

# PENERAPAN *K-NEAREST NEIGHBOR IMPUTATION* UNTUK PERBAIKAN *MISSING VALUE* PADA KLASIFIKASI PENDUDUK KURANG MAMPU MENGGUNAKAN METODE *NAÏVE BAYES*

Deni Arifianto, M.Kom<sup>1</sup>, dan Anik Nur Novitasari Eka Septianingrum<sup>2</sup>

<sup>1</sup>Universitas Muhammadiyah Jember, Jember, Indonesia

<sup>2</sup>Universitas Muhammadiyah Jember, Jember, Indonesia

deniarifianto@unmuhjember.ac.id, aniknumovitasari@gmail.com

Abstrak, Pada tahun 2016 di Kabupaten Jember masih tercatat 265.000 jiwa penduduk kurang mampu dimana klasifikasi penduduk kurang mampu tersebut didasarkan pada beberapa parameter yang diantaranya terdiri dari pendapatan, jumlah makan sehari, pendidikan, luas bangunan, program yang diterima, status kepemilikan rumah, sumber penerangan dan MCK. Untuk mengklasifikasi data penduduk kurang mampu tersebut bisa digunakan metode *Naïve Bayes*. Akan tetapi dalam prosesnya terdapat kendala yaitu tidak lengkapnya data-data dari kelurahan yang membuat proses klasifikasi tidak dapat dilakukan secara optimal. Untuk itu dibutuhkan preproses untuk dapat melengkapi data-data kosong tersebut. Untuk memprediksi *missing value* pada data penduduk kurang mampu tersebut. Metode *k-nearest neighbor* akan diaplikasikan guna melengkapi *missing value* dalam data penduduk tersebut dan akan dikombinasikan dengan klasifikasi *Naïve Bayes* dalam pengambilan keputusan dapat membantu mengatasi permasalahan *missing value* dalam data penduduk kurang mampu di Kabupaten Jember sekaligus mengetahui pengaruh data yang kosong bila dibandingkan parameter-parameter lain terhadap pengambilan keputusan. Setelah ujicoba dilakukan, maka mengetahui tingkat akurasi metode *k-nearest neighbor* dalam memprediksikan data kosong sebesar 80,4% sehingga dikatakan cukup baik untuk dapat diterapkan pada data yang selanjutnya akan diklasifikasikan menggunakan metode *Naïve Bayes*. Pada proses klasifikasi penduduk kurang mampu menggunakan metode *Naïve Bayes* setelah divalidasi memperoleh hasil cukup tinggi yaitu akurasi sebesar 87% sedangkan proses klasifikasi yang memanfaatkan *k-nearest neighbor* untuk prediksi *missing value* mendapatkan nilai akurasi yaitu 85%.

## LATAR BELAKANG

Penduduk kurang mampu merupakan suatu masalah yang dialami beberapa Negara, termasuk Indonesia. Masalah kemiskinan adalah masalah lama yang belum terselesaikan. Meskipun sudah begitu banyak bantuan bagi penduduk kurang mampu yang di berikan oleh pemerintah, seperti halnya Jaminan Kesehatan Nasional, Kartu Indonesia Sehat (KIS), Kartu Indonesia Pintar (KIP) serta jenis bantuan lainnya, namun hal itu masih belum mampu mengatasi sepenuhnya masalah kemiskinan di Indonesia. Salah satu contohnya, pada tahun 2016 di Kabupaten Jember masih tercatat 265.000 jiwa penduduk kurang mampu dimana klasifikasi penduduk kurang mampu tersebut didasarkan pada beberapa parameter yang diantaranya terdiri dari pendapatan, jumlah makan sehari, pendidikan, luas bangunan, program yang diterima, status kepemilikan rumah, sumber penerangan dan MCK.

Dari penelitian yang dilakukan oleh Haditsah Annur (2018) menyatakan bahwa kinerja algoritma *Naïve Bayes* terhadap dataset yang telah diambil diperoleh tingkat akurasi sebesar 73% atau termasuk

dalam kategori bagus (*good*). Namun, penelitian tersebut masih memiliki kekurangan. Kekurangan penelitian tersebut salah satunya adalah tidak ada *pre-processing* untuk melengkapi data masukan yang kosong (*missing value*). Sedangkan di lapangan, data survei atau data di lapangan sering kali tidak lengkap atau yang sering disebut *missing value* atau data kosong yang terdapat di antara beberapa parameter. Data kosong seringkali terjadi akibat tidak adanya respon terhadap item-item yang dibutuhkan. Kurangnya data di salah satu parameter sangat berpengaruh pada parameter-parameter yang lain. Kurang lengkapnya data ini selalu menjadi permasalahan umum dalam melakukan survey dan menyebabkan data penelitian menjadi tidak lengkap.

Untuk mengklasifikasi data penduduk kurang mampu tersebut bisa digunakan beberapa metode, misalnya metode *Naïve Bayes*. *Naïve Bayes* adalah metode sederhana klasifikasi probabilitas yang menghitung serangkaian probabilitas dengan menambahkan frekuensi dan kombinasi nilai dari dataset yang diberikan. Algoritma teorema bayes digunakan dalam metode ini dan mengasumsikan

semua atribut independen yang diberikan oleh nilai pada variabel kelas (Patil dan Sherekar, 2013).

Peneliti merasa perlu untuk melakukan laporan tentang definisi di atas yaitu tentang *missing value*, untuk memprediksi *missing value* pada data penduduk kurang mampu tersebut. Untuk perbaikan *missing value* dalam data penduduk tersebut, akan menggunakan metode *k-nearest neighbor*. Penanganan missing data dengan KNN (*K-Nearest Neighbor*) diawali dengan menentukan sejumlah tetangga terdekat atau observasi terdekat yang disimbolkan dengan K, kemudian menghitung jarak terkecil dari setiap observasi yang tidak mengandung missing data (Martha Shantika dkk, 2018).

