p-ISSN 2085-7829
e-ISSN 2798-3455

Klasifikasi *Naïve* Bayes Pada Data Status Kesejahteraan Rumah Tangga Penerima Manfaat di Kecamatan Samarinda Ilir Tahun 2023

Naïve Bayes Classification of Welfare Status Data of Beneficiary Households in Samarinda Ilir District in 2023

Indah Cahyani Lupinda[1], Rito Goejantoro[2], Memi Nor Hayati[3a)], Aji Syarif Hidayatullah[4]
[1,2,3] Program Studi Statistika, Universitas Mulawarman, Indonesia
[4] Program Studi Administrasi Publik, Universitas Widya Gama Mahakam, Indonesia

[a)] Corresponding author: meminorhayati@ fmipa.unmul.ac.id

Abstract

Data mining is the process of extracting useful information and patterns from very large amounts of data. Based on the task or work performed, data mining is divided into cluster analysis, association analysis, anomaly detection, and predictive modeling. Predictive modeling consists of two types, namely regression and classification. Classification is a method for determining the membership of an object in a class based on available data. There are several methods for classification, one of which is naïve Bayes with the advantages of being easy to build and having good performance. This study was conducted using data on the welfare status of beneficiary households in Samarinda Ilir District in 2023 using the naive Bayes classification method which aims to determine the accuracy of the naive Bayes classification on the welfare status data of beneficiary households in Samarinda Ilir District in 2023 and to assist the government in determining the welfare status of households eligible to receive benefits so that they are right on target. Based on the research results, it can be seen that the accuracy level of the naïve Bayes classification on this data is 0.8316 or 83.16%. The results of accuracy measurements show that the naïve Bayes classification of this data has a fairly high level of accuracy.

Keywords: *accuracy, classification, naïve bayes*

1. Pendahuluan

Data mining merupakan proses penggalian informasi dan pola yang bermanfaat dari data yang sangat besar. Berdasarkan tugas atau pekerjaan yang dilakukan, *data mining* terbagi menjadi analisis klaster, analisis asosiasi, deteksi anomali, dan pemodelan prediktif. Pemodelan prediktif terdiri dari regresi dan klasifikasi (Suyanto, 2019). Klasifikasi adalah suatu proses untuk memperoleh sejumlah model atau fungsi yang dapat menggambarkan, mengenali, dan membedakan kelas data yang bertujuan untuk menggunakan model yang diperoleh agar dapat memprediksi kelas objek yang kelasnya belum diketahui. Beberapa data dengan struktur data yang mirip akan memiliki klasifikasi yang mirip pula. Proses klasifikasi diawali dengan membangun data yang dikenal dengan data *training* yang target kelas atau nilai target telah diketahui kemudian untuk memperoleh kelas dari data *testing* dapat menggunakan algoritma dalam *data mining* (Han dkk., 2012). Data *training* biasanya digunakan untuk membangun model, pola, ataupun fungsi dari sekumpulan data dan apabila model telah diperoleh maka dapat divalidasi dan ditentukan keakuratannya menggunakan data *testing* untuk menentukan data-data yang diuji tersebut masuk ke dalam kelas yang mana (Arhami & Nasir, 2020). Terdapat beberapa algoritma klasifikasi yang dapat digunakan, salah satunya adalah *naïve* Bayes. Algoritma *naïve* Bayes adalah salah satu algoritma klasifikasi berbasis probabilitas/peluang yang digunakan untuk memprediksi probabilitas/peluang keanggotaan suatu kelas di masa depan dengan memanfaatkan data di masa lalu atau biasa dikenal dengan teorema Bayes yang mengasumsikan semua variabel independen (Pramana dkk., 2018).

Algoritma *naïve* Bayes hanya membutuhkan jumlah data *training* yang kecil untuk menentukan estimasi parameter yang dibutuhkan dalam proses klasifikasi sehingga klasifikasi *naïve* Bayes merupakan klasifikasi yang mudah dibangun dan memiliki kinerja yang baik. Model klasifikasi yang baik dapat diketahui dengan melihat tingkat akurasi dalam memprediksi berdasarkan kategori respon (Herman dkk., 2018).

