UNP JOURNAL OF STATISTICS AND DATA SCIENCE
Volume 2 Nomor 2 Tahun 2024 Halaman 187-195.

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

Classification of Program Keluarga Harapan Recipient Households in Padang City Using K-Nearest Neighbors

Yurivo Rianda Saputra, Syafriandi*, Dony Permana, Zilrahmi
Departemen Statistika, Universitas Negeri Padang, Padang, Indonesia
*Corresponding author: syafriandi_math@fmipa.unp.ac.id

Submitted : 17 Mei 2024
Revised : 30 Mei 2024
Accepted : 31 Mei 2024

ABSTRACT

Program Keluarga Harapan (PKH) is a social assistance program from the government aimed at providing social protection in the central government's efforts to promote social welfareas. PKH provides benefits to poor families, especially pregnant women and children, by utilizing various health and education services available. PKH benefits also include people with disabilities and the elderly by maintaining their level of social welfare in accordance with the Constitution and the Nawacita of the Republic of Indonesia. The implementation of PKH that experiences distribution errors needs to be classified to ensure its proper distribution. Classification is performed by comparing the number of neighbors (k) in K-Nearest Neighbors (KNN). The Synthetic Minority Oversampling Technique Edited Nearest Neighbors (SMOTEENN) is applied to balance classes in the target classification and Recursive Feature Elimination Cross Validation (RFECV) is applied to select attributes in the dataset used. The data source was obtained from SUSENAS 2023 data in Padang City. The research results show that KNN with k = 3 is a good algorithm for classifying households recieiving PKH using 10 attributes. KNN with k = 3 achieves an Accuracy of 91,12%, Precision of 89,29%, and Recall of 96,77%.

Keywords: *Program Keluarga Harapan, K-Nearest Neighbors, Recursive Feature Elimination Cross Validation.*

I. PENDAHULUAN

Badan Pusat Statistik (BPS) memproyeksikan jumlah penduduk Indonesia tahun 2023 sebanyak 278,8 juta jiwa. Penduduk Indonesia sangat beragam dengan berbagai latar belakang budaya, etnis, sosial, dan ekonomi. Beragam permasalahan juga dihadapi oleh pemerintah dan masyarakat Indonesia, terutama permasalahan dalam aspek ekonomi seperti kemiskinan. BPS mencatat pada tahun 2023, persentase penduduk miskin di Indonesia sebesar 9,36% yang mengalami penurunan sebesar 0,18% dibandingkan tahun 2022. Sedangkan di Provinsi Sumatera Barat persentase penduduk miskin pada tahun 2023 sebesar 5,95% yang mengalami kenaikan sebesar 0,03% dibandingkan tahun 2022. Sedangkan di Kota Padang pada tahun 2023, persentase penduduk miskin sebesar 4,17% yang mengalami penurunan sebesar 0,09% dibandingkan tahun 2022. Kota Padang merupakan kota dengan jumlah penduduk miskin terbanyak di Provinsi Sumatera Barat tahun 2023 dengan jumlah 41,97 ribu jiwa.

Angka kemiskinan yang mengalami naik turun, akan mengakibatkan kesulitan untuk mendapatkan kesetaraan dan kesejahteraan kehidupan masyarakat. Berdasarkan Undang-undang Republik Indonesia Nomor 11 tahun 2009, pemerintah berupaya untuk melakukan penyelenggaraan kesejahteraan sosial untuk memenuhi kebutuhan dasar setiap warga negara, yang meliputi rehabilitasi sosial, jaminan sosial, pemberdayaan sosial, dan perlindungan sosial. Perlindungan sosial merupakan usaha pemerintah yang bertujuan untuk memberikan dukungan kepada masyarakat agar mampu mengatasi berbagai kerentanan atau guncangan yang mungkin terjadi sepanjang perjalanan kehidupan. Salah satu bentuk perlindungan sosial adalah program pemberian bantuan yang bersifat tanpa iuran yang bersumber dari APBN dan/atau APBD dengan sasaran masyarakat miskin dan rentan. Salah satu program bantuan sosial yang diberikan oleh pemerintah adalah Program Keluarga Harapan (PKH). Tahun 2021, Badan Pemeriksa Keuangan (BPK) menemukan kesalahan dalam penyaluran bantuan sosial dalam program perlindungan sosial pemerintah yang menyebabkan kerugian negara hingga Rp 6,93 triliun (Sembiring, 2022). Untuk mengantisipasi agar PKH tersalurkan kepada yang benar membutuhkan, maka diperlukan suatu sistem untuk mengambil keputusan agar penyaluran PKH tepat sasaran.