Dengan memperbaiki missing value menggunakan metode k-nearest neighbor yang dikombinasikan dengan klasifikasi *naïve bayes* dalam pengambilan keputusan, diharapkan dapat membantu mengatasi permasalahan *missing value* dalam data penduduk kurang mampu di Kabupaten Jember sekaligus mengetahui pengaruh data yang kosong bila dibandingkan parameter-parameter lain terhadap pengambilan keputusan.

## PENDUDUK KURANG MAMPU

Kemiskinan diartikan sebagai standar hidup yang lebih rendah dari standar hidup kebanyakan. Dengan kata lain, tingkat kelangkaan komoditas dibandingkan dengan kondisi kehidupan umum di masyarakat yang bersangkutan. Kemiskinan juga dapat diartikan secara ekonomi sebagai tidak adanya sumber daya yang dapat digunakan untuk meningkatkan kesejahteraan sekelompok orang.

Berdasarkan Badan Pusat Statistik, Kemiskinan adalah kegagalan untuk memenuhi kebutuhan dasar minimum, termasuk kebutuhan makanan atau selain kebutuhan makanan. Kemiskinan dapat diukur dengan membandingkan tingkat konsumsi seseorang dengan garis kemiskinan atau jumlah rupiah yang dihabiskan per bulan untuk konsumsi. Sedangkan masyarakat miskin adalah orang-orang yang memiliki garis kemiskinan di bawah perkiraan pengeluaran per kapita per bulan. Penentuan ukuran garis kemiskinan lokal adalah Rp. 7.050 per orang per hari pengukuran garis kemiskinan, termasuk kebutuhan nutrisi dan bukan makanan. Persyaratan makanan minimum termasuk 2.100 kilokalori per orang per hari. Garis kemiskinan non makanan adalah kebutuhan minimum untuk perumahan, pendidikan, dan kesehatan (rendahnya konsumsi makanan bergizi, kurangnya sarana kesehatan (Supriatna, 2017).

Dari survey di Kantor Badan Pusat Statistik Kabupaten Jember, data pendukung untuk penentuan

penduduk kurang mampu menggunakan parameter-parameter diantaranya adalah kondisi rumah, umur, pendidikan, jumlah keluarga, pengeluaran, kesehatan dan penghasilan (BPS, 2016).

## METODE K-NEAREST NEIGHBOR

Adalah Sebuah sistem yang menggunakan algoritma pembelajaran terawasi dan bertujuan menemukan pola data baru dengan menghubungkan pola data yang ada dengan data baru. Ada dua jenis algoritma NN yaitu 1NN dan KNN. 1NN atau Nearest Neighbor adalah pendekatan yang melakukan klasifikasi pada 1 data terdekat, sedangkan KNN adalah pendekatan yang melakukan klasifikasi pada K data terdekat, dengan  $K > 1$ .

KNN merupakan pendekatan yang digunakan untuk mengidentifikasi objek berdasarkan informasi tertentu yang merupakan jarak terdekat ke objek. Pada klasifikasi, KNN bekerja dengan menghitung jarak antara data baru (data testing) dengan data yang sudah diketahui kelasnya (data training) menggunakan jarak euclidian. Penanganan missing data dengan KNN diawali dengan menentukan sejumlah tetangga terdekat atau observasi terdekat yang disimbolkan dengan K, kemudian menghitung jarak terkecil dari setiap observasi yang tidak mengandung missing data. Langkah-langkah imputasi missing data dengan metode KNN adalah sebagai berikut:

1. Tentukan parameter K, K adalah jumlah observasi terdekat atau tetangga terdekat yang akan digunakan.
2. Menghitung jarak antara observasi yang mengandung missing data dengan observasi lengkap pada variabel ke-j yang tidak mengandung missing data dengan variabel j lainnya yang bersesuaian dengan rumus jarak euclidian yaitu:

$$d(x_a, x_b) = \sqrt{\sum_{j=1}^m (x_{aj} - x_{bj})^2}$$

$d(x_a, x_b)$  adalah Kesenjangan antara hasil yang berisi data yang hilang dan bukan data yang hilang,  $x_{aj}$  adalah nilai dari variabel ke-j pada setiap observasi yang mengandung missing data dengan  $j = 1, 2, \dots, m$ ,  $x_{bj}$  adalah nilai dari variabel lainnya pada setiap observasi yang tidak mengandung missing data dengan  $j = 1, 2, \dots, m$ .

3. Urutkan jarak berdasarkan observasi yang memiliki nilai jarak terbesar sampai observasi yang memiliki nilai jarak terkecil.
4. Menentukan sejumlah K observasi terdekat berdasarkan nilai jarak terkecil.
5. Melakukan

imputasi missing data dengan menghitung nilai weight mean estimation pada K observasi terdekat yang tidak mengandung nilai missing data dengan rumus:

$$x_j = \frac{\sum_{k=1}^k w_k v_k}{\sum_{k=1}^k w_k}$$

dimana  $x_j$  adalah estimasi rata-rata berbobot,  $v_k$  adalah nilai pada data lengkap pada variabel yang mengandung missing data berdasarkan observasi dari k, K adalah jumlah observasi terdekat yang digunakan, k adalah observasi dari K,  $w_k$  adalah bobot observasi tetangga terdekat ke-K dengan rumus

$$w_k = \frac{1}{d(x_{ak} x_{bk})^2}$$

dimana  $d(x_{ak} x_{bk})$  adalah jarak observasi K (Martha Shantika dkk, 2018).