Berdasarkan hasil survei kemiskinan Kota Samarinda tahun 2023, Kecamatan Samarinda Ilir merupakan kecamatan dengan selisih terbesar antara rumah tangga masuk kriteria miskin dan tidak miskin, di mana sebanyak 2.158 rumah tangga termasuk kriteria miskin dan sebanyak 1.301 rumah tangga tidak termasuk kriteria miskin (Hayati dkk., 2023). Pemerintah berperan besar dalam meningkatkan kesejahteraan masyarakat agar terhindar dari kemiskinan melalui berbagai program pembangunan, namun dalam pelaksanaannya seringkali ditemukan kesalahan dalam menentukan status kesejahteraan masyarakat sehingga program pemerintah tidak tepat sasaran (Asj'ari, 2015). Berdasarkan uraian tersebut, untuk membantu pemerintah dalam menentukan status kesejahteraan masyarakat agar dapat mengetahui rumah tangga layak menerima manfaat diperlukan suatu sistem pengolah data dengan salah satu metode klasifikasi yaitu naïve Bayes.

Berdasarkan penelitian yang telah dilakukan sebelumnya tentang perbandingan metode *Naïve* Bayes

p-ISSN 2085-7829
e-ISSN 2798-3455

dan *K-Nearest Neighbor* (K-NN) pada data status pembayaran Pajak Pertambahan Nilai (PPN) di Kantor Pelayanan Pajak (KPP) Pratama Samarinda Ulu, diperoleh hasil bahwa *Naïve* Bayes menunjukkan kesalahan klasifikasi dalam memprediksi sebesar 17,01% sedangkan metode K-NN menunjukkan kesalahan klasifikasi dalam memprediksi sebesar 19,51%. Hal ini menunjukkan bahwa *Naïve* Bayes bekerja lebih baik dibandingkan K-NN dalam mengklasifikasikan status pembayaran PPN di KPP Pratama Samarinda Ulu berdasarkan nilai APER yang rendah (Rahmaulidyah dkk., 2021). Berdasarkan uraian yang telah dipaparkan, oleh karena itu dilakukan penelitian ini yang bertujuan untuk mengetahui ketepatan klasifikasi *Naïve* Bayes pada data status kesejahteraan rumah tangga penerima manfaat di Kecamatan Samarinda Ilir Tahun 2023.

2. Metodologi Penelitian

Penelitian yang dilakukan menggunakan data sekunder dari hasil Survei Kemiskinan di Kota Samarinda Tahun 2023.

Tabel 1. Variabel Penelitian

Notasi	Variabel Penelitian
C	Status rumah tangga
X_1	Status bangunan
X_2	Status lahan
X_3	Jenis lantai
X_4	Jenis dinding
X_5	Jenis atap
X_6	Sumber Penerangan
X_7	Persepsi ekonomi
X_8	Kepemilikan fasilitas tempat buang air besar
X_9	Tempat pembuangan akhir tinja
X_{10}	Kepemilikan aset bergerak
X_{11}	Kepemilikan aset tidak bergerak
X_{12}	Kepemilikan usaha
X_{13}	Pengeluaran rumah tangga per bulan
X_{14}	Luas Bangunan

Teknik analisis data pada penelitian ini adalah klasifikasi status kesejahteraan rumah tangga menggunakan *naïve* Bayes. Penelitian ini menggunakan bantuan *software* komputer yaitu Microsoft Excel untuk analisis statistika deskriptif dan *Rstudio* untuk proses klasifikasi. Adapun tahapan dalam teknik analisis data yang dilakukan peneliti adalah sebagai berikut:

1. Melakukan statistika deskriptif dengan membuat penyajian data dalam bentuk diagram lingkaran pada variabel status rumah tangga.
2. Melakukan randomisasi data agar semua data memiliki kesempatan yang sama untuk menjadi data *training* dan data *testing*.
3. Membagi data menjadi data *training* dan data *testing* dengan proporsi 80:20 dengan cara memilih 80% dari data urutan teratas hasil randomisasi menjadi data *training* dan sebanyak 20% data urutan terbawah hasil randomisasi menjadi data *testing*.
4. Melakukan klasifikasi *naive* Bayes pada data. Langkah-langkah yang dilakukan dalam mengolah data pengklasifikasian dengan *naïve* Bayes menggunakan bantuan *software* R adalah sebagai berikut:
   a. Menghitung nilai probabilitas awal (*prior*) kelas menggunakan data *training* berdasarkan persamaan (1) (Suntoro, 2012).