DOI: https://doi.org/10.24036/ujsds/vol2-iss2/167

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

Sistem yang dapat dimanfaatkan agar tidak terjadi kesalahan penyaluran PKH adalah dengan memanfaatkan algoritma *machine learning* dalam klasifikasi. Menurut Nelli (2015), *machine learning* adalah disiplin yang berhubungan dengan studi tentang metode untuk pengenalan pola dalam kumpulan data untuk analisis data. Secara khusus, ini berkaitan dengan pengembangan algoritma yang belajar dari data dan membuat prediksi. Berdasarkan teknik pembelajarannya, *machine learning* dapat dibedakan menjadi *supervised learning* dan *unsupervised learning*. Menurut Nelli (2015), *supervised learning* adalah metode dimana dataset pelatihan berisi atribut yang digunakan untuk memprediksi suatu target. *Supervised learning* merupakan pembelajaran yang menggunakan dataset yang telah memiliki label sebagai keluaran dari proses pembelajaran. *Supervised learning* terbagi dua yaitu, regresi yang menghasilkan keluaran berupa nilai kontinu dan klasifikasi yang menghasilkan keluaran berupa klasifikasi pada dua atau lebih kategori. Algoritma klasifikasi yang dapat digunakan adalah Regresi Logistik, *Decision Tree*, *Naïve Bayes Cassifier* (NBC), *Support Vector Machine* (SVM), *Random Forest*, dan *K-Nearest Neighbors* (KNN). *Unsupervised learning* adalah metode dimana dataset pelatihan terdiri dari serangkaian nilai atribut tanpa adanya target. *Unsupervised learning* dapat digunakan untuk pengelompokan data (*clustering*) dan reduksi dimensi.

Pemilihan algoritma klasifikasi yang tepat dapat meningkatkan keakuratan keputusan yang diambil. Menurut Bhatia, (2010), KNN merupakan algoritma yang sederhana dan mudah dipelajari serta efektif jika data pelatihan berukuran besar. Menurut Tang dkk., (2016), KNN dinilai sebagai algoritma yang sederhana dalam pengolahan data *training* dan data *testing* dalam data yang berukuran besar. Menurut Tharwat dkk., (2018), KNN merupakan algoritma klasifikasi yang berbasis contoh atau non parametrik dan dianggap teknik paling sederhana dalam data mining. KNN merupakan teknik pengklasifikasian data yang memiliki konsistensi yang kuat, dengan cara menghitung kedekatan antara data baru dengan data lama yang disebut dengan k. Menurut Jayadi dkk., (2023), nilai *k* yang terbaik tergantung pada data, secara umum nilai *k* yang tinggi akan mengurangi efek noise. Pada data yang berukuran besar sebaiknya memilih nilai *k* yang kecil. Nilai *k* sangat menentukan hasil klasifikasi, sehingga diperlukan kehati-hatian dalam menetepkannya.

Penelitian yang dilakukan oleh Indrayanti dkk., (2017), KNN diterapkan pada dataset penyakit diabetes di India dalam mengklasifikasikan penyakit diabetes mellitus. Penelitian tersebut menghitung nilai *k* paling optimum pada KNN untuk klasifikasi penyakit diabetes mellitus. Penelitian lain juga dilakukan oleh Karomi (2015), KNN diterapkan pada dataset penerimaan mahasiswa baru di STMIK Widya Pratama Pekalongan dalam klasifikasi herregistrasi mahasiswa, serta dilakukan optimalisasi nilai *k*. Penelitian yang mengimplementasikan algoritma KNN dilakukan oleh Sumarlin (2015), klasifikasi dengan KNN digunakan untuk pendukung keputusan penerima beasiswa Peningkatan Prestasi Akademik (PPA) dan Bantuan Belajar Mahasiswa (BBM). Berdasarkan penelitian terdahulu, penelitian ini melakukan klasifikasi penerima PKH di Kota Padang menggunakan algoritma KNN.

Kesalahan penyaluran PKH yang tidak tepat sasaran, menjadi permasalahan yang perlu diatasi oleh pemerintah, agar sesuai dengan tujuan yang telah dirumuskan dalam konstitusi. Pemanfaatan algoritma klasifikasi KNN diharapkan dapat memberikan saran dan membantu pemerintah dalam mengatasi permasalahan tersebut. Tujuan dari penelitian ini adalah membantu pemerintah dalam menetapkan penerima PKH di Kota Padang dengan menggunakan algoritma KNN, agar penerima PKH tepat sasaran.