### METODE NAÏVE BAYES

Bayes adalah teknik prediksi probabilistik dasar berdasarkan aplikasi luas teorema Bayes (atau aturan Bayes) dengan kata lain, *Naïve Bayes*, model yang digunakan adalah "model fungsi independensi." Fitur-fitur dalam data tidak ditautkan ke apakah fitur-fitur lain dalam data yang sama atau tidak. Teorema Bayes dengan persamaan umum berikut untuk perkiraan gaji didasarkan pada:

$$P(H|E) = \frac{P(E|H) \cdot P(H)}{P(E)}$$

Penjelasan dari formula tersebut adalah sebagai berikut :

Parameter	Keterangan
P(H E)	Probalitas akhir bersyarat (conditional probability) suatu hipotesis H terjadi jika diberikan bukti (evidence) E terjadi.
P(E H)	Probabilitas sebuah bukti E terjadi akan mempengaruhi hipotesis H.
P(E)	Probabilitas awal (priori) bukti E terjadi tanpa memandang hipotesis/bukti lain.

Bayes didasarkan pada beberapa bukti (E) yang dapat ditemukan berdasarkan hasil dari hipotesis atau kejadian (H). Hukum Bayes mencakup beberapa hal penting, yaitu:

1. Probabilitas kesimpulan sebelum bukti dipenuhi adalah probabilitas pertama H atau P (H);

2. Probabilitas kesimpulan setelah bukti ditemukan adalah kemungkinan akhir H atau P (H E).

#### 2.5.1 Karateristik *Naïve Bayes*

Kategori *Naïve Bayes* didasarkan pada teori probabilitas, yang mempertimbangkan semua karakteristik data sebagai bukti probabilitas. Karenanya *Naïve Bayes* menawarkan fitur-fitur berikut:

1. Metode *Naïve Bayes* sangat bergantung pada pengetahuan yang terisolasi, biasanya data dengan fitur yang berbeda (outliner). *Naïve Bayes* juga dapat mengelola nilai atribut yang salah dengan menyisihkan data pelatihan selama pembuatan model dan implementasi prediksi.
2. Tangguh menghadapi karakteristik yang tidak signifikan.

Atribut yang memiliki korelasi dapat menurunkan kualitas kategori Naïve Bayes karena tidak memiliki otonomi. (F. Arfiana, 2014).

### CONFUSION MATRIKS

Dalam penelitian ini menggunakan *Confusion Matriks* untuk mengukur hasil klasifikasi. Diantaranya yaitu akurasi, akurasi merupakan pengukuran bagaimana kekakuratan dari suatu klasifikasi. Suatu klasifikasi dikatakan baik jika nilai akurasi mendekati 100%, baik dari percobaan pertama hingga seterusnya. Berikut Tabel *Confusion Matriks*.

Tabel 2.1 Tabel Confusion Matriks

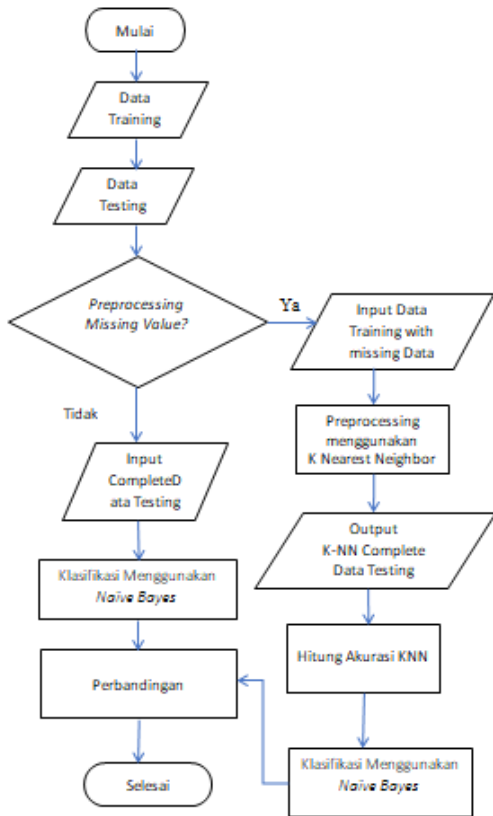
Classification	Predicted class	
	Class = Yes	Class = No
Class = Yes	a (true positive - TP)	c (false negative - FN)
Class = No	c (false positive - FP)	d (true negative - TN)

Dimana untuk mencari nilainya dengan rumus :

$$\text{Akurasi} = \frac{TP+TN}{TP+TN+FP+FN} * 100\%$$

### DESAIN

Proses desain merupakan proses penggambaran, perancangan dan pembuatan penerapan Algoritma KNN untuk *melengkapi missing value* dan *Naïve Bayes* untuk klafisikasi data penduduk miskin. Berikut merupakan alur system pada penelitian ini:



Gambar 3.1 Alur Sistem

Pada gambar 3.1 menggambarkan alur pada data set dilakukan *preprocessing* dengan melengkapi data kosong menggunakan *k-nearest neighbor*. Data yang telah lengkap di proses kembali dengan metode *naive bayes* untuk mendapatkan keputusan pada data penduduk kurang mampu.

#### Proses Perhitungan K-Nearest Neighbor

Proses metode *k-nearest neighbor* digunakan untuk mengatasi adanya *missing value*, data yang digunakan sebanyak 15 data penduduk kurang mampu dan 3 data tidak lengkap. Contoh perhitungan dengan metode *k-nearest neighbor* adalah sebagai berikut:

- Hitung jumlah data yang dipakai yaitu sebanyak 15 data penduduk kurang mampu. Dimana terdapat data kosong yang akan diproses terlebih dahulu menggunakan metode *k-nearest neighbor*.
- Menghitung nilai bobot disetiap atribut.

