PROSIDING Seminar Nasional SISFOTEK ISSN 2597-3584 (Media Online)
Sistem Informasi dan Teknologi
website : www.seminar.iaii.or.id | email : seminar@iaii.or.id

Klasifikasi Metode Naïve Bayes dan K-NearestNeighbor untuk Menentukan Keluarga Tidak Mampu

Riza Marsuciati[1], Gagah Gumelar[2] , Rudy Prietno[3]
[1,2,3]Magister Teknik Informatika, Fakultas Ilmu Komputer, Universitas AMIKOM Yogyakarta
[1]riza.marsuciati@students.amikom.ac.id, [2]gagah.gumelar@students.amikom.ac.id, [3]rudy.prietno@students.amikom.ac.id

Abstract

The problem of poverty has a criticalrole in social life, especially for the government associated with all forms of programs to eradicate poverty. The classification of low-income families also serves as a point to prioritize all forms of assistance in government programs. In these problems, it is quite apparent that the distribution ofaid is not well-targeted. In this study, we are lookingfor for the classification method with the best performance in classifying low-income families. Thisstudy limits the classification method to the Naïve Bayes classification method and the k-NearestNeighbor classification method. The dataset used is more than 800 families spread overtwo labels with 12 parameters, where 30 percent of family data is used as training data, and 70 percent of family data becomes as test data. The test results show that the average accuracy of the Naïve Bayes calcification method is 82.68%, while the K- Nearest Neighbor classification method is 85.57%. This studyconcludes that the best method for classifying low- income families is the Naïve Bayes method.

Keywords: Comparison of Classifications, NaïveBayes, K- Nearest Neighbor

Abstrak

Masalah kemiskinan memiliki peran penting dalam kehidupan sosial, terutama bagi pemerintah terkait dengan segala bentuk program pengentasan kemiskinan. Pengelompokan keluarga berpenghasilan rendah juga menjadi acuan untuk memprioritaskan segala bentuk bantuan dalam program pemerintah. Dalam permasalahan tersebut, terlihat jelas bahwa penyaluran bantuan tidak tepat sasaran. Dalam penelitian ini, kami mencari metode klasifikasi dengan kinerja terbaik dalam mengklasifikasikan keluarga berpenghasilan rendah. Penelitian ini membatasi metode klasifikasi pada metode klasifikasi Naïve Bayes dan metode klasifikasi k-Nearest Neighbor. Dataset yang digunakan adalah lebih dari 800 keluarga yang tersebar pada dua label dengan 12 parameter, dimana 30 persen data keluarga digunakan sebagai data latih, dan 70 persen data keluarga menjadi data uji. Hasil pengujian menunjukkan bahwa rata-rata akurasi metode kalsifikasi Naïve Bayes adalah 82,68%, sedangkan metode klasifikasi K- Nearest Neighbor adalah 85,57%. Penelitian ini menyimpulkan bahwa metode terbaik untuk mengklasifikasikan keluarga berpenghasilan rendah adalah metode Naïve Bayes.

Kata kunci: *Perbandingan Klasifikasi, Naïve Bayes, K- Nearest Neighbor*

1. Pendahuluan

Kemiskinan merupakan isu kritis dan fundamental dalam pembangunan. Mengurangi jumlah kemiskinan merupakan tantangan yang signifikan bagi setiap negara [1]. Menurut Badan Pusat Statistik (BPS) pada 2019, jumlah penduduk miskin di Indonesia pada September 2019 sebanyak 24,79 juta jiwa, turun 0,36 juta jiwa pada Maret 2019, dan menurun 0,88 juta jiwapada September 2018 [2]. Identifikasi masalah dan sumber masalah dalam pembahasan ini terkait dengan masalah pengelompokan keluarga berpenghasilan rendah karena kemiskinan merupakan masalah yang sulit untuk diatasi. Oleh karena itu, klasifikasi berguna untuk menentukan apakah suatu keluarga lemah atau tidak. Selanjutnya, klasifikasi keluarga berpenghasilan rendah dapat dimanfaatkan oleh pemerintah dalam berbagai jenis bantuan, seperti Raskin, Kartu Indonesia Pintar, dan bantuan lainnya.

