

PERBANDINGAN KINERJA ALGORITMA NAÏVE BAYES DAN C4.5 PADA SISTEM WEB KLASIFIKASI KELAYAKAN PKH

Jupriyanto², Jamaludin Apandi¹, Anderias Eko Wijaya^{*1}, Rian Hermawan², Timbo Faritcan Parlaungan Siallagan³, Kodar Udoyono³, Hermansyah Nur Ahmad¹

Teknik Informatika, Universitas Mandiri¹
Sistem Informasi, Universitas Mandiri²
Teknik Komputer dan Jaringan, Universitas Mandiri³

E-mail: jupriyanto@universitasmandiri.ac.id², jamaludinapandi29@gmail.com¹,
ekowjy09@gmail.com^{*1}, rian@universitasmandiri.ac.id², timbo@universitasmandiri.ac.id³
kodar@universitasmandiri.ac.id³, hermansyahna@gmail.com¹

Received: 17-03-2025 | Accepted: 29-03-2025 | Published: 01-04-2025

Abstrak

Penelitian ini membahas pengembangan sistem klasifikasi kelayakan peserta Program Keluarga Harapan (PKH) berbasis web dengan membandingkan dua algoritma data mining, yaitu C4.5 dan Naïve Bayes. Data yang digunakan mencakup sejumlah atribut penentu kelayakan bantuan sosial. Algoritma C4.5 digunakan untuk membentuk pohon keputusan yang mudah dipahami, sedangkan Naïve Bayes dimanfaatkan untuk klasifikasi berbasis probabilitas. Hasil pengujian menunjukkan bahwa algoritma Naïve Bayes memiliki tingkat akurasi tertinggi sebesar 98%, unggul dalam kecepatan pemrosesan terhadap dataset berskala besar. Sementara itu, algoritma C4.5 mencapai akurasi sebesar 93,33% dan memberikan kelebihan dalam hal interpretabilitas melalui visualisasi pohon keputusan. Kedua algoritma terbukti efektif dalam mengklasifikasikan kelayakan penerima PKH, dan dapat diimplementasikan dalam sistem informasi bantuan sosial untuk meningkatkan akurasi serta efisiensi proses seleksi penerima. Penelitian ini menyimpulkan bahwa pemilihan algoritma dapat disesuaikan dengan kebutuhan sistem, apakah lebih menekankan pada kecepatan atau kemudahan interpretasi.

Kata Kunci: Algoritma C4.5, Algoritma Naïve Bayes, Program Keluarga Harapan (PKH).

Abstract

This study discusses the development of a web-based classification system for determining the eligibility of recipients of the Family Hope Program (PKH), by comparing two data mining algorithms: C4.5 and Naïve Bayes. The dataset used includes various attributes relevant to eligibility assessment for social assistance. The C4.5 algorithm is employed to generate an interpretable decision tree, while the Naïve Bayes algorithm is used for probabilistic classification. The results show that Naïve Bayes achieved the highest accuracy at 98%, excelling in processing large datasets more efficiently. Meanwhile, C4.5 achieved an accuracy of 93.33% and offered better interpretability through its decision tree visualization. Both algorithms proved effective in classifying PKH eligibility and can be implemented in social assistance information systems to improve the accuracy and efficiency of the beneficiary selection process. This research concludes that the choice of algorithm should be based on system priorities—whether the focus is on processing speed or result interpretability.

Keywords: C4.5 algorithm, Naïve Bayes algorithm, Program Keluarga Harapan (PKH).

1. Pendahuluan

Program Keluarga Harapan (PKH) adalah salah satu bentuk bantuan sosial dari pemerintah Indonesia yang bertujuan untuk meningkatkan kesejahteraan keluarga miskin. Salah satu tantangan besar dalam implementasi PKH adalah penentuan penerima bantuan yang tepat dan akurat. Proses ini melibatkan banyak kriteria dan seringkali terkendala oleh ketidaklengkapan atau ketidakakuratan



data yang ada. Oleh karena itu, dibutuhkan sistem yang efisien dan akurat dalam menentukan kelayakan penerima bantuan.

Dalam hal ini, data mining telah menjadi salah satu solusi yang efektif, terutama dengan menggunakan algoritma pembelajaran mesin yang dapat mengklasifikasikan data berdasarkan pola yang ditemukan. Algoritma C4.5 adalah salah satu metode yang sering digunakan untuk klasifikasi karena kemampuannya dalam membangun pohon keputusan yang mudah dipahami dan efisien dalam menangani data besar. Algoritma ini telah terbukti efektif dalam berbagai aplikasi, seperti menentukan kelayakan penerima Bantuan Langsung Tunai (BLT) [1], [2], [3], dan dalam mengukur tingkat kepuasan pasien [2]. Penelitian sebelumnya juga menunjukkan bahwa C4.5 dapat diandalkan dalam menentukan penerima bantuan sosial di berbagai daerah [1].

Sementara itu, algoritma Naïve Bayes juga populer digunakan dalam klasifikasi, terutama ketika data memiliki ketergantungan rendah antar atribut. Metode ini bekerja dengan probabilitas berdasarkan Teorema Bayes, menjadikannya efisien untuk klasifikasi pada data besar yang terdistribusi secara tidak merata. Penelitian yang dilakukan oleh Rahmadani et al. [5] menunjukkan bahwa Naïve Bayes dapat memberikan hasil yang cukup baik dalam menentukan kelayakan penerima bantuan sosial.