Tabel 3.1 Pembobotan Pendapatan

PENDAPATAN		
Kriteria	Penamaan	Bobot

>1500000	Tinggi	4
1000000-1500000	Sedang	3
600000-1000000	Cukup	2
<600000	Kurang	1

Tabel 3.2 Pembobotan Status Kepemilikan Rumah

STATUS KEPEMILIKAN RUMAH		
Kriteria	Penamaan	Bobot
Milik Sendiri	Milik Sendiri	3
Sewa	Sewa	2
Menumpang	Menumpang	1

Tabel 3.3 Pembobotan Program yang diterima

PROGRAM YANG DITERIMA		
Kriteria	Penamaan	Bobot
<2	Sedikit	4
2	Kurang	3
3	Cukup	2
>3	Banyak	1

Tabel 3.4 Pembobotan Luas Bangunan Rumah

LUAS BANGUNAN RUMAH		
Kriteria	Penamaan	Bobot
>12 M	Sangat Luas	4
ANTARA 10 - 12M	Luas	3
ANTARA 8 - 10M	Sedang	2
< 8M	Sempit	1

Tabel 3.5 Pembobotan Pendidikan

PENDIDIKAN		
Kriteria	Penamaan	Bobot
SMA	SMA	4
SMP	SMP	3
SD	SD	2
TIDAK SEKOLAH	TIDAK SEKOLAH	1

Tabel 3.6 Pembobotan Jumlah Makan Sehari

JUMLAH MAKAN SEHARI		
Kriteria	Penamaan	Bobot
>3	Lebih	4
3	Sedang	3
2	Cukup	2
<2	Kurang	1

Tabel 3.7 Pembobotan MCK

MCK		
Kriteria	Penamaan	Bobot
MEMILIKI DAN LAYAK	Punya Layak	4
MEMILIKI TETAPI TIDAK LAYAK	Punya Tidak Layak	3
MCK UMUM	Umum	2
TIDAK MEMILIKI	Tidak Punya	1

Tabel 3.8 Pembobotan Sumber Penerangan

SUMBER PENERANGAN		
Kriteria	Penamaan	Bobot
>450WATT MILIK SENDIRI	>450Watt Milik Sendiri	4
450 WATT MILIK SENDIRI	450 Watt Milik Sendiri	3
MENUMPANG	Menumpang	2
TIDAK MEMILIKI	Tidak Memiliki	1

- b. Menghitung jarak antara observasi yang mengandung *missing value* data dengan observasi yang tidak mengandung *missing value*.

Tabel 3.9 Contoh Data

No	Nama	Pendidikan	Program Yang Diterima	Status Rumah	Luas Bangunan	Fasilitas MCK	Sumber Penerangan	Jumlah Makan Sehari	Pendapatan	Kesimpulan
1	KASMLA	4	2	1	3	2	2	3	4	layak
2	HASANAH	1	1	1	2	4	1	1	2	tidak layak
3	A. JAMILA	2	1	1	4	2	2	2	3	tidak layak
4	SUPIYATUN	3	4	3	3	4	2	3	3	layak
5	MISTATIK	2	3	3	3	2	1	3	4	layak
6	LATIFAH	3	1	2	2	1	4	4	2	layak
7	SULIHATI	3	3	2	2	1	3	3	4	tidak layak
8	SALEH	3	2	2	4	4	3	4	1	layak
9	arwi	4	4	2	3	2	3	1	4	layak
10	supiyani	2	3	1	3	2	2	3	3	layak
11	huri	4	4	3	1	1	1	4	1	tidak layak
12	yuliani harisa	1	3	1	4	2	1	3	3	layak
13	EKA NOVI	2	4	2	1	3	3	3	3	layak
14	siti rudyana	3	2	3	3	3	4	3	2	tidak layak
15	misiyati	2	4	3	4	2	3	2	4	layak

D(program yang diterima) Kasmia=

$$\sqrt{(1-4)^2 + (0-2)^2 + (1-1)^2 + (2-3)^2 + (4-4)^2} = 4,3588$$

D(program yang diterima) A. Jamila=

$$\sqrt{(1-2)^2 + (0-1)^2 + (1-1)^2 + (2-4)^2 + (4-4)^2} = 2,2366$$

D(program yang diterima) Sapiyatun=

$$\sqrt{(1-3)^2 + (0-4)^2 + (1-3)^2 + (2-3)^2 + (4-4)^2} = 4,5677$$

Setelah dihitung observasi maka mengurutkan jarak berdasarkan observasi yang memiliki nilai jarak terbesar sampai observasi yang memiliki nilai jarak terkecil.

- Menentukan K observasi terdekat berdasarkan nilai terkecil.
- Melakukan imputasi missing data dengan menghitung nilai weight mean estimation pada K observasi terdekat yang tidak mengandung nilai missing data.

### HASIL DAN PEMBAHASAN

Data yang digunakan dalam penelitian ini sebanyak 600 data yang terdiri dari:

- 316 data layak
- 184 data tidak layak
- 100 data validasi yang terdiri dari layak dan tidak layak

Berikut tahapan penelitian dan uji coba, Gambar 4.1 memperlihatkan kondisi data sebagai berikut :