II. METODE PENELITIAN

Jenis penelitian ini merupakan penelitian terapan. Penerapan algoritma KNN terhadap klasifikasi rumah tangga penerima PKH di Kota Padang, dengan berbagai k pada data yang diperoleh dari data SUSENAS tahun 2023. Atribut yang digunakan terdiri atas 16 atribut dan 1 target, yang disajikan pada Tabel 1.

Tabel 1. Atribut Penelitian

No	Atribut	Skala	Kategori
(1)	(2)	(3)	(4)
1	Status kepemilikan rumah (X_1)	Nominal	Milik sendiri; Kontrak/Sewa; Bebas sewa; Dinas; dan Lainnya
2	Banyak ART (X_2)	Rasio	-
3	Luas lantai rumah (X_3)	Rasio	-
4	Kepemilikan rumah lain (X_4)	Nominal	Ya dan Tidak
5	Bahan atap rumah (X_5)	Nominal	Beton; Genteng; Seng; Asbes; Bambu; Kayu/sirap; Jerami/ijuk/daun-daunan/rumbia; dan Lainnya
6	Bahan dinding rumah (X_6)	Nominal	Tembok; Plesteran anyaman bambu/kawat; Kayu/papan; Anyaman bambu; Batang kayu; Bambu; dan Lainnya

DOI: https://doi.org/10.24036/ujsds/vol2-iss2/167

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

7	Bahan lantai rumah (X_7)	Nominal	Marmer/granit; Keramik; Parket/vinil/karpet; Ubin/tegel/teraso; Kayu/papan; Semen/bata merah; Bambu; Tanah; dan Lainnya
(1)	(2)	(3)	(4)
8	Daya listrik (X_8)	Nominal	450 watt; 900 watt; dan 1.300 watt atau lebih
9	Sumber utama air minum (X_9)	Nominal	Air kemasan bermerk; Air isi ulang; Leding; Sumur bor/pompa; Sumur terlindung; Sumur tak terlindung; Mata air terlindung; Mata air tak terlindung; Air permukaan (sungai, danau/waduk, kolam, irigasi) ; Air hujan; dan Lainnya
10	Bahan bakar masak (X_{10})	Nominal	Tidak memasak di rumah; Listrik; Elpiji 5,5 kg/bluegaz; Elpiji 12 kg; Elpiji 3 kg; Gas kota; Biogas; Minyak tanah; Briket; Arang; Kayu bakar; dan Lainnya
11	Kepemilikan kulkas (X_{11})	Nominal	Ya dan Tidak
12	Kepemilikan AC (X_{12})	Nominal	Ya dan Tidak
13	Kepemilikan telepon (X_{13})	Nominal	Ya dan Tidak
14	Kepemilikan motor (X_{14})	Nominal	Ya dan Tidak
15	Kepemilikan mobil (X_{15})	Nominal	Ya dan Tidak
16	Kepemilikan TV (X_{16})	Nominal	Ya dan Tidak
17	Menerima PKH (Y)	Nominal	Ya dan Tidak

A. Tahapan Analisis

Penelitian ini melakukan perbandingan nilai k dalam algoritma *K-Nearest Neighbors* (KNN) untuk mengklasifikasikan rumah tangga penerima PKH di Kota Padang. Penelitian menerapkan metode *Synthetic Minority Oversampling Technique Edited Nearest Neighbors* (SMOTEENN) dalam menyeimbangkan kelas pada target dan *Recursive Feature Elimination Cross Validation* (RFECV) dalam menyeleksi atribut pada dataset. Tahapan analisis yang digunakan adalah sebagai berikut.

1. *Preprocessing* data dengan melakukan pelabelan data dan pemeriksaan data yang kosong.
2. Melakukan visualisasi data pada atribut target untuk melihat proporsi data.
3. Mengatasi ketidakseimbangan data dengan metode SMOTEENN.

   Menurut Gu dkk., (2016), Klasifikasi yang dilakukan pada data yang tidak seimbang, algoritma klasifikasi akan menghasilkan akurasi yang lebih tinggi untuk kelas mayoritas daripada kelas minoritas. Menurut Batista dkk., (2004), SMOTEENN merupakan metode *hybrid sampling* gabungan antara metode *Synthetic Minority Oversampling Technique* (SMOTE) untuk membangkitkan data kelas minoritas dan *Edited Nearest Neighbors* (ENN) untuk menghapus data kelas mayoritas. Tahapan dari SMOTEENN sebagai berikut.

   a. Mengidentifikasi kelas yang menjadi kelas minoritas dan kelas mayoritas dari dataset yang digunakan.
   b. Menghitung jarak antar data untuk menentukan tetangga terdekat dari data tersebut. Menurut Cost & Salzberg (1993), Perhitungan jarak pada data dapat menggunakan *euclidean distance* untuk data numerik dan *Value Difference Metric* (VDM) untuk data kategorik. Menurut Permana dkk., (2023), rumus *euclidean distance* dapat dilihat pada Persamaan (1) dan Menurut Cost & Salzberg (1993), rumus VDM dapat dilihat pada Persamaan (2).