Perbandingan antara C4.5 dan Naïve Bayes dalam berbagai aplikasi data mining, terutama dalam konteks klasifikasi, telah dilakukan dalam beberapa studi sebelumnya. Misalnya, Bhat dan Bhat [6] serta Sinaga dan Winanjaya [1] menguji kedua algoritma ini dalam berbagai kasus klasifikasi dan menemukan perbedaan yang signifikan dalam hal akurasi dan efisiensi. Rasika [7] juga mengaplikasikan C4.5 untuk menentukan faktor kelayakan penerima bantuan langsung tunai (BLT-DD), yang memberikan gambaran lebih dalam mengenai kelebihan algoritma ini dalam konteks sosial [8].

Selain itu, penelitian oleh Quinlan [9] dan Hastie et al. [10] memberikan penjelasan mendalam mengenai kelebihan dan kekurangan masing-masing algoritma dalam berbagai aplikasi data mining. Penelitian oleh Wibowo [11] dan Chandra [12] juga menunjukkan penerapan metode data mining dalam seleksi dan pengklasifikasian data, meskipun tidak langsung berfokus pada bantuan sosial.

Dengan mempertimbangkan keunggulan dan karakteristik masing-masing algoritma, penelitian ini bertujuan untuk membandingkan efektivitas algoritma Naïve Bayes dan C4.5 dalam menentukan kelayakan penerima bantuan Program Keluarga Harapan. Melalui sistem berbasis web, diharapkan proses klasifikasi dapat dilakukan dengan lebih cepat, akurat, dan efisien, sehingga dapat mendukung distribusi bantuan yang lebih tepat sasaran dan berkelanjutan.

2. Analisa dan Pembahasan

2.1. Metodologi Penelitian

1) Jenis Penelitian

Penelitian ini menggunakan pendekatan kuantitatif dengan metode deskriptif dan eksperimental. Penelitian ini bertujuan untuk menganalisis kelayakan penerima Program Keluarga Harapan (PKH) menggunakan dua algoritma data mining, yaitu Naïve Bayes dan C4.5.

2) Sumber Data

Data yang digunakan dalam penelitian ini diperoleh dari survei lapangan di desa yang mencakup informasi sosial dan ekonomi keluarga penerima PKH, termasuk penghasilan, jumlah anggota keluarga, status pekerjaan, pendidikan, dan kondisi kesehatan. Data yang telah terkumpul kemudian divalidasi dan dikonsolidasikan menjadi dataset yang dapat digunakan untuk analisis.

3) Teknik Pengolahan Data

Data yang telah diperoleh melalui survei akan diproses dan dianalisis menggunakan algoritma Naïve Bayes dan C4.5 untuk mengklasifikasikan keluarga yang layak menerima bantuan berdasarkan kriteria yang telah ditentukan.

- Naïve Bayes digunakan untuk mengklasifikasikan keluarga berdasarkan probabilitas kondisi sosial dan ekonomi.
- C4.5 digunakan untuk membangun pohon keputusan yang dapat menggambarkan hubungan antara variabel-variabel yang mempengaruhi kelayakan penerima bantuan.

4) Langkah-langkah Analisis

- Persiapan Data: Data yang diperoleh melalui survei lapangan akan disaring dan disusun untuk memenuhi format yang diperlukan oleh algoritma.
- Pengolahan Data: Setelah data dipersiapkan, algoritma Naïve Bayes dan C4.5 akan diterapkan untuk menganalisis kelayakan penerima PKH.
- Evaluasi Model: Hasil dari kedua algoritma akan dievaluasi berdasarkan tingkat akurasi dan efisiensi dalam mengklasifikasikan penerima bantuan.

5) Evaluasi dan Pengujian

Model yang dibangun akan diuji untuk menilai akurasi prediksi penerima bantuan dengan membandingkan hasil klasifikasi dengan data sebenarnya (ground truth). Evaluasi dilakukan dengan menggunakan akurasi, presisi, recall, dan F1-score.

2.2. Tahapan Pemodelan Menggunakan Algoritma Naïve Bayes dan C4.5:

a. Teknik Pengambilan Data

Pengambilan data untuk Program Keluarga Harapan (PKH) di desa melibatkan beberapa langkah untuk memastikan data yang akurat, yaitu:

- Pendataan Awal: Melakukan survei lapangan, wawancara dengan kepala keluarga atau anggota keluarga, dan observasi kondisi rumah serta aset keluarga untuk memperoleh informasi sosial dan ekonomi.
- Validasi Data: Data yang terkumpul divalidasi untuk memastikan akurasi, melalui pengecekan ulang atau menggunakan data dari instansi terkait.
- Konsolidasi Data: Data yang telah diverifikasi kemudian disusun menjadi dataset penerima PKH yang valid.

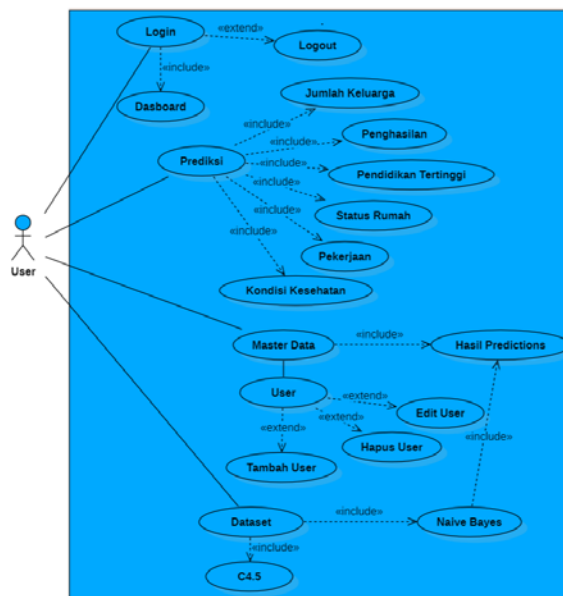
b. Penentuan Kriteria Penerima

Berdasarkan data yang dikumpulkan, kriteria kelayakan penerima PKH ditentukan, seperti jumlah anggota keluarga, penghasilan, pendidikan, status rumah, pekerjaan, dan kondisi kesehatan. Algoritma Naïve Bayes atau C4.5 kemudian digunakan untuk mengklasifikasikan keluarga yang layak menerima bantuan, berdasarkan kriteria tersebut.