$$P(C_g) = \frac{n_g}{N} \tag{1}$$

   b. Menghitung probabilitas setiap variabel bebas pada setiap kelas berdasarkan persamaan (2) untuk data bertipe kualitatif,

$$P(X_a \mid C, X_b) = P(X_a \mid C) \tag{2}$$

   dan persamaan (3) untuk data bertipe kuantitatif (Prasetyo, 2012).

$$P(X = x_a \mid C = c_g) = \frac{1}{\sqrt{2\pi}\sigma_g} e^{-\left(\frac{(x_a-\mu_g)^2}{2\sigma_g^2}\right)} \tag{3}$$

   c. Menghitung perkalian antara *prior* dan probabilitas setiap variabel bebas pada setiap kelas berdasarkan persamaan (4).

p-ISSN 2085-7829
e-ISSN 2798-3455

$$P(C \mid X_1, X_2, ..., X_m) = P(C) \prod_{a=1}^{m} P(X_a \mid C) \quad (4)$$

d. Menghitung probabilitas akhir (*posterior*) pada masing-masing kelas berdasarkan persamaan (5).

$$P(C \mid X_1, X_2, ..., X_m) = \frac{P(C) P(X_1, X_2, ..., X_m \mid C)}{P(X_1, X_2, ..., X_m)} \quad (5)$$

e. Menentukan hasil klasifikasi objek dengan melihat nilai maksimum pada kedua kelas.

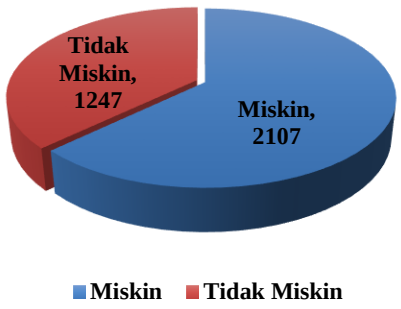
5. Mengevaluasi hasil klasifikasi menggunakan persamaan (6) (Handayanto & Herlawati, 2020).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (6)$$

3. Hasil dan Pembahasan

3.1 Analisis Statistika Deskriptif

Gambaran umum mengenai status rumah tangga pada rumah tangga penerima manfaat di Kecamatan Samarinda Ilir Tahun 2023 dengan status rumah tangga miskin dan tidak miskin dapat dilihat pada Gambar 1.



Gambar 1. Statistika deskriptif status rumah tangga

Berdasarkan Gambar 1 dapat dilihat sebanyak 2.107 atau 63% rumah tangga termasuk kriteria rumah tangga miskin dan sebanyak 1.247 atau 37% rumah tangga termasuk tidak miskin. Dari Gambar 1 dapat diketahui bahwa rumah tangga miskin lebih banyak dibandingkan rumah tangga tidak miskin.

3.2 Membagi Data *Training* dan Data *Testing*

Langkah awal sebelum melakukan klasifikasi adalah membagi data menjadi data *training* dan data *testing* berdasarkan hasil randomisasi. Adapun data yang digunakan pada proses klasifikasi yaitu data rumah tangga penerima manfaat di Kecamatan Samarinda Ilir dengan data *training* 80% sebanyak 2.683 sampel, sedangkan untuk data *testing* 20% sebanyak 671 sampel.

3.3 Klasifikasi *Naïve* Bayes

a. Klasifikasi *Naive* Bayes

1. Menghitung probabilitas *prior* status rumah tangga

Adapun nilai probabilitas *prior* setiap kelompok menggunakan persamaan (1) sebagai berikut:

a). Kelompok pertama (Rumah tangga berstatus miskin)

$$P(C_1) = \frac{n_1}{N} = \frac{1.690}{2.683} = 0,62989$$

Jadi, probabilitas *prior* rumah tangga berstatus miskin adalah sebesar 0,62989.

b). Kelompok kedua (Rumah tangga berstatus tidak miskin)

$$P(C_2) = \frac{n_2}{N} = \frac{993}{2.683} = 0,37011$$

Jadi, probabilitas *prior* rumah tangga berstatus tidak miskin adalah sebesar 0,37011.