Paper ini akan membahas tentang klasifikasi keluarga berpenghasilan rendah. Beberapa penelitian sebelumnya membahas tentang perbandingan performansi antara beberapa metode klasifikasi yaitu metode klasifikasi fuzzy, naïve bayes, k-nearest neighbour, decision tree, dan reverse DBSCAN. Dari hasil perbandingan tersebut, Naïve Bayes merupakan salah satu algoritma klasifikasi yang berkinerja terbaik [3] [4] [5].

Pembuatan paper ini bertujuan untuk membandingkan dua metode klasifikasi yaitu metode klasifikasi naïve bayes dengan metode klasifikasi k- Nearest Neighbor.

2. Metode Penelitian

2.1 Klasifikasi Pendekatan

Klasifikasi adalah metode penambangan data yang digunakan untuk menganalisis kumpulan data tertentu dan mengelompokkannya ke dalam kelas atau label. Pemberian kelas atau label penting karena akan mempengaruhi keakuratan dataset [6].

Parameter yang digunakan dalam penelitian ini adalah status bangunan, luas lantai, tipe lantai, tipe dinding, tipe atap, jumlah kamar tidur, sumber air minum, cara memperoleh air minum, sumber penerangan primer, bahan bakar / energi untuk memasak, penggunaan Fasilitas pembuangan air berlimpah, tempat tinja berakhir dengan dua label, yaitu miskin dan tidak membutuhkan. Dalam klasifikasi terdapat dua langkah yaitu penentuan data latih, dan yang kedua adalah pengujian data berdasarkan data latih atau bisa disebut data uji [7].

Dalam penelitian ini peneliti menggunakan dua metode klasifikasi yaitu naïve bayes dan k- Nearest Neighbor, dimana penelitian ini adalah mencari metode klasifikasi dengan kinerja terbaik untuk menentukan keluarga berpenghasilan rendah.

2.1 Naïve Bayes

Naïve Bayes adalah klasifikasi probabilistiksederhana yang menghitung sekumpulan probabilitas dengan menambahkan frekuensi dan kombinasi nilai dari kumpulan data tertentu. Algoritma ini menggunakan teorema Bayes dan mengasumsikan semua atribut independen atau tidak interdependen yang diberikan oleh nilai variabel kelas [8]. Klasifikasi akan memilih klasifikasi yang paling dekat dengan Vnb dengan atribut yang diberikan $\alpha 1$,

$a2$, $a3$,…, an Rumus penghitungan Vnb adalah sebagai berikut:

$$V_{nb} = argmax_{v_j \in V} P(v_j) \prod P(a_i|v_j)$$

The estimate $P(a_1|v_j)$ uses estimation m as follows [7]:

$$P(a_1|v_j) = \frac{n_c + mp}{n + m}$$

Information:

n = the amount of training data which is $v = v_j$

n_c = number of examples for which $v = v_j$ dan $a = a_i$

p = a estimated priority for $P(a_1|v_j)$

m = an equivalent sample size

m = the equivalent sample size

2.2 k-Nearest Neighbor

k-Nearest Neighbor merupakan metode non parametrik yang digunakan dalam perancangan keluarga berpenghasilan rendah untuk menentukan kelas dari setiap data uji berdasarkan nilai k terdekat dari data latih, dan mayoritas kelas adalah sama dengankelas k tetangga ini [9]. Cara kerja algoritma k-Nearest Neighbor ditunjukkan pada gambar dibawah ini :

start

look for positive k values with new sampleschoose the k value closest to the sample

In calculating the similarity can use Euclidean distance, which is the separation between two focuses, for example, X1 and X2 where X1 = (x10, x11, x12, ..., x1n) and X2 (x20, x21, x22, ... x2n)

$$dist(X_1X_2) = \sqrt{\sum_{i=1}^{n}(X_{1i} - X_{2i})\,2}$$

Min-max normalization to change the value of v from attribute A to v 'in the range 0

$$v' = \frac{v - min_A}{max_A - min_A}$$

To get the best value of k, it is necessary to add k values starting from k = 1.