$$d(x_i, y_i) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{1}$$

$$d(x_i, y_i) = \sum_{j=1}^{k} \left| \frac{C_{xi}}{C_x} - \frac{C_{yi}}{C_y} \right|^d \tag{2}$$

   Keterangan :

   $d(x_i, y_i)$: jarak antara amatan x dan y pada peubah ke-i
   x_i : data amatan x pada peubah ke-i
   y_i : data amatan y pada peubah ke-i
   n : jumlah peubah atau amatan
   $d(x_i, y_i)$: jarak antara kategori x dan y pada atribut ke-i
   k : banyaknya kelas pada peubah respon
   C_{xi} : banyaknya kategori x pada kelas ke-i
   C_{yi} : banyaknya kategori y pada kelas ke-i

   c. Membangkitkan data buatan pada kelas minoritas. Menurut Cost & Salzberg (1993), pembangkitan data sintetis dapat menggunakan rumus pada Persamaan (3).

$$X_{new} = X_i + (X_k - X_i)\delta \tag{3}$$

DOI: https://doi.org/10.24036/ujsds/vol2-iss2/167

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

Keterangan :
X_{new} : data sintetis baru
X_i : data dari kelas minoritas
X_k : data dari k tetangga terdekat yang memiliki jarak terdekat dengan X_i
δ : bilangan acak antara 0 dan 1

d. Menghapus data yang bersifat *noise* pada kelas mayoritas.
e. Menggabungkan data hasil SMOTEENN.

4. Melakukan pembagian data penelitian menjadi 75% data *training* dan 25% data *testing*.
5. Melakukan seleksi atribut pada data dengan metode RFECV.

Menurut Guyon dkk., (2002), *Recursive Feature Elimination* (RFE) merupakan bentuk penerapan dari *backward feature elimination*. RFE bekerja dengan cara mengurangi atribut yang tidak mempengaruhi akurasi model sampai jumlah Atribut yang diinginkan. Dalam beberapa kasus jumlah atribut yang mempengaruhi akurasi model belum diketahui. Penggabungan RFE dengan *Cross Validation* (CV) dapat digunakan untuk memilih jumlah atribut yang maksimal dalam mempengaruhi model tanpa harus menentukan jumlah atribut terlebih dahulu. Menurut Guyon dkk., (2002), RFECV memerlukan nilai koefisien dari masing-masing atribut untuk menentukan tingkat kepentingan atribut Menurut Zhang dkk., (2013), perangkingan atribut dilakukan dengan nilai koefisien atribut yang diperoleh dengan menggunakan algoritma *Support Vector Machine* (SVM) Linear. Tahapan dari SVM-RFECV sebagai berikut.

a. Membagi data menjadi *k-fold*. Menurut Bramer (2007), nilai *k* yang biasa digunakan antara 5 atau 10, dimana 1 subset akan dijadikan sebagai data *testing* dan k-1 subset sebagai data *training*. Masing-masing subset akan menjadi data *testing* dalam proses pemodelan.
b. Melatih data *training* dengan algoritma *Support Vector Machine* (SVM) Linear.
   1) Membentuk *hyperplane* terbaik dengan menghitung *margin* terbesar dengan Persamaan (4).

$$\sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j \vec{x}_i . \vec{x}_j \tag{4}$$

   2) Penghitungan margin terbesar menghasilkan nilai koefisien lagrange (α_i).
   3) Nilai $\vec{w}$ diperoleh dengan menggunakan nilai α_i dengan Persamaan (5).

$$\vec{w} = \sum_{i=1}^{l} \alpha_i y_i x_i \tag{5}$$

   4) Nilai b diperoleh dengan menggunakan nilai $\vec{w}$ dengan Persamaan (6).

$$b = \frac{1}{2} \vec{w} [x_r + x_s] \tag{6}$$

   5) Melakukan klasifikasi pada data *testing* dengan Persamaan (7).

$$f(x_{new}) = \sum_{i=1}^{ns} \alpha_i y_i x_i x_{new} + b \tag{7}$$

c. Mengevaluasi kinerja SVM Linear dengan nilai F1-*score*.
d. Melakukan iterasi sebanyak k dan menghitung rata-rata F1-*score.*
e. Mengurutkan atribut berdasarkan nilai koefisiennya dan menghapus atribut dengan koefisien terkecil.
f. Melakukan langkah b sampai e sampai 1 atribut tersisa.
g. Diperoleh rata-rata F1-*score* untuk masing-masing jumlah atribut. Jumlah atribut dengan rata-rata F1-*score* terbesar menjadi jumlah atribut yang digunakan selanjutnya.