2. Menghitung probabilitas setiap variabel bebas pada setiap kelas

Tahapan selanjutnya yaitu menghitung probabilitas setiap variabel bebas pada kedua kelompok yang dilakukan pada data *testing* berdasarkan data *training* menggunakan persamaan (2) untuk variabel data kualitatif dan persamaan (3) untuk variabel data kuantitatif. Pada data *testing* pertama diketahui

p-ISSN 2085-7829
e-ISSN 2798-3455

rumah tangga dengan status bangunan (X_1) kontrak atau sewa, status lahan (X_2) milik orang lain, jenis lantai (X_3) ubin/tegel/teraso, jenis dinding (X_4) tembok, jenis atap (X_5) seng, sumber penerangan (X_6) listrik PLN ≤ 900 Watt, persepsi ekonomi (X_7) tidak khawatir, memiliki fasilitas tempat buang air besar (X_8) sendiri, dengan tempat pembuangan akhir tinja (X_9) di tangki, memiliki aset bergerak (X_{10}), tidak memiliki aset tidak bergerak (X_{11}), tidak memiliki usaha (X_{12}), pengeluaran rumah tangga (X_{13}) sebesar Rp2.000.000 < Pengeluaran ≤ Rp2.500.000 per bulan, dan luas bangunan (X_{14}) sebesar 79 m^2. Adapun perhitungan nilai probabilitas setiap variabel bebas pada kedua kelompok sebagai berikut:

Tabel 2. Perhitungan Probabilitas Setiap Variabel Bebas Pada Kedua Kelompok

Variabel Bebas	Kategori	Status Rumah Tangga	
		Miskin	Tidak Miskin
X_1	Kontrak atau sewa	0,22485	0,20947
X_2	Milik orang lain	0,48462	0,45720
X_3	Ubin/Tegel/Teraso	0,00237	0,02417
X_4	Tembok	0,05385	0,46123
X_5	Seng	0,98935	0,96475
X_6	Listrik PLN ≤ 900 Watt	0,84970	0,83585
X_7	Tidak khawatir	0,90888	0,97080
X_8	Sendiri	0,90059	0,97583
X_9	Tangki	0,81361	0,92044
X_{10}	Memiliki aset bergerak	0,98462	0,99899
X_{11}	Tidak memiliki aset tidak bergerak	0,97929	0,89023
X_{12}	Tidak memiliki usaha	0,88698	0,61128
X_{13}	Rp2.000.000 < Pengeluaran ≤ Rp2.500.000	0,16272	0,22558
X_{14}	79 m^2	0,00791	0,01016

3. Menghitung perkalian *prior* dan probabilitas setiap variabel bebas pada setiap kelas berdasarkan persamaan (4)
   c). Kelompok pertama (Rumah tangga berstatus miskin)

$$P(C_1 \mid X) = P(C_1)\prod_{a=1}^{14} P(X_a \mid C_1) = 5,39\times10^{-9}$$

   d). Kelompok kedua (Rumah tangga berstatus tidak miskin)

$$P(C_2 \mid X) = P(C_2)\prod_{a=1}^{14} P(X_a \mid C_2) = 3,461\times10^{-7}$$

4. Menghitung probabilitas *posterior* pada masing-masing kelas berdasarkan persamaan (5)
   e). Kelompok pertama (Rumah tangga berstatus miskin)

$$P(C_1 \mid X) = \frac{5,39\times10^{-9}}{5,39\times10^{-9} + 3,461\times10^{-7}} = 0,01533$$

   f). Kelompok kedua (Rumah tangga berstatus tidak miskin)

$$P(C_2 \mid X) = \frac{3,461\times10^{-7}}{5,39\times10^{-9} + 3,461\times10^{-7}} = 0,98467$$

Berdasarkan perhitungan probabilitas *posterior* pada kedua kelompok tersebut dapat diketahui bahwa kelas yang memiliki nilai probabilitas terbesar adalah kelas tidak miskin yaitu sebesar 0,98467, sehingga dapat disimpulkan bahwa kasus seperti pada data *testing* pertama diprediksi masuk dalam kelas tidak miskin artinya rumah tangga tersebut termasuk rumah tangga dengan status tidak miskin.

