). *Buku Pedoman Manajemen Penyakit Tidak Menular*. Jakarta: Direktorat Jenderal Pencegahan dan Pengendalian Penyakit.

[2] Kementerian Kesehatan RI. (2020). *Tetap Produktif, Cegah, dan Atasi Diabetes Melitus*. Jakarta: Pusat Data dan Informasi.

[3] Suyanto. (2017). *Data Mining untuk Klasifikasi dan Klasterisasi Data*. Bandung: Informatika.

[4] Han, J, Kamber, M, Pei, J. (2012). *Data Mining: Concepts and Techniques. Third Edition.* USA: Elsevier.

[5] Sabransyah, M., Nasution, Y. N, Amijaya, F. D. T. (2017). Aplikasi Metode *Naïve* Bayes dalam Memprediksi Risiko Penyakit Jantung. *Jurnal Eksponensial,* 111-117.

[6] Lutfi, M., dan Rizal, S. (2018). Penerapan Algoritma Naïve Bayes untuk Prediksi Penerimaan Siswa Baru di SMK Al-Amien Wonorejo. *Jurnal Explore It,* 14-17.

[7] Nangi, J., Ransi, N., Apriliana. (2017). Implementasi *Text Mining* Klasifikasi Skripsi Menggunakan Metode *Naïve* Bayes *Classifier*. *Jurnal Semantik,* 187-194.