2.3 Terkait Pekerjaan

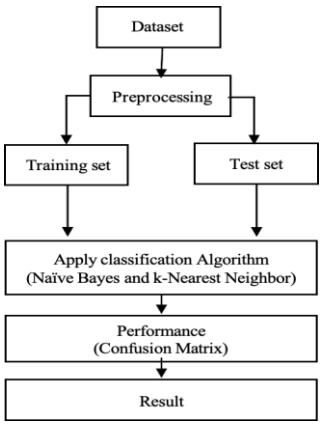
Pada penelitian sebelumnya pada tahun 2017 membahas perbandingan performansi dan optimasi dua metode klasifikasi yaitu k-NN dan Naïve Bayes, menggunakan data yang diperoleh dari TREC Legal Trac dengan total lebih dari 3000 dokumen teks dan lebih dari 20 klasifikasi tetapi hanya enam klasifikasi yang dipilih saja. Hasil penelitian ini menunjukkan bahwa nilai optimal untuk k pada k-NN terjadi pada k = 13. Dengan menggunakan nilai k tersebut, rata-rata akrual mencapai 55,17 persen, lebih baik daripada menggunakan Naïve Bayes yaitu 39,01 persen [10].

Pada penelitian sebelumnya, penelitian klasifikasi menggunakan tiga metode yaitu k-NN, Naïve Bayes, dan Reverse DBSCAN, membandingkan performansi akurasi dengan melakukan pre-processing data untuk diolah dengan hasil terbaik yang diperoleh dengan metode naïve bayes [4].

Pada tahun 2016 dilakukan studi banding metode klasifikasi k-NN, Naïve Bayes, dan Decision Tree ditinjau dari kelemahan dan kekuatan masing-masing metode klasifikasi.

3. Implementasi dan pembahasan

3.1 Alur penelitian



Gambar 1. Alur penelitian

a. Persiapan Data

Data yang akan diolah berasal dari Desa Tamanwinangun, Kebumen, Jawa Tengah, Indonesia. Data yang diperoleh berjumlah lebih dari 861 keluarga dengan 26 parameter yaitu nomor kartu keluarga, kode provinsi, kode kabupaten, kode kecamatan, kode desa, alamat, nama rukun tetangga, nama kepala rumah tangga, jumlah anggota rumah tangga. , jumlah keluarga, status bangunan, luas lantai, tipe lantai, tipe dinding, tipe atap, jumlah kamar tidur, sumber air minum, cara mendapatkan air minum, sumber penerangan utama, bahan bakar / energi untuk memasak, penggunaan fasilitas buang air besar, sarana pembuangan sampah, status, pengelolaan, kabupaten, dan desa.

b. Proses Awal

Sebelum dataset dihitung data yang diperoleh perlu terlebih dahulu diproses terlebih dahulu, karena tidak semua data yang diperoleh digunakan dalam perhitungan di software RapidMiner. Ada dua tahapan yang dilakukan dalam proses preprocessing ini, yaitu tahap pembersihan data dan tahap pemilihan data.

Pembersihan Data

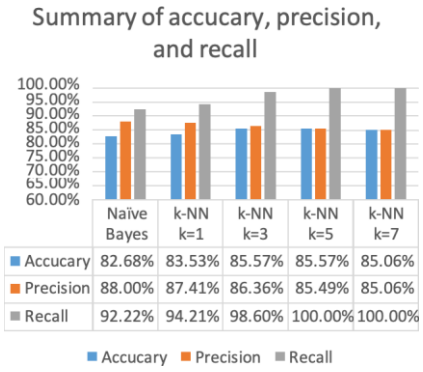
Pada dataset yang telah didapatkan terdapat 26 parameter, namun dalam penggunaan model klasifikasi naïve bayes dan juga k-nearest neighbour perlu dilakukan pemilihan parameter agar hasil perhitungan kedua model tersebut maksimal. Fungsi pembersihan data untuk memilih dari 26 parameter menjadi 12 parameter yang diinginkan digunakan untuk mengklasifikasikan naïve bayes dan k-nearest neighbour.

Pemilihan Data

Dari dataset tersebut terdapat beberapa record yang tidak memiliki nilai untuk masing- masing parameternya untuk menghindari kesalahan atau menurunkan performansi model klasifikasi naïve bayes dan k-nearest neighbour, sehingga perlu dilakukan dataset dengan melakukan pemilihan data dengan mengisi data dengan nilai parameter kosong dengan nilai sudah terdaftar dalam pilihan.