3.4 Evaluasi Hasil Klasifikasi

Setelah dilakukan klasifikasi rumah tangga penerima manfaat menggunakan metode *naïve* Bayes, selanjutnya dilakukan evaluasi menggunakan data *testing*. Evaluasi yang digunakan dalam penelitian ini adalah matriks konfusi. Evaluasi matriks konfusi akan membentuk matriks yang terdiri dari *True Positives, False Negatives, False Positives, dan True Negatives*. Untuk matriks konfusi dari data dapat dilihat pada Tabel 3.

Tabel 3. Matriks Konfusi Klasifikasi *Naive* Bayes

Kelas Prediksi	Kelas Aktual	
	Miskin	Tidak Miskin
Miskin	390	86
Tidak Miskin	27	168

p-ISSN 2085-7829
e-ISSN 2798-3455

Dari matriks konfusi tersebut dapat dihitung nilai *Accuracy* menggunakan persamaan (6). Perhitungan *Accuracy* dapat dilihat pada Tabel 4.

Tabel 4. Nilai *Accuracy*

Metode	**Nilai *Accuracy***
Klasifikasi *naïve* Bayes pada data status kesejahteraan rumah tangga penerima manfaat di Kecamatan Samarinda Ilir Tahun 2023	0,8316

4. Kesimpulan

Berdasarkan hasil analisis dan pembahasan yang telah dilakukan, maka dapat disimpulkan bahwa pada klasifikasi *naïve* Bayes diperoleh nilai *accuracy* yang merupakan persentase dari jumlah data yang diprediksi secara benar sebesar 0,8316 atau 83,16%.

5. Daftar Pustaka

E. Prasetyo. (2012). *Data Mining: Konsep dan Aplikasi Menggunakan Matlab.* Yogyakarta: Andi Offset.

F. Asj'ari. (2015). Pengaruh Pertumbuhan Ekonomi Terhadap Kesejahteraan Keluarga Bukan Pegawai Negeri Sipil di Surabaya. *Majalah Ekonomi.* Vol 20. No 1. 87-96.

F. N. Rahmaulidyah, M. N. Hayati dan R. Goejantoro. (2021). Perbandingan Metode Klasifikasi Naive Bayes dan K-Nearest Neighbor Pada Data Status Pembayaran Pajak Pertambahan Nilai di Kantor Pelayanan Pajak Pratama Samarinda Ulu. *Jurnal EKSPONENSIAL.* Vol 12. No 2. 161-165.

J. Han, M. Kamber dan J. Pei. (2012). *Data Mining Concept and* Techniques. California: Morgan Kaufmann.

J. Suntoro. (2012). *Data Mining: Algoritma dan Implementasi Menggunakan Bahasa Pemrograman PHP*. Semarang: Universitas Semarang.

M. Arhami dan M. Nasir. (2020). *Data Mining: Algoritma dan Implementasi*. Yogyakarta: ANDI.

M. N. Hayati, A. Sofyan, A. A. Lepu, A. S. Hidayatullah, A. S. Hartoyo, B. D. Ariyanto dan Udin. (2023). *Hasil Survei Kemiskinan Kota Samarinda Tahun 2023*, Samarinda: DISKOMINFO.

R. T. Handayanto dan Herlawati. (2020). *Data Mining dan Machine Learning Menggunakan Matlab dan Python*, Bandung: INFORMATIKA.

S. Pramana, B. Yuniarto, S. Mariyah, I. Santoso dan R. Nooraeni. (2018). *Data Mining dengan R*: *Konsep serta Implementasi*. Bogor: IN MEDIA.

Suyanto. (2019). *Data Mining Untuk Klasifikasi dan Klasterisasi Data*. Bandung: INFORMATIKA