2.3. Pemodelan Sistem

a. Use Case Diagram

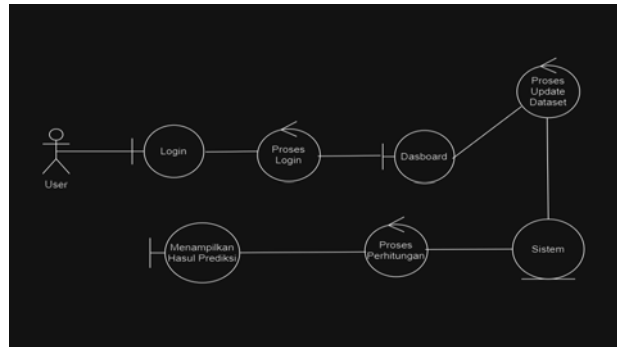
Berikut diagram *use case* gambar 3.1 pada aplikasi Penerapan *Algoritma Naïve Bayes* dan C4.5 pada Sistem Perbandingan klasifikasi data PKH.



Gambar 3.1 Use Case Diagram

b. *Communication Diagram*

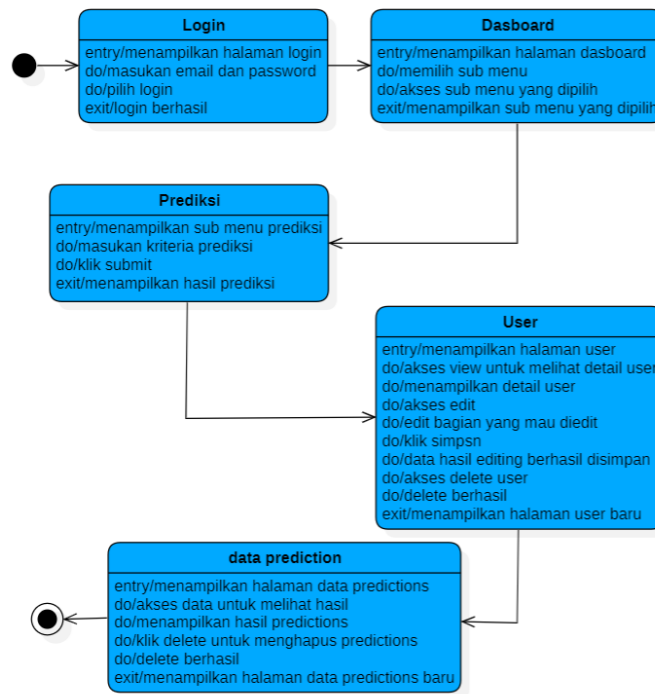
Berikut diagram *communication* gambar 3.2 pada aplikasi Penerapan Algoritma Naïve Bayes dan C4.5 pada Sistem Perbandingan klasifikasi data PKH (studi kasus: Desa Tenjolaya).



Gambar 3.2 *Communication Diagram*

c. *Statechart Diagram*

Berikut diagram *statechart* gambar 3.3 pada aplikasi Penerapan Algoritma Naïve Bayes dan C4.5 pada Sistem Perbandingan klasifikasi data PKH (studi kasus: Desa Tenjolaya).



Gambar 3.3 *Statechart Diagram*

2.4. Perhitungan Algoritma *Naive Bayes*

Berikut ini adalah perhitungan manual penyelesaian menggunakan algoritma Naïve Bayes dalam Klasifikasi kelayakan.

1. Menghitung Probabilitas Prior

setiap kategori, yang menjadi kategori ada 6 kategori yaitu kelas jumlah keluarga, penghasilan per bulan, pendidikan tertinggi, setatus rumah, pekerjaan, kondisi Kesehatan

Untuk menghitung probabilitas Priorr dengan rumus

$$Prob = \frac{frekuensi}{jumlah.data}$$

Probabilitas prior untuk masing-masing anggota yaitu:

$$P(\text{setatus penerima} = \text{ya}) = 0,50$$

$$P(\text{setatus penerima} = \text{tidak}) = 0,50$$

2. Menghitung Probilitas Kondisi (Likelihood)

melibatkan enam variabel yaitu jumlah keluarga, penghasilan per bulan, pendidikan tertinggi, status rumah, pekerjaan, dan kondisi kesehatan, di mana perhitungan probabilitas kondisi sama seperti probabilitas prior tetapi melibatkan anggota atau kelas target.

Rumus Probabilitas Kondisi (Likelihood)

Probabilitas kondisi untuk setiap kategori dihitung dengan membagi frekuensi kategori tersebut dengan total frekuensi dalam kelas yang sama.

Probabilitas "Ya" untuk setiap kategori:

Probabilitas "Ya" untuk "kecil":

$$Probilitas (kecil | Ya) = \frac{Frekuensi\ kecil\ (Ya)}{Total\ Frekuensi\ (Ya)} = \frac{5}{30} = 0,16$$

Probabilitas "Ya" untuk "sedang":

$$Probilitas (sedang | Ya) = \frac{Frekuensi\ sedang\ (Ya)}{Total\ Frekuensi\ sedang\ (Ya)} = \frac{24}{30} = 0,8$$

Probabilitas "Ya" untuk "besar":

$$Probilitas (besar | Ya) = \frac{Frekuensi\ besar\ (Ya)}{Total\ Frekuensi\ besar\ (Ya)} = \frac{1}{30} = 0,03$$

Probabilitas "Tidak" untuk setiap kategori:

Probabilitas "Tidak" untuk "kecil":

$$Probilitas (kecil | Tidak) = \frac{Frekuensi\ kecil\ (tidak)}{Total\ Frekuensi\ (tidak)} = \frac{10}{30} = 0,33$$

Probabilitas "Tidak" untuk "Sedang":

$$Probilitas (sedang | Tidak) = \frac{Frekuensi\ sedang\ (tidak)}{Total\ Frekuensi\ (tidak)} = \frac{10}{30} = 0,33$$

Probabilitas "Tidak" untuk "besar":

$$Probilitas (besar | Tidak) = \frac{Frekuensi\ besar\ (tidak)}{Total\ Frekuensi\ (tidak)} = \frac{10}{30} = 0,33$$

1. Probilitas kondisi jumlah keluarga

Tabel 3.1 Probilitas Jumlah Keluarga

Anggota jumlah keluarga	Frekuensi		Probabilitas	
	Ya	Tidak	Ya	Tidak
kecil	5	10	0,16	0,33
Sedang	24	10	0,8	0,33
besar	1	10	0,03	0,33
Total	30	30	1	1

2. Probilitas kondisi pengasilan

Tabel 3.2 Probilitas Penghasilan

penghasilan per bulan	Frekuensi		Probabilitas	
	Ya	Tidak	Ya	Tidak
kurang	23	0	0,76	0,0
Sedang	9	9	0,3	0,3
tinggi	0	23	0,0	0,76
Total	30	30	1	1

penghasilan per bulan kecil

$$P(\text{penghasilanPerBulan}=\text{kurang} = \text{ya}) = 0,69$$

$$P(\text{penghasilanPerBulan}=\text{kurang} = \text{tidak}) = 0,03$$

penghasilan per bulan sedang

$$P(\text{penghasilanPerBulan}=\text{sedang} = \text{ya}) = 0,27$$

$$P(\text{penghasilanPerBulan}=\text{sedang} = \text{tidak}) = 0,27$$

penghasilan per bulan tinggi

$P(\text{penghasilanPerBulan}=\text{tinggi} = \text{ya}) = 0,03$

$P(\text{penghasilanPerBulan}=\text{tinggi} = \text{tidak}) = 0,69$

3. Probabilitas kondisi Pendidikan

Tabel 3.3 Probabilitas Pendidikan

pendidikan tertinggi	Frekuensi		Probabilitas	
	Ya	Tidak	Ya	Tidak
sd	15	6	0,42	0,17
smp	13	13	0,37	0,37
sma	1	9	0,02	0,25
tidak sekolah	5	4	0,14	0,11
sarjana	1	3	0,02	0,08
Total	35	35	1	1

4. Probabilitas setatus rumah

Tabel 3.4 Probabilitas Pendidikan

setatus rumah	Frekuensi		Probabilitas	
	Ya	Tidak	Ya	Tidak
kurang	8	5	0,24	0,15
Sedang	24	13	0,72	0,39
baik	1	15	0,03	0,45
Total	33	33	1	1

5. Probabilitas setatus pekerjaan

Tabel 3.5 Probabilitas Pekerjaan

pekerjaan	Frekuensi		Probabilitas	
	Ya	Tidak	Ya	Tidak
lain-lain	23	20	0,69	0,60
petani	9	6	0,27	0,18
wirusaha	1	7	0,03	0,21
Total	33	33	1	1

6. Probabilitas setatus Kesehatan

Tabel 3.6 Probabilitas kesehatan

setatus kesehatan	Frekuensi		Probabilitas	
	Ya	Tidak	Ya	Tidak
kurang	8	4	0,24	0,12
cukup	24	15	0,72	0,45
baik	1	14	0,03	0,42
Total	33	33	1	1

3. Perhitungan Klasifikasi

Setelah didapatkan nilai untuk probabilitas prior dan kondisi maka untuk klasifikasi pada Perhitungan Manual Naives Bayes sebagai berikut.

Untuk menghitung probabilitas suatu kategori (Ya/Tidak) berdasarkan kondisi-kondisi yang diberikan, kita menggunakan rumus sebagai berikut:

$$P(\text{Ya} | \text{Kondisi}) = \frac{P(\text{Kondisi} | \text{Ya}) \times P(\text{Ya})}{P(\text{Kondisi})}$$

$$P(\text{Tidak} | \text{Kondisi}) = \frac{P(\text{Kondisi} | \text{Tidak}) \times P(\text{Tidak})}{P(\text{Kondisi})}$$

Namun, karena $P(\text{Kondisi})$ sama untuk kedua kasus, kita dapat membandingkan $P(\text{Kondisi} | \text{Ya}) \times P(\text{Ya})$ dengan $P(\text{Kondisi} | \text{Tidak}) \times P(\text{Tidak})$.

3.1 Tingkat Akurasi

Untuk menghitung tingkat Accuracy, Precision, Recall, dan F1-Score, kita perlu terlebih dahulu menghitung jumlah True Positives (TP), True Negatives (TN), False Positives (FP), dan False Negatives (FN) berdasarkan data.

Menentukan TP, TN, FP, dan FN

- True Positives (TP): Kategori yang benar dan diprediksi benar.
- True Negatives (TN): Kategori yang benar dan diprediksi benar.
- False Positives (FP): Kategori yang salah namun diprediksi benar.
- False Negatives (FN): Kategori yang benar namun diprediksi salah.