6. Melakukan klasifikasi menggunakan algoritma KNN. Tahapan dari KNN sebagai berikut.
   a. Menentukan nilai ketetanggaan (k).

      Tidak ada aturan pasti dalam menentukan nilai *k*, namun pada data yang terdapat 2 kelas klasifikasi, nilai *k* yang digunakan adalah ganjil agar tidak terjadi keambiguan dalam klasifikasi.

   b. Menghitung jarak antara data *training* dan *testing* dengan Persamaan (1).
   c. Mengurutkan jarak yang terbentuk dari yang terkecil sampai jarak terbesar.
   d. Menentukan jarak terdekat sesuai dengan ukuran k.
   e. Menentukan jumlah masing-masing kelas yang termasuk dalam jarak terdekat.
   f. Menetapkan klasifikasi data *training* sesuai dengan jumlah kelas terbanyak.
7. Melakukan evaluasi klasifikasi.

DOI: https://doi.org/10.24036/ujsds/vol2-iss2/167

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

Evaluasi klasifikasi dapat dihitung dengan menggunakan *confusion matrix*. Menurut Han dkk., (2012), *confusion matrix* dapat diartikan sebagai suatu alat yang berfungsi untuk melakukan analisis apakah algoritma *machine learning* dapat mengklasifikasikan dengan baik dalam mengenali *tuple* dari kelas yang berbeda. Tabel *confusion matrix* terdapat pada Tabel 2. Evaluasi klasifikasi dengan memperhatikan nilai *accuracy* yang diperoleh dengan Persamaan (8).

Tabel 2. *Confusion Matrix*

Classification		**Prediksi**	
		Positif	**Negatif**
Observasi	**Positif**	True Positive (TP)	False Negative (FN)
	Negatif	False Positive (FP)	True Negative (TN)

Keterangan :
TP adalah banyaknya data yang sebenarnya positif dan diprediksi sebagai data positif.
TN adalah banyaknya data yang sebenarnya negatif dan diprediksi sebagai data negatif.
FP adalah banyaknya data yang sebenarnya negatif dan diprediksi sebagai data positif.
FN adalah banyaknya data yang sebenarnya positif dan diprediksi sebagai data negatif.
Informasi yang diperoleh sebagai berikut.

a. *Accuracy* digunakan untuk mengukur sejauh mana model klasifikasi berhasil dalam memprediksi dengan benar seluruh kelas target. *Accuracy* dapat dihitung dengan rumus pada Persamaan (8).

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \tag{8}$$

b. *Precision* digunakan untuk mengukur persentase kemampuan model dalam memprediksi kelas positif dengan benar dari semua prediksi positif. *Precision* dapat dihitung dengan rumus pada Persamaan (9).

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{9}$$

c. *Recall* digunakan untuk mengukur persentase kemampuan model dalam memprediksi kelas positif. *Recall* dapat dihitung dengan rumus pada Persamaan (10).

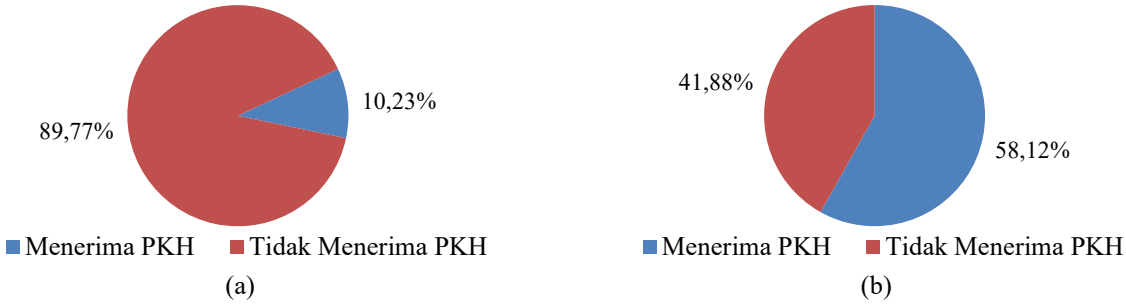
$$Recall = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{10}$$

8. Membandingkan nilai *Accuracy*, *Precision*, dan *Recall* dari klasifikasi dengan KNN.

III. HASIL DAN PEMBAHASAN

A. *Synthetic Minority Oversampling Technique Edited Nearest Neighbors*

Visualisasi data pada atribut target dilakukan untuk melihat proporsi data sebelum dilakukan proses penyeimbangan dengan metode SMOTEENN. Visualisasi atribut target sebelum dan sesudah dilakukan penyeimbangan terdapat pada Gambar 1.