Mr. Nama	pendidikan	prgrn_olomc	sts_rumah	luas_bnr	fas_mnk	ombr_penerangan	mkn_seha	pendapatan	kependudukan
1.UTIMASAH	SD	2	Milik Sendiri	Antara 8-11Memiliki Tidak Layak	450Watt Milik Sendiri	3	kurang 600.000	layak	
2.HOTIMAH	SD	2	Milik Sendiri	Antara 8-11Memiliki Tidak Layak	450Watt Milik Sendiri	3	kurang 600.000	layak	
3.BANTEM	SD	2	Milik Sendiri	Antara 8-11Memiliki Tidak Layak	450Watt Milik Sendiri	3	kurang 600.000	layak	
4.SAGDA	SD	2	Milik Sendiri	kurang 8MMemiliki Tidak Layak	Tidak Memiliki	3	kurang 600.000	layak	
5.KASATUN	Tidak Sekolah	3	Milik Sendiri	Antara 8-11Memiliki Tidak Layak	450Watt Milik Sendiri	3	kurang 600.000	layak	
6.SIRI WATI	SD	lebih 3	Milik Sendiri	Antara 8-11Memiliki Tidak Layak	450Watt Milik Sendiri	3	kurang 600.000	layak	
7.SAKANE	SD	2	Milik Sendiri	Antara 8-11Memiliki Tidak Layak	450Watt Milik Sendiri	3	kurang 600.000	layak	
8.PATIMAH	SD	lebih 3	Milik Sendiri	kurang 8MMemiliki Tidak Layak	450Watt Milik Sendiri	3	kurang 600.000	layak	
9.PARIDA	SD	kurang 2	Milik Sendiri	Antara 8-11Memiliki Tidak Layak	450Watt Milik Sendiri	lebih 3	kurang 600.000	layak	
10.TIKH	SD	2	Milik Sendiri	Antara 8-11Memiliki Tidak Layak	450Watt Milik Sendiri	3	kurang 600.000	layak	
11.SULATI	SD	2	Milik Sendiri	Antara 8-11Memiliki Tidak Layak	450Watt Milik Sendiri	3	600.000 - 1.000.000	tidak layak	
12.SUPAN	SD	2	Milik Sendiri	Memiliki Tidak Layak	450Watt Milik Sendiri	3	kurang 600.000	layak	
13.KUMARAH	SD	2	Milik Sendiri	Antara 8-11Memiliki Tidak Layak	450Watt Milik Sendiri	3	1.000.000 - 1.500.000	layak	
14.SULIA PURWATI	SD	2	Milik Sendiri	Antara 8-11Memiliki dan Layak	450Watt Milik Sendiri	3	kurang 600.000	layak	
15.SUNDAI	Tidak Sekolah	2	Milik Sendiri	lebih 12MMemiliki Tidak Layak	450Watt Milik Sendiri	3	kurang 600.000	layak	
16.MARYAMAH	SD	2	Milik Sendiri	Antara 8-11Memiliki Tidak Layak	450Watt Milik Sendiri	3	kurang 600.000	layak	
17.LUMATI	SD	2	Milik Sendiri	Antara 8-11Memiliki Tidak Layak	450Watt Milik Sendiri	3	kurang 600.000	layak	
18.SYITRAH	Tidak Sekolah	3	Memumpang	kurang 8MMemiliki Tidak Layak	450Watt Milik Sendiri	kurang 2	kurang 600.000	layak	
19.TUM	Tidak Sekolah	2	Milik Sendiri	Antara 8-11Memiliki Tidak Layak	450Watt Milik Sendiri	3	kurang 600.000	layak	
20.SAMA	SD	2	Milik Sendiri	Memiliki Tidak Layak	450Watt Milik Sendiri	3	1.000.000 - 1.500.000	layak	

Gambar 4.2 Data Training

Dimana data terdiri dari 8 parameter yang telah ditentukan untuk proses klasifikasi. Pada proses implementasi data terlebih dahulu dikelompokkan/ dikategorikan, kemudian data dapat disimpan dengan format “.xls” dan data siap untuk diimport kedalam sistem yang telah dibuat. Dari data training tersebut ada beberapa data yang sengaja dikosongkan agar dapat mengetahui tingkat akurasi dari metode *k-nearest neighbor* dalam menangani masalah *missing value* dengan menggunakan data kosong berjumlah 240 data kosong dan 260 data lengkap. Implementasi Perangkat Lunak

Implementasi merupakan suatu kegiatan yang dilakukan untuk menguji data dan menerapkan pada sistem, sehingga akan diketahui apakah sistem yang dibuat benar-benar sesuai dengan yang direncanakan. Pada implementasi perangkat lunak ini akan dijelaskan bagaimana program sistem ini bekerja dengan menampilkan berbagai form yang telah dibuat.

Interface program berisi potongan gambar dari Tugas Akhir Penerapan *K-Nearest Neighbor Imputation* untuk Perbaikan *Missing Value* pada Klasifikasi Penduduk Kurang Mampu menggunakan Metode *Naive Bayes*. Berikut ini Gambar 4.2 menggambarkan program pada saat dijalankan:

ID	Nama	Nilai	Aksi
1	Pendapatan	kurang 600.000	kurang
2	Pendapatan	600.000 - 1.000.000	Cukup
3	Pendapatan	1.000.000 - 1.500.000	Sehingga
4	Pendidikan	lebih 3	Tinggi
5	Pendidikan	Tidak Sekolah	Tidak Sekolah
6	Pendidikan	SD	SD
7	Pendidikan	SDP	SDP
8	Pendidikan	SMK	SMK
9	Program Lang Diranca	lebih 1	Berapah
10	Program Lang Diranca	3	Cukup

Gambar 4.2 Halaman Master Kriteria

Pada halaman ini user dapat menambah dan mengedit kriteria yang digunakan untuk proses klasifikasi data penduduk kurang mampu. Fungsi dari master kriteria ini untuk mengkategorikan data asli yang sudah ada.

ID	Nama	Aksi
1	JEMBER SD 1 (JEMR)	+
2	JEMBER SD 1 (JEMR)	-
3	JEMBER SD 2 (JEMR)	+
4	JEMBER SD 2 (JEMR)	-
5	JEMBER SD 3 (JEMR)	+
6	JEMBER SD 3 (JEMR)	-
7	JEMBER SD 4 (JEMR)	+
8	JEMBER SD 4 (JEMR)	-
9	JEMBER SD 5 (JEMR)	+
10	JEMBER SD 5 (JEMR)	-

Gambar 4.3 Halaman Master Data

Gambar 4.3 berfungsi untuk input data penduduk kurang mampu. Data yang akan dimasukkan hanya file berekstensi xls. Saat proses import data asli yang berupa angka akan terkategori. Data yang terkategori yaitu pendapatan, program yang diterima, luas bangunan rumah dan jumlah makan dalam sehari. Setelah data diinput kedalam aplikasi maka dapat diproses lebih lanjut di menu data detail seperti pada gambar 4.4 dibawah ini.