Tabel 3.7 Tingkat Akurasi

no	ya	tidak	fakta	klasifikasi	prediksi	Kategori yang	Kategori yg	True/False
						Benar	Diprediksi	
1	0.00119113	0.00046286	Ya	Ya	sesuai	Ya	Ya	TP
2	0.00000037	0.00015324	Tidak	Tidak	sesuai	Tidak	Tidak	TN
3	0.01217597	0.00006743	Ya	Ya	sesuai	Ya	Ya	TP
4	0.01382137	0.00005212	Ya	Ya	sesuai	Ya	Ya	TP
5	0.00000000	0.00102698	Tidak	Tidak	sesuai	Tidak	Tidak	TN
....
....
59	0.00000000	0.00110003	Tidak	Tidak	sesuai	Tidak	Tidak	TN
60	0.00000003	0.00293423	Tidak	Tidak	sesuai	Tidak	Tidak	TN

Jumlah TP, TN, FP, FN:

- TP (True Positives): 26
- TN (True Negatives): 26
- FP (False Positives): 0
- FN (False Negatives): 0

1. Menghitung Numerik

1. Accuracy:

$$\text{Accuracy} = \frac{TP+TN}{\text{Total}}$$

$$\text{Accuracy} = \frac{26+26}{60} = \frac{52}{60} = 0.867$$

2. Precision:

$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{Precision} = \frac{26}{26+0} = 1$$

3. Recall:

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

$$\text{Recall} = \frac{26}{26+0} = 1$$

4. F1-Score:

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

$$\text{F1-Score} = 2 \times \frac{1 \times 1}{1+1} = 1$$

Hasil Metrik:

- Accuracy: 0.980 (98%)
- Precision: 1 (100%)
- Recall: 1 (100%)
- F1-Score: 1 (100%)

Data menunjukkan kinerja yang sangat baik dengan precision, recall, dan F1-Score masing-masing 100%, dan akurasi 98%.

3.8 Perhitungan Algoritma C4.5

NO	Jumlah Keluarga	Penghasilan	Pendidikan	Sts Rumah	Pekerjaan	Kesehatan	Sts Penerima
1	Keluarga Kecil	Sedang	SMP	Sedang	lain-lain	Kurang	Ya
2	Keluarga Besar	Sedang	Tidak Sekolah	Sedang	Wirausaha	Baik	Tidak
3	Keluarga Sedang	Kurang	SMP	Kurang	lain-lain	Cukup	Ya
4	Keluarga Sedang	Kurang	Tidak Sekolah	Sedang	lain-lain	Cukup	Ya
5	Keluarga Kecil	Tinggi	SMA	Baik	Wirausaha	Baik	Tidak
....
....
59	Keluarga Kecil	Tinggi	SMA	Baik	Wirausaha	Baik	Tidak
60	Keluarga Sedang	Tinggi	SMP	Baik	lain-lain	Baik	Tidak

Mengelompokkan setiap variabel untuk mendapatkan atribut yang akan digunakan untuk menghitung entropy dari kasus agar mendapatkan akar dan cabang dari penerima bantuan PKH.

Tabel 3.9 Perhitungan manual

keterangan	Jumlah Kasus	Status (Ya)	Status (Tidak)	Entropy	Gain
Total	60	30	30	1	
Jumlah Keluarga					0,1945957
Keluarga Kecil	15	5	10	0,918295834	
Keluarga Sedang	34	24	10	0,873981048	
Keluarga Besar	11	1	10	0,439496987	
Penghasilan					0,7509775
Kurang	22	23	0	0	
Sedang	16	9	9	0,933834373	
Tinggi	22	0	23	0	
Pendidikan					0,2217565
SD	19	14	5	0,831474388	
SMP	24	12	12	1	
SMA	8	0	8	0	
Tidak Sekolah	7	4	3	0,985228136	
SARJANA	2	0	2	0	
Status Rumah					0,2855718
Kurang	11	7	4	0,945660305	
Sedang	35	23	12	0,927526588	
Baik	14	0	14	0	
Pekerjaan					0,3527269
lain-lain	41	22	19	0,996134484	
petani	6	8	5	-0,33418799	
Wirausaha	13	0	6	0	
Kesehatan					0,2630368
Kurang	10	7	3	0,881290899	
Cukup	37	23	14	0,956888666	
Baik	13	0	13	0	

1. menghitung nilai entropy dari semua atribut

Berikut adalah nilai entropy:

Pertama adalah mencari entropy dari total data yang ada, jumlah data yang digunakan adalah 60 data terbagi menjadi 30 “Ya” dan 30 “Tidak”, berikut contoh perhitungannya:

a. *Entropy* Total

$$= Entropy(S) \sum_{i=1}^{n} (-P_i \log_2 P_i)$$

$$= ((-30/60) * \log_2(30/60)) + ((-30/60) * \log_2(30/60))$$

$$= 1$$

b. *Entropy* Jumlah Keluarga

- Keluarga kecil = 0,918295834
- keluarga Sedang = 0,873981048
- Keluarga Besar = 0,439496987

c. *Entropy* Penghasilan

- Kurang = 0
- Sedang = 0,933834373
- Tinggi = 0

d. *Entropy* Pendidikan

- Sd = 0,831474388
- Smp = 1
- Sma = 0
- Tidak Sekolah = 0,985228136
- Sarjana = 0

e. *Entropy* Setatus Rumah

- Kurang = 0,945660305
- Sedang = 0,927526588
- Baik = 0

f. *Entropy* Pekerjaan

- Lain-lain = 0,996134484
- Petani = 0,961236605
- Wirausaha = 0

g. *Entropy* Kesehatan

- Kurang = 0,881290899
- Cukup = 0,956888666
- Baik = 0

2. Menghitung Nilai Gain Dari Semua Atribut

Berikut adalah nilai gain dari setiap atribut:

$$Gain = Entropy(S) - \sum_{i=1}^{|S|} P_i * Entropy(S_i)$$

a. Nilai Gain Jumlah Keluarga

$$= 0,194595667$$

b. Nilai Gain Penghasilan

$$= 0,7509775$$

c. Nilai Gain Pendidikan

$$= 0,221756495$$

d. Nilai Gain Setatus Rumah

$$= 0,285571768$$

e. Nilai Gain Pekerjaan

$$= 0,111040172$$

f. Nilai Gain Kesehatan

$$= 0,26303684$$

Berdasarkan perhitungan diatas dipilih atribut yang memiliki nilai gain tertinggi untuk dijadikan sebagai akar. Pada tabel 3.72 terlihat atribut penghasilan akan dipilih sebagai akar. Terdapat 3 atribut pada penghasilan yaitu kurang, sedang, tinggi. dari ketiga atribut tersebut, nilai atribut kurang sudah mengklasifikasikan kasus menjadi 1 yaitu keputusannya Ya dan atribut kurang sudah

mengklasifikasikan kasus menjadi Tidak, sehingga tidak perlu dilakukan perhitungan lagi. Untuk nilai atribut sedang masih perlu dilakukan perhitungan lagi.

Selanjutnya menghitung nilai dari atribut jumlah keluarga yang memiliki nilai gain lebih besar dan hasil perhitungan dapat dilihat pada tabel

Tabel 3.10 Node 1.1

keterangan	Jumlah Kasus	Status (Ya)	Status (Tidak)	Entropy	Gain
Total	60	30	30	1	
Jumlah Keluarga					0,8622556
Keluarga Kecil	2	2	0	0	
Keluarga Sedang	9	6	3	0,918295834	
Keluarga Besar	5	0	5	0	
Pendidikan					0,7883856
SD	6	5	1	0,650022422	
SMP	7	2	5	0,863120569	
SMA	0	0	0	0	
Tidak Sekolah	3	1	2	0,918295834	
SARJANA	0	0	0	0	
Status Rumah					0,7917321
Kurang	1	0	1	0	
Sedang	13	8	5	0,961236605	
Baik	2	0	2	0	
Pekerjaan					0,7513386
lain-lain	9	5	4	0,99107606	
petani	6	3	3	1	
Wirausaha	1	0	1	0	
Kesehatan					0,7748371
Kurang	6	4	2	0,918295834	
Cukup	8	4	4	1	
Baik	2	0	2	0	

Selanjutnya menghitung nilai dari atribut pendidikan untuk dijadikan node percabangan dari nilai atribut sd, smp dan tidak sekolah. Perhitungan dilakukan dengan cara sama dan hasil perhitungan dapat dilihat pada tabel 3.11.

Tabel 3.11 Node 1.111

keterangan	Jumlah Kasus	Status (Ya)	Status (Tidak)	Entropy	Gain
Total	60	30	30	1	
Pendidikan					0,4034203
SD	18	14	4	0,764204507	
SMP	18	12	6	0,918295834	
SMA	3	0	3	0	
Tidak Sekolah	6	4	2	0,918295834	
SARJANA	1	0	1	0	
Pekerjaan					0,3244838
lain-lain	35	22	13	0,951762676	
petani	10	8	2	0,721928095	
Wirausaha	1	0	1	0	
Kesehatan					0,3959724
Kurang	10	7	3	0,881290899	
Cukup	32	23	9	0,857148437	
Baik	4	0	4	0	

Dari hasil perhitungan entropy dan information gain yang didapat kemudian diolah kedalam decision tree, berikut hasil pohon keputusan (Decision Tree).

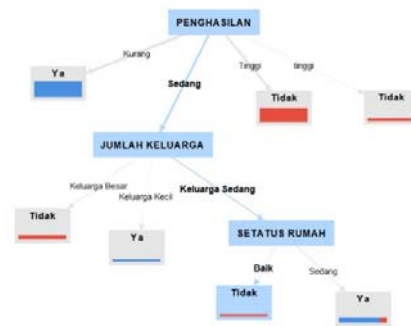
1. Entropy dan Gain:

Tabel 3.12 Entropy dan Gain

Atribut	Entropy	Gain
Jumlah Keluarga	0,918 - 0,439	0,1946
Penghasilan	0 - 0,933	0,751
Pendidikan	0 - 0,985	0,2218
Status Rumah	0 - 0,946	0,2856
Pekerjaan	0 - 0,996	0,3527
Kesehatan	0 - 0,956	0,263

2. Pohon Keputusan C4.5:

Berdasarkan nilai gain tertinggi, atribut Penghasilan dengan gain 0,7509775 menjadi akar (root) dari pohon keputusan, diikuti oleh atribut lain yang memiliki gain tertinggi berikutnya.



3.4 Pohon Keputusan

3. Evaluasi Model:

Jika model sudah dibangun dan diuji dengan data uji (testing data), kita bisa menghitung Accuracy, Precision, Recall, dan F1-Score.

4. Menghitung Accuracy, Precision, Recall, dan F1-Score:

Untuk menghitung metrik tersebut, Anda memerlukan informasi mengenai:

- True Positives (TP): Kasus yang diprediksi "Ya" dan memang benar "Ya".
- True Negatives (TN): Kasus yang diprediksi "Tidak" dan memang benar "Tidak".
- False Positives (FP): Kasus yang diprediksi "Ya" tetapi seharusnya "Tidak".
- False Negatives (FN): Kasus yang diprediksi "Tidak" tetapi seharusnya "Ya".

$$\text{Accuracy} \left(\frac{25 + 28}{25 + 28 + 5 + 2} \right) = \left(\frac{53}{60} \right) = 93\%$$

$$\text{precision} \left(\frac{25}{25 + 5} \right) = \left(\frac{25}{30} \right) = 83,3\%$$

$$\text{Recall} \left(\frac{25}{25 + 2} \right) = \left(\frac{25}{27} \right) = 92,6\%$$

$$\text{F1 - Score} \left(\frac{2 \times 0,833 \times 0,926}{0,833 + 0,926} \right) = 87,7\%$$

5. Support:

Support adalah jumlah kasus aktual dari setiap kelas dalam dataset. Misalnya, dalam data Anda, terdapat 30 kasus "Ya" dan 30 kasus "Tidak".