Gambar 1. (a) Atribut Target Sebelum Diseimbangkan dan (b) Atribut Target Setelah Diseimbangkan

Antar kelas pada atribut target memiliki perbandingan yang jauh. Dapat dikatakan data awal yang digunakan mengalami ketidakseimbangan dan perlu dilakukan penanganan. Rumah tangga yang tidak menerima PKH menjadi kelas mayoritas dengan persentase sebesar 89,77% atau 702 objek amatan. Rumah tangga yang menerima PKH menjadi kelas minoritas, dengan persentase sebesar 10,23% atau 80 objek amatan. Setelah metode SMOTEEN diterapkan pada data awal diperoleh data yang cukup seimbang dengan perbandingan kelas yang tidak terlalu jauh. Sebesar 41,88% merupakan data rumah tangga yang tidak menerima PKH atau sebanyak 433 objek amatan dan sebesar 58,12%

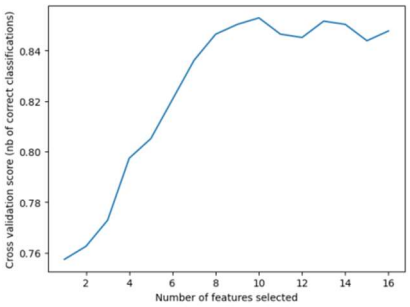
DOI: https://doi.org/10.24036/ujsds/vol2-iss2/167

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

merupakan data rumah tangga yang menerima PKH atau sebanyak 601 objek amatan. Setelah penyeimbangan diperoleh peningkatan objek amatan dalam data, pada data awal terdapat sebanyak 782 objek amatan, setelah diseimbangkan menjadi 1034 objek amatan.

B. *Recursive Feature Elimination Cross Validation*

RFECV merupakan tahapan memilih atribut data yang banyak kemudian menjadi lebih sedikit dengan tujuan menghilangkan atribut yang tidak memiliki kontribusi atau pengaruh terhadap klasifikasi. Dengan menggunakan metode RFECV, atribut data yang berjumlah 16 atribut kemudian akan dilihat pada jumlah atribut dan atribut yang mana saja yang paling optimal untuk digunakan untuk tahap klasifikasi, hasil dari metode RFECV terdapat pada Gambar 2.



Gambar 2. Hasil Seleksi Atribut Menggunakan RFECV

Dari Gambar 2 diperoleh nilai *cross validation* paling tinggi diperoleh ketika jumlah atribut yang digunakan 10, dengan nilai c*ross validation* sebesar 85,29%. Sepuluh atribut yang optimal tersebut adalah X_1, X_3, X_5, X_8, X_{11}, X_{12}, X_{13}, X_{14}, X_{15}, dan X_{16}. Dengan atribut yang diperoleh dibentuk tabel frekuensi dari kategori dalam masing-masing atribut. Tabel frekuensi untuk atribut kategorik terdapat pada Tabel 3.

Tabel 3. Frekuensi Kategori Masing-masing Atribut

Atribut	Keterangan	Menerima	Tidak Menerima
Status kepemilikan rumah (X_1)	Milik sendiri	51	414
	Kontrak/Sewa	6	82
	Bebas sewa	23	193
	Dinas	0	13
	Lainnya	0	0
Bahan atap rumah (X_5)	Beton	0	2
	Genteng	1	26
	Seng	79	665
	Asbes	0	6
	Bambu	0	3
	Kayu/sirap	0	0
	Jerami/ijuk/daun-daunan/rumbia	0	0
	Lainnya	0	0
Daya listrik (X_8)	450 watt	19	78
	900 watt	58	411
	1.300 watt atau lebih	3	213
Kepemilikan kulkas (X_{11})	Memiliki Kulkas	60	615
	Tidak Memiliki Kulkas	20	87
Kepemilikan AC (X_{12})	Memiliki AC	1	157
	Tidak Memiliki AC	79	545
Kepemilikan telepon (X_{13})	Memiliki Telepon Rumah	0	28
	Tidak Memiliki Telepon Rumah	80	674
Kepemilikan motor (X_{14})	Memiliki Motor	71	609
	Tidak Memiliki Motor	9	93
Kepemilikan mobil (X_{15})	Memiliki Mobil	1	217
	Tidak Memiliki Mobil	79	485
Kepemilikan TV (X_{16})	Memiliki TV	19	323

DOI: https://doi.org/10.24036/ujsds/vol2-iss2/167

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

	Tidak Memiliki TV	61	379

Pada Tabel 3 diperoleh sebagian besar rumah tangga penerima PKH di Kota Padang memiliki rumah dengan status milik sendiri, beratap seng, daya listrik 900 watt, memiliki kulkas dan motor, serta tidak memiliki AC, telepon, mobil, dan TV. Selanjutnya, deskripsi data untuk atribut numerik terdapat pada Tabel 4.