No	Nama	Aksi
1	JEMBER 50 1 (KNN)	Proses NB, Proses KNN + NB, Detail, Hapus
2	JEMBER 50 1 (NB)	Proses NB, Proses KNN + NB, Detail, Hapus
3	JEMBER 50 2 (KNN)	Proses NB, Proses KNN + NB, Detail, Hapus
4	JEMBER 50 2 (NB)	Proses NB, Proses KNN + NB, Detail, Hapus
5	JEMBER 50 3 (KNN)	Proses NB, Proses KNN + NB, Detail, Hapus
6	JEMBER 50 3 (NB)	Proses NB, Proses KNN + NB, Detail, Hapus
7	JEMBER 50 4 (KNN)	Proses NB, Proses KNN + NB, Detail, Hapus
8	JEMBER 50 4 (NB)	Proses NB, Proses KNN + NB, Detail, Hapus
9	JEMBER 50 5 (KNN)	Proses NB, Proses KNN + NB, Detail, Hapus
10	JEMBER 50 5 (NB)	Proses NB, Proses KNN + NB, Detail, Hapus

Gambar 4.4 Halaman Data Detail

Pada menu data detail dapat melihat data yang telah diimport kedalam aplikasi dengan klik tombol detail. Kemudian data yang ada dapat diproses dengan 2 tombol yaitu Proses NB dan Proses KNN+NB. Tombol Proses NB berfungsi untuk klasifikasi data penduduk kurang mampu yang lengkap dengan menggunakan metode *naïve bayes*, namun tombol KNN+NB berfungsi untuk menghitung data yang tidak lengkap dan setelah data lengkap dilanjutkan proses klasifikasi data menggunakan metode *naïve bayes*. Berikutnya user dapat melihat hasil yang telah diproses pada Gambar 4.6.

No	Nama	Pendidikan	Program Yang Diterima	Status Rumah	Luas Bangunan	Fasilitas MCK	Sumber Penerangan	Jumlah Makanan	Pendapatan	Kesimpulan
1	JUMANAH	2	3	3	3	2	2	2	4	layak
2	HOTMAH	3	2	3	3	2	2	2	4	layak
3	SANPYEM	3	2	3	3	2	2	2	4	layak
4	SAODA	3	2	3	4	2	2	2	4	layak
5	KASIATUN	4	3	3	3	2	2	2	4	layak
6	SRI WIATI	3	4	3	3	2	2	2	4	layak
7	SANWINE	3	2	3	1	2	2	2	4	layak
8	FATMAH	3	4	3	4	2	2	2	4	layak
9	FARDA	3	1	3	3	2	2	1	4	layak
10	TIMA	3	2	3	1	2	2	2	4	layak

Gambar 4.5 Hasil Bobot

Pada halaman ini user dapat melihat data asli yang telah dibobotkan sesuai kriteria yang ada pada menu master kriteria. Setelah itu user dapat melihat hasil dari proses klasifikasi penduduk kurang mampu menggunakan metode *naïve bayes*.

Perhitungan *naïve bayes* dimulai dari mencari nilai class data seperti pada Gambar 4.6.

No	Nama	Jumlah
1	layak	38
2	tidak layak	12

Gambar 4.6 Hasil Class Data

Dalam proses *naïve bayes*, class data adalah proses paling awal yang harus dilakukan. Dimana user dapat mengetahui nilai class data dari data yang sudah diinput pada menu master data. Setelah nilai class data sudah ada maka proses selanjutnya mencari nilai probabilitas dari setiap kriteria. Berikut Gambar 4.7 menggambarkan perhitungan probabilitas data klasifikasi penduduk kurang mampu.

No	Jenis	Layak	Tidak Layak	(P) Layak	(P) Tidak Layak
1	Mak Sendiri	28	9	0.73684210526315	0.75
2	Sewa	3	2	0.078947368421053	0.16666666666667
3	Menumpang	4	1	0.10526315789474	0.083333333333333

Gambar 4.7 Hasil Probabilitas

Pada menu ini user dapat melihat nilai probabilitas dari masing-masing kriteria. Setelah itu user dapat melihat nilai dari perhitungannya *naïve bayes* seperti pada Gambar 4.8.

No	Nama	Pendidikan	Status Rumah	Luas Bangunan	Fasilitas MCK	Sumber Penerangan	Jumlah Makanan	Probabilitas	Layak	Tidak Layak	Data Real	Data Model	Condition
1	JUMANAH	2	3	3	2	2	2	4	0.0227948852234	0.000400000000000	layak	layak	True-False
2	HOTMAH	3	2	3	3	2	2	4	0.0003388276212	0.004600000000000	layak	layak	True-False
3	SANPYEM	3	2	3	3	2	2	4	0.0003388276212	0.004600000000000	layak	layak	True-False
4	SAODA	3	2	3	4	2	2	4	0.0220200776217	0.004600000000000	layak	layak	True-False
5	KASIATUN	4	3	3	3	2	2	4	0.00272076475837	0.000300000000000	layak	layak	True-False
6	SRI WIATI	3	4	3	3	2	2	4	0.0140248888882	0.000370714640000	layak	layak	True-False
7	SANWINE	3	2	3	1	2	2	4	0.000100000000000	0.000100000000000	layak	layak	True-False
8	FATMAH	3	4	3	4	2	2	4	0.000400000000000	0.000370714640000	layak	layak	True-False
9	FARDA	3	1	3	3	2	2	4	0.000400000000000	0.000400000000000	layak	layak	True-False
10	TIMA	3	2	3	1	2	2	4	0.000300000000000	0.000300000000000	layak	layak	True-False

Gambar 4.8 Hasil Perhitungan Naïve Bayes

Pada menu hasil lalu klik *naïve bayes* menampilkan nilai hasil perhitungan *naive bayes*. Serta dapat melihat hasil dari perhitungan *naive bayes* dengan data asli dan hasil confusion matrixnya.