Ringkasan:

Hasil evaluasi model C4.5 dapat disajikan seperti berikut:

Accuracy: 93%

Precision: 83.3%

Recall: 92.6%

F1-Score: 87.7%



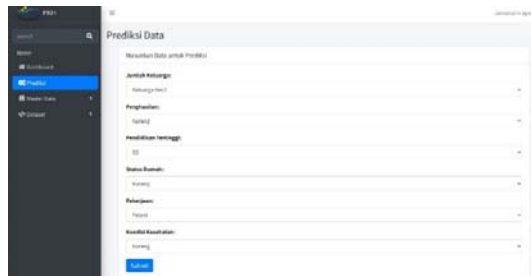
3. Hasil dan Pengujian

Halaman dashboard merupakan halaman pertama yang tampil setelah user melakukan login. Adapun implementasi dari halaman dashboard dapat dilihat pada Gambar 3.1 Halaman Dashboard.



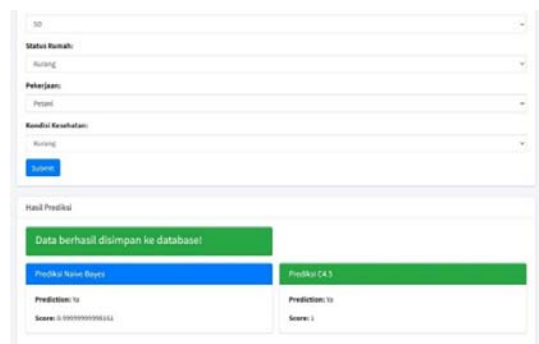
Gambar 3.11 Halaman Dashboard

Halaman Prediksi merupakan halaman untuk memprediksi kelayakan. Adapun implementasi dari halaman prediksi dapat dilihat pada Gambar 3.2 Halaman Prediksi.



Gambar 3.2 Halaman Prediksi

Halaman hasil perbandingan Prediksi merupakan halaman untuk menampilkan hasil prediksi algoritma naïve bayes dan C4.5 dan disini dapat melihat score dari kedua algoritma. Adapun implementasi dari halaman hasil perbandingan prediksi dapat dilihat pada Gambar 3.3



Gambar 3.4 Halaman Hasil Perbandingan Prediksi

Halaman Sub Menu predictions merupakan halaman yang berisi data yang sudah di prediksi terlihat pada Gambar 3.4

Kategori Keluarga	Penghasilan	Pendidikan	Maka Besar	Pekerjaan	Kategori Kesehatan	Prediksi Model Bayes	Probabilitas	Score Model Bayes	Score	Aksi
Keluarga Besar	Tringgi	SD	Baik	Musikista	Kurang	Tidak	Tidak	3	3	delete
Keluarga Kecil	Kurang	SD	Kurang	pakani	Kurang	Ya	Ya	3	3	delete
Keluarga Kecil	Kurang	SD	Kurang	pakani	Kurang	Ya	Ya	3	3	delete
Keluarga Kecil	Kurang	SD	Kurang	pakani	Kurang	Ya	Ya	3	3	delete
Keluarga Kecil	Kurang	SD	Kurang	pakani	Kurang	Ya	Ya	3	3	delete
Keluarga Kecil	Kurang	SD	Kurang	pakani	Kurang	Ya	Ya	3	3	delete

Gambar 3.5 Halaman Sub Menu Data (Predictions).

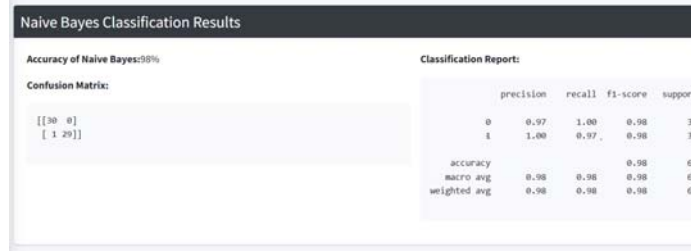
Halaman Sub Menu Data Latih merupakan halaman yang berisi data latih dimana user dapat melakukan tambah, edit, hapus data dan melakukan import data berupa csv. Adapun implementasi dari halaman sub menu data latih dapat dilihat pada Gambar 3.5 Halaman Sub Menu Data (Data Latih).

NAMA	NIK	ALAMAT	JUMLAH KELUARGA	PENGHASILAN	PENDIDIKAN TERTINGGI	STATUS	PEKERJAAN	KEMUDI KESEHATAN
AIN ANJAL	32102040707002	KP PANGKALAN BUNDA RT 023	Keluarga Sedang	Kurang	SD	Sedang	Siswa	Kurang
AIN DAMBATI	32102040874004	KP PANGKALAN BUNDA RT 023	Keluarga Sedang	Kurang	Tidak Sekolah	Sedang	Siswa	Cukup
AIN DAMBATI	32102040836003	KP PANGKALAN BUNDA RT 023	Keluarga Sedang	Kurang	SD	Kurang	Siswa	Cukup
AIN MAHSUDIN	32102040899000	KP PANGKALAN BUNDA RT 023	Keluarga Sedang	Tringgi	SD	Baik	pakani	Baik
AIN YUNARATI	32102040899000	KP PANGKALAN BUNDA RT 023	Keluarga Sedang	Tringgi	SD	Baik	Musikista	Baik
AINUL AZIZ	32102041100000	KP PANGKALAN BUNDA RT 023	Keluarga Besar	Sedang	SD	Sedang	Siswa	Kurang

Gambar 3.6 Halaman Sub Menu Data (Latih).