Tabel 4. Karakteristik Atribut Numerik

	PKH	Rata-rata	Standar Deviasi	Nilai Minimum	Q_1	Q_2	Q_3	Nilai Maksimum
Luas lantai rumah (X_3)	Menerima	71,87	40,33	9	48	62	96	225
	Tidak Menerima	93,84	70,27	5	48	80	119,25	620

Pada Tabel 4 diperoleh rumah tangga yang penerima PKH di Kota Padang, memiliki rumah dengan rata-rata luas lantai sebesar 71,87m^2, 25% memiliki rumah dengan luas lantai ≤ 48 m^2 dan 25% memiliki rumah dengan luas lantai ≥ 96 m^2. Rumah tangga yang tidak penerima PKH di Kota Padang, memiliki rumah dengan rata-rata luas lantai sebesar 71,87m^2, 25% memiliki rumah dengan luas lantai ≤ 48 m^2 dan 25% memiliki rumah dengan luas lantai ≥ 119,25 m^2.

C. *K-Nearest Neighbors*

Klasifikasi rumah tangga penerima PKH di Kota Padang dengan menggunakan algoritma KNN dilakukan dengan berbagai nilai *k*. Nilai *k* yang digunakan merupakan angka ganjil dari 3 sampai 15, karena kelas atribut target pada data berjumlah 2. *Confusion matrix* dari algoritma KNN terdapat pada Tabel 5.

Tabel 5. Nilai *Confusion Matrix* dari KNN

Algoritma	K	*Confusion Matrix*			
		***True Positive* (TP)**	***False Positive* (FP)**	***True Negative* (TN)**	***False Negative* (FN)**
KNN	3	150	18	86	5
	5	144	19	85	11
	7	146	25	79	9
	9	145	28	76	10
	11	145	29	75	10
	13	142	25	79	13
	15	143	26	78	12

Nilai *accuracy*, *precision,* dan *recall* dari masing-masing k dalam KNN, yang terdapat pada Tabel 6.

Tabel 6. Nilai *Accuracy*, *Precision,* dan *Recall* dari KNN

K	3	5	7	9	11	13	15
Accuracy	91,12*	88,42	86,87	85,33	84,94	85,33	85,33
Precision	89,29*	88,34	85,38	83,82	83,33	85,03	84,62
Recall	96,77*	92,90	94,14	93,55	93,55	91,61	92,26

Pada Tabel 6 diperoleh algoritma KNN dengan nilai *k* = 3 memiliki nilai *accuracy*, *precision,* dan *recall* tertinggi. Algoritma KNN dengan nilai *k* = 3 memperoleh nilai *accuracy* sebesar 91,12%, artinya model mampu mengklasifikasikan rumah tangga penerima PKH dan tidak menerima PKH di Kota Padang dengan benar sebesar 91,12%. *Precision* sebesar 89,29%, artinya algoritma KNN memprediksi rumah tangga penerima PKH dengan benar dari semua prediksi rumah tangga penerima PKH sebesar 89,29%. *Recall,* sebesar 96,77%, artinya algoritma KNN mampu memprediksi rumah tangga penerima PKH sebesar 96,77%.

IV. KESIMPULAN

Berdasarkan hasil perbandingan algoritma KNN dengan berbagai k, diperoleh KNN dengan nilai *k* = 3 dengan 10 atribut menjadi algoritma terbaik dalam klasifikasi rumah tangga penerima PKH di Kota Padang. Atribut tersebut adalah status kepemilikan rumah, luas lantai rumah, bahan atap rumah, daya listrik, kepemilikan kulkas, kepemilikan AC, kepemilikan telepon, kepemilikan motor, kepemilikan mobil, dan kepemilikan TV. KNN dengan nilai *k* = 3, mampu dengan sangat baik mengklasifikasikan rumah tangga penerima PKH yakni sebesar 91,12%, dan mampu untuk memprediksi rumah tangga penerima PKH sebesar 96,77%.

DOI: https://doi.org/10.24036/ujsds/vol2-iss2/167

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

UCAPAN TERIMA KASIH

Terima kasih kepada Badan Pusat Statistik (BPS) Sumatera Barat sebagai lembaga penyedia data SUSENAS 2023 yang digunakan dalam penelitian ini.