### 4.3 Uji Coba

Untuk mengatasi data kosong tersebut menggunakan metode *k-nearest neighbor* dengan mencari nilai K yang paling optimal. Pada Tabel 4.1 dapat mengetahui hasil akurasi dari tiap K yang telah di coba seperti dibawah ini:

Tabel 4.1 Nilai K dari K-Nearest Neighbor

Nilai K	Akurasi
K = 3	True: 184 data False: 56 data <b>Akurasi 76,6%</b>
K = 5	True: 168 data False: 72 data <b>Akurasi 70%</b>
K = 7	True: 193 data False: 47 data <b>Akurasi 80,4%</b>
K = 9	True: 190 data False: 50 data <b>Akurasi 79,1%</b>

Sumber: Martha Shantika dkk, 2018

Dari percobaan diatas diketahui hasil masing-masing K=3, K=5, K=7, K=9, Terbukti bahwa pada K=7 mendapatkan nilai True: 193 data dan nilai False: 47 data sehingga akurasi tertinggi yaitu 80,4%, maka K=7 dapat digunakan untuk mengatasi *missing value* pada proses perhitungan menggunakan metode *k-nearest neighbor*.

Tahap ini adalah untuk mengetahui tingkat akurasi dari metode-metode yang digunakan dalam proses perhitungan sebuah data. Maka pengujian ini dibutuhkan beberapa cara salah satunya menggunakan *k-fold cross validation*.

#### K-Fold Validation

Dalam melakukan pengujian dengan menggunakan *k-fold validation*, pembagian data

training dan data testing dilakukan dengan mengambil k sama dengan 2, 4, 5 dan 10. Jumlah data yang digunakan adalah sebanyak 500 data dan 100 data untuk validasi. Berikut hasil dari tiap folds yang telah dilakukan:

#### 2-Fold

Pada tabel 4.2 adalah proses validasi menggunakan 250 data testing dan 250 data training

Uji	Naive Bayes	K-Nearest Neighbor + Naive Bayes	Selisih
I	TP = 143, TN = 65 FP = 21, FN = 21 <b>Akurasi 83,2%</b>	TP = 139, TN = 66 FP = 15, FN = 30 <b>Akurasi 80,2%</b>	3%
II	TP = 131, TN = 86 FP = 12, FN = 21 <b>Akurasi 86,8%</b>	TP = 150, TN = 64 FP = 14, FN = 22 <b>Akurasi 85,6%</b>	1,2%

#### 4-Fold

Pada tabel 4.3 adalah proses validasi menggunakan 125 data testing dan 125 data training

Uji	Naive Bayes	K-Nearest Neighbor + Naive Bayes	Selisih
I	TP = 60, TN = 43 FP = 13, FN = 7 <b>Akurasi 82,4%</b>	TP = 54, TN = 48 FP = 10, FN = 13 <b>Akurasi 81,6%</b>	0,8%
II	TP = 77, TN = 30 FP = 4, FN = 14 <b>Akurasi 85,6%</b>	TP = 83, TN = 29 FP = 5, FN = 8 <b>Akurasi 89,6%</b>	4%
III	TP = 57, TN = 52 FP = 7, FN = 9 <b>Akurasi 87,2%</b>	TP = 50, TN = 55 FP = 6, FN = 16 <b>Akurasi 84%</b>	3,2%



<b>IV</b>	TP = 78, TN = 30 FP = 7, FN = 10 <b>Akurasi 86,4%</b>	TP = 75, TN = 28 FP = 4, FN = 12 <b>Akurasi 84,8%</b>	<b>1,6%</b>
-----------	---	---	-------------

**5-Fold**

Pada tabel 4.4 adalah proses validasi menggunakan 100 data testing dan 100 data training

Uji	<i>Naive Bayes</i>	<i>K-Nearest Neighbour + Naive Bayes</i>	Selisih
<b>I</b>	TP = 60, TN = 28 FP = 6, FN = 6 <b>Akurasi 88%</b>	TP = 50, TN = 29 FP = 5, FN = 16 <b>Akurasi 79%</b>	<b>1%</b>
<b>II</b>	TP = 64, TN = 24 FP = 2, FN = 10 <b>Akurasi 88%</b>	TP = 55, TN = 30 FP = 8, FN = 7 <b>Akurasi 85%</b>	<b>3%</b>
<b>III</b>	TP = 48, TN = 35 FP = 10, FN = 7 <b>Akurasi 83%</b>	TP = 40, TN = 41 FP = 9, FN = 10 <b>Akurasi 81%</b>	<b>2%</b>
<b>IV</b>	TP = 46, TN = 42 FP = 5, FN = 7 <b>Akurasi 88%</b>	TP = 47, TN = 38 FP = 7, FN = 8 <b>Akurasi 85%</b>	<b>3%</b>
<b>V</b>	TP = 55, TN = 30 FP = 7, FN = 8 <b>Akurasi 85%</b>	TP = 58, TN = 25 FP = 4, FN = 5 <b>Akurasi 83%</b>	<b>2%</b>

**10-Fold**

Pada tabel 4.5 adalah proses validasi menggunakan 50 data testing dan 50 data training

Uji	<i>Naive Bayes</i>	<i>K-Nearest Neighbour + Naive Bayes</i>	Selisih
<b>I</b>	TP = 33, TN = 8 FP = 4, FN = 5 <b>Akurasi 82%</b>	TP = 33, TN = 6 FP = 6, FN = 5 <b>Akurasi 78%</b>	<b>4%</b>
<b>II</b>	TP = 26, TN = 17	TP = 24, TN = 16	<b>6%</b>