Halaman Sub Menu Data Latih merupakan halaman yang berisi data latih dimana user dapat melakukan tambah, edit, hapus data dan melakukan import data berupa csv. Adapun implementasi dari halaman sub menu data latih dapat dilihat pada Gambar 3.6 Halaman Sub Menu Data (Data Latih).

Hasil pengujian algoritma naïve bayes



Naive Bayes Classification Results

Accuracy of Naive Bayes: 98%

Confusion Matrix:

```
[[ 30  0]
 [ 1 29]]
```

Classification Report:

	precision	recall	f1-score	support
0	0.97	1.00	0.98	30
1	1.00	0.97	0.98	30
accuracy			0.98	60
macro avg	0.98	0.98	0.98	60
weighted avg	0.98	0.98	0.98	60

Gambar 4. 2 Tabel Pengujian Sistem Berdasarkan Akurasi NB

Hasil pengujian algoritma C4.5



C4.5 (Decision Tree) Classification Results

Accuracy of C4.5: 0.9333333333333333

Confusion Matrix:

```
[[ 29  1]
 [ 3 27]]
```

Classification Report:

	precision	recall	f1-score	support
0	0.91	0.97	0.94	30
1	0.96	0.90	0.93	30
accuracy			0.93	60
macro avg	0.94	0.93	0.93	60
weighted avg	0.94	0.93	0.93	60

Gambar 4. 3 Tabel Pengujian Sistem Berdasarkan Akurasi C4.5

4. Kesimpulan

Berdasarkan hasil penelitian dan laporan tugas akhir yang telah dilakukan, maka dapat ditarik kesimpulan bahwa Machine learning Penyortiran Buah Naga Berbasis Internet of Things Menggunakan Metode K-Means Pada Platform Blynk, maka penulis dapat menarik beberapa kesimpulan berdasarkan dari hasil penelitian sebagai berikut :

- Alat ini dapat membantu para petani buah dalam proses penyortiran.
- Sistem pendeteksi kualitas berat dan warna dapat berhasil terealisasi menggunakan sensor Load Cell untuk menghitung berat dan sensor TCS230 untuk menentukan warna. Dengan akurasi keberhasilan penimbangan sebesar 99% dan hanya memiliki error sebesar 1%. Serta sensor TCS3200 dapat mendeteksi warna dengan baik.
- Sistem penyortiran buah naga yang penulis kerjakan telah berfungsi dan berjalan baik sesuai dengan harapan.
- Data yang didapat oleh alat dapat dihitung menggunakan Algoritma K-means dengan benar.

Daftar Pustaka

- [1] R. W. Sinaga and R. Winanjaya, "Analisis Data Mining Menentukan Penerima Bantuan Langsung Tunai pada Desa Pamatang Purba dengan Algoritma C4.5," Jurnal Teknologi dan Sistem Informasi, vol. 3, no. 1, pp. 1–9, 2021.
- [2] G. Ramadhan et al., "Penerapan Data Mining Menggunakan Algoritma C4.5 Dalam Mengukur Tingkat Kepuasan Pasien BPJS," Prosiding Seminar Nasional Riset dan Information Science (SENARIS), vol. 2, pp. 376–385, 2020.
- [3] L. Bachtiar and M. Mahradianur, "Analisis Data Mining Menggunakan Metode Algoritma C4.5 Menentukan Penerima Bantuan Langsung Tunai," Jurnal Informatika, vol. 10, no. 1, pp. 1–9, 2023.
- [4] Muhamad, M., Windarto, A. P., & Suhada, S. (2019). Penerapan Algoritma C4.5 Pada Klasifikasi Potensi Siswa Drop Out. KOMIK (Konferensi Nasional Teknologi Informasi Dan Komputer), 3(1), 1–8. <https://doi.org/10.30865/komik.v3i1.1688>
- [5] N. Rahmadani, R. Risnawati, and M. D. Sena, "Penerapan Algoritma Naïve Bayes dalam Penentuan Kelayakan Penerima Bantuan Program Keluarga Harapan," Jurnal Teknisi, vol. 3, no. 2, pp. 40–48, 2023.

- [6] R. S. Bhat and R. Bhat, "A Comparative Analysis of Machine Learning Algorithms for Classification Purpose," *Procedia Computer Science*, vol. 199, pp. 372–380, 2022. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1877050922021159>
- [7] S. Rasika, "Penerapan Algoritma C4.5 dalam Menentukan Faktor Kelayakan Penerima Bantuan Langsung Tunai (BLT-DD)," *Jurnal Ilmiah Tunas Bangsa*, vol. 1, no. 1, pp. 45–52, 2022.
- [8] E. Fitri, "Perbandingan Algoritma C4.5 Dan Naïve Bayes Untuk Menentukan Kelayakan Penerima Bantuan Program Keluarga Harapan," *Jurnal Sistemasi Sistem Informasi*, vol. 9, No. 1, pp. 103-115, 2020.
- [9] J. R. Quinlan, "Improved Use of Continuous Attributes in C4.5," *Journal of Artificial Intelligence Research*, vol. 4, pp. 77–90, 1996. [Online]. Available: <https://jair.org/index.php/jair/article/view/10157>
- [10] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. Springer, 2009.
- [11] T. Wibowo, "Penerapan Data Mining Pemilihan Siswa Kelas Unggulan dengan Metode K-Means Clustering di SMP N 02 Tasikmadu," *Jurnal Ilmu Komputer*, vol. 5, no. 1, pp. 27–36, 2018.
- [12] A. Chandra, "Penerapan Data Mining Menggunakan Pohon Keputusan," *Jurnal Ilmu Komputer*, vol. 6, no. 3, pp. 1–6, 2017.