DAFTAR PUSTAKA

Badan Pusat Statistik. (2023). *Jumlah Penduduk Pertengahan Tahun (Ribu Jiwa), 2022-2023*. https://www.bps.go.id/id/statistics-table/2/MTk3NSMy/jumlah-penduduk-pertengahan-tahun--ribu-jiwa-.html

Batista, G. E. A. P. A., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter*, *6*(1), 20–29. https://doi.org/10.1145/1007730.1007735

Bhatia, N. (2010). Survey of Nearest Neighbor Condensing Techniques. *International Journal of Computer Science and Information Security*, *8*, 302–305.

Bramer, M. (2007). Principles of Data Mining. In *Springer*. London: Springer.

Cost, S., & Salzberg, S. (1993). A weighted nearest neighbor algorithm for learning with symbolic features. *Machine Learning*, *10*(1), 57–78. https://doi.org/10.1007/bf00993481

Gu, Q., Wang, X. M., Wu, Z., Ning, B., & Xin, C. S. (2016). An improved SMOTE algorithm based on genetic algorithm for imbalanced data classification. *Journal of Digital Information Management*, *14*(2), 92–103.

Guyon, I., Weston, J., & Barnhill, S. (2002). Gene Selection for Cancer Classification Using DCA. In *Machine Learning*, 46, 389–422. https://doi.org/https://doi.org/10.1023/A:1012487302797

Han, J., Kamber, M., & Pei, J. (2012). *Data Mining Concepts and Techniques* (3 ed.). Morgan Kaufmann.

Indrayanti, Sugianti, D., & Karomi, M. A. Al. (2017). Optimasi Parameter K Pada Algoritma K-Nearest Neighbouru Untuk Klasifikasi Penyakit Diabetes Mellitus. *Prosiding SNATIF*, *4*, 823–830.

Jayadi, B. V., Handhayani, T., & Lauro, M. D. (2023). Perbandingan Knn Dan Svm Untuk Klasifikasi Kualitas Udara Di Jakarta. *Jurnal Ilmu Komputer dan Sistem Informasi*, *11*(2), 1-7. https://doi.org/10.24912/jiksi.v11i2.26006

Karomi, M. A. Al. (2015). Optimasi Parameter K Pada Algoritma KNN Untuk Klasifikasi Heregistrasi Mahasiswa. *IC-Tech*, *10*(1), 28–33.

Nelli, F. (2015). Python Data Analytics:Data Analysis and Science Using Pandas, matplotlib, and the Python Programming Language. In *Apress*. New York: Apress.

Permana, A. A., S, W., Santoso, L. W., Wibowo, G. W. N., Wardhani, A. K., Rahmaddeni, Wahidin, A. J., Yuliastuti, G. E., Elisawati, Wijayanti, R. R., & Abdurrasyid. (2023). *Machine Learning*. Global Eksekutif Teknologi.

Sembiring, L. J. (2022). *6 Kesalahan Penyaluran Bansos Ini Bikin Negara Tekor Rp 6,9 T*. https://www.cnbcindonesia.com/news/20220525103456-4-341744/6-kesalahan-penyaluran-bansos-ini-bikin-negara-tekor-rp-69-t

Sumarlin. (2015). Implementasi Algoritma K-Nearest Neighbor Sebagai Pendukung Keputusan Klasifikasi Penerima Beasiswa PPA dan BBM. *Jurnal Sistem Informasi Bisnis*, 5, 52–62.

Tang, Y., Jing, L., Li, H., & Atkinson, P. M. (2016). A multiple-point spatially weighted k-NN method for object-based classification. *International Journal of Applied Earth Observation and Geoinformation*, *52*, 263–274. https://doi.org/10.1016/j.jag.2016.06.017

Tharwat, A., Mahdi, H., Elhoseny, M., & Hassanien, A. E. (2018). Recognizing human activity in mobile crowdsensing environment using optimized k -NN algorithm. *Expert Systems with Applications*, *107*, 32–44. https://doi.org/10.1016/j.eswa.2018.04.017

Zhang, F., Kaufman, H. L., Deng, Y., & Drabier, R. (2013). Recursive SVM Biomarker Selection for Early Detection of Breast Cancer in Peripheral Blood. *BMC Medical Genomics*, *6*(SUPPL.1), S4. https://doi.org/10.1186/1755-8794-6-S1-S4

DOI: https://doi.org/10.24036/ujsds/vol2-iss2/167

ISSN(*Print*) : 3025-5511
ISSN(*Online*): 2985-475X

DOI: https://doi.org/10.24036/ujsds/vol2-iss2/167