	FP = 5, FN = 2 <b>Akurasi 86%</b>	FP = 6, FN = 4 <b>Akurasi 80%</b>	
<b>III</b>	TP = 29, TN = 16 FP = 2, FN = 3 <b>Akurasi 90%</b>	TP = 26, TN = 16 FP = 2, FN = 6 <b>Akurasi 84%</b>	<b>6%</b>
<b>IV</b>	TP = 39, TN = 6 FP = 2, FN = 3 <b>Akurasi 90%</b>	TP = 30, TN = 12 FP = 1, FN = 7 <b>Akurasi 84%</b>	<b>6%</b>
<b>V</b>	TP = 22, TN = 22 FP = 4, FN = 2 <b>Akurasi 88%</b>	TP = 20, TN = 23 FP = 3, FN = 4 <b>Akurasi 86%</b>	<b>2%</b>
<b>VI</b>	TP = 28, TN = 13 FP = 6, FN = 3 <b>Akurasi 82%</b>	TP = 30, TN = 10 FP = 9, FN = 1 <b>Akurasi 80%</b>	<b>2%</b>
<b>VII</b>	TP = 22, TN = 20 FP = 5, FN = 3 <b>Akurasi 84%</b>	TP = 23, TN = 22 FP = 3, FN = 2 <b>Akurasi 90%</b>	<b>6%</b>
<b>VII I</b>	TP = 25, TN = 20 FP = 1, FN = 4 <b>Akurasi 90%</b>	TP = 26, TN = 15 FP = 7, FN = 2 <b>Akurasi 82%</b>	<b>8%</b>
<b>IX</b>	TP = 25, TN = 15 FP = 1, FN = 4 <b>Akurasi 80%</b>	TP = 29, TN = 14 FP = 2, FN = 5 <b>Akurasi 86%</b>	<b>6%</b>

<b>X</b>	TP = 31, TN = 15 FP = 1, FN = 3 <b>Akurasi 92%</b>	TP = 31, TN = 13 FP = 3, FN = 3 <b>Akurasi 88%</b>	<b>4%</b>
----------	--	--	-----------

Dalam melakukan pengujian diatas menggunakan k-fold 2, 4, 5, dan 10-fold mendapatkan hasil tertinggi dari proses *naïve bayes* yaitu di 10-fold yang terjadi diskenario ke-10 dengan nilai True Positif = 31, True Negatif = 15, False Positif = 1, False Negatif = 3 sehingga memperoleh nilai akurasi 92%. Pada proses perhitungan *k-nearest neighbor* mendapatkan hasil nilai tertinggi yaitu di 10-fold yang terjadi diskenario ke- 7 True Positif = 23, True Negatif = 22, False Postif = 3, False Negatif = 2 sehingga nilai akurasi dari proses tersebut sebesar 90%. Maka perhitungan diatas dapat mengetahui selisih dari perhitungan *naïve bayes* setelah memanfaatkan *k-nearest neighbour* rata-rata 3,5%.

#### Hasil Validation

Setelah mendapatkan hasil dari masing-masing *k-fold*, maka hasil tersebut divalidasi dengan data yang ditetapkan sebagai data validasi yaitu sebanyak 100data. Berikut Tabel 4.5 hasil proses *Validation*:

Tabel 4.6 Hasil Proses Validation

Uji	<i>Naïve Bayes</i>	<i>K-Nearest Neighbour + Naïve Bayes</i>
<b>I</b>	TP = 43, TN = 44 FP = 9, FN = 4 <b>Akurasi 87%</b>	TP = 43, TN = 42 FP = 6, FN = 9 <b>Akurasi 85%</b>

Maka akurasi yang didapat dari perhitungan validasi metode *naïve bayes* memperoleh nilai akurasi sebesar 87% dan proses perhitungan metode *k-nearest neighbor + naïve bayes* mendapatkan akurasi sebesar 85%. Dari hasil tersebut terdapat selisih sebesar 2% .

## DAFTAR PUSTAKA

- Badan Pusat Statistik.2016. *Berita Resmi Statistik*. Jember: BPS Jember
- Farhangfar Alireza, Dkk. 2006. *Impact of imputation of missing values on classification error for discrete data*. University of Alberta. Jurnal
- F, Arfiana. 2014. *Klasifikasi Kendaraan Roda Empat Menggunakan Metode Naïve Bayes*. Bandung: Universitas Widyatama.
- Haditsah, A. 2018. *Klasifikasi Masyarakat Miskin Menggunakan Naïve Bayes*. Universitas Ichsan Gorontalo. Jurnal.
- North, Matthew A. 2012. *Data Mining for the Masses*.USA: Global Text Project
- Martha Shantika, Susanti, Sulistianingsih Evy.2018. *K Nearest Neighbor Dalam Imputasi Missing Data*. Buletin Ilmiah Math. Stat. dan Terapannya (Bimaster).
- Patil dan Shrekar, 2013.*Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification*.Sant Gadgebaba Amravati University, Amravati.Jurnal.
- Prasetyo E, 2012. *Data Mining: Konsep dan Aplikasi menggunakan Matlab*, 1 ed. Yogyakarta: Andi Offset.
- Saputra, A., 2013.*Membangun Aplikasi bioskop dan sms untuk panduan skripsi*. Penerbit : PT Elex Media Komputindo , Jakarta.
- Setiawan, Bambang, Widjaja, Raden Sjarief and Nugroho, Setyo, 2009. *Perancangan Sistem Pendukung Keputusan (SPK) untuk Menentukan Kelaiklautan Kapal*.Surabaya: Program Studi MMT-ITS, 2009. Prosiding Seminar Nasional Manajemen Teknologi X. pp. C-14-1.
- Supriatna. 2017. *Pengertian Penduduk Kurang Mampu.Analisis Pengaruh Jumlah Penduduk, Tingkat*

*Pendidikan, Dan Kesehatan Terhadap Tingkat Kemiskinan Di Kabupaten Sumbawa Tahun 2010-2015.* Universitas Samawa. Sumbawa.Jurnal.

Witten, Ian H., dan Eibe Frank. 2005.*Data Mining.* Amsterdam: Morgan Kaufman.