c. Metodologi

Metode klasifikasi penelitian yang digunakan adalah metode klasifikasi naïve bayes dan k-nearestneighbour. Data latih akan menggunakan 30 persen dari data yang didapat, dan sisanya akan digunakan sebagai data uji sebesar 70 persen dari data yang didapat. Pengolahan data menggunakan aplikasi software RapidMiner dengan menambahkan validasi berupa performance untuk melihat keakuratan kedua metode klasifikasi tersebut. Dalam performance, pengujian akan menggunakan metode confusion matrix yang terdiri dari akurasi, presisi, dan recall. yang mana dalam akurasi matriks kinerja digunakan untuk menguji seberapa akurat model klasifikasi. kemudian untuk menguji keakuratan data yang diberikan oleh model klasifikasi menggunakan ketelitian matriks kinerja.dan yang terakhir adalah matriks performance recall yang digunakan untuk menguji keberhasilan suatu model klasifikasi untuk menemukan kembali informasi. Kemudian akan dibuat grafik untuk menganalisis hasilnya.



	Naïve Bayes	k-NN k=1	k-NN k=3	k-NN k=5	k-NN k=7
Accucary	82.68%	83.53%	85.57%	85.57%	85.06%
Precision	88.00%	87.41%	86.36%	85.49%	85.06%
Recall	92.22%	94.21%	98.60%	100.00%	100.00%

Gambar 2 Akurasi Perhitungan

3.2 Pembahasan

Sebelum melakukan pengujian menggunakan aplikasi data RapidMiner yang akan digunakan sebagai data pengujian dan data latih terlebih dahulu dilakukan preprocessing yaitu pembersihan data terlebih dahulu dari data asli yang asli terdapat 26 parameter yang direduksi menjadi 12 parameter. Parameter yang diambil disesuaikan dengan kebutuhan metode klasifikasi naive bayes dan k-nearest neighbour. Kemudian dilakukan pemilihan data yaitu mengganti beberapa nilai yang kosong berdasarkan pilihan masing-masing parameter. Tujuannya adalah untuk mengurangi kesalahan dan memaksimalkan kinerja.

Kemudian menyiapkan dataset untuk pelatihan sebesar 30% dari total data. Dan juga data yang digunakan untuk pengujian adalah 70% dari total data. Setelah data siap digunakan, selanjutnya dilakukan pengujian menggunakan metode klasifikasi naïve bayes dan juga k- nearest neighbour. Dalam pengujian juga ditambahkan pengukuran kinerja dengan menggunakan metode matriks konfusi yang akan menghasilkan akurasi, presisi, dan recall.

Setelah dilakukan pengujian hasilnya dapat dilihat pada Gambar 3 menunjukkan bahwa menggunakan metode klasifikasi naïve bayes dari segi akurasi tidak lebih baik dari pada metode klasifikasi k-NN dengan k =

1. K-NN dengan k = 3 mengalami peningkatan dalam hal akurasi jika dibandingkan dengan k1 dan seterusnya. Untuk hasil dari k-NN mulailah menggunakan k = 7 sehingga menghasilkan nilai yang sama. Jika Anda melihat hasil yang paling optimal dari segi keakuratannya yaitu pada k = 3.

Ketepatan metode klasifikasi naïve bayes paling tinggi dibandingkan metode klasifikasi k-NN dengan nilai k berapapun. Dan untuk metode k-NN presisi tertinggi terjadi pada k1.

Nilai recall metode klasifikasi naïve bayes palingrendah jika dibandingkan dengan metode klasifikasi k-

NN dengan nilai k1. Ingat pada k-NN stabil dari k = 5 hingga k = 7.

4. Kesimpulan

Berdasarkan hasil penelitian yang dilakukan terhadap dataset keluarga miskin. Dengan membandingkan dua algoritma klasifikasi yaitu naïve bayes dan k-nn dengan nilai k = 1 - k = 7 maka hasil akurasi yang paling optimal adalah dengan menggunakan metode klasifikasi k-NN pada k = 3 dan k = 5 sebesar 85,57%. Presisi dan recall pada k- NN mulai stabil saat menggunakan nilai k = 7 dan seterusnya. Sedangkan untuk naïve bayes, nilai presisi tertinggi adalah 88.00%.

Untuk penelitian selanjutnya agar mendapatkan hasil yang lebih optimal maka perlu dilakukan preprocessing data dengan menggunakan transformasi dan rekayasa fitur. itu juga dapat menggunakan metode klasifikasi yangberbeda seperti jaringan saraf dan C4.5.

Kritik dan Saran :

Paper yang telah dibuat ini jauh dari kata sempurna karena keterbatasan pengetahuan tim penulis dan saat pembuatan paper ini tim penulis sedang isolasi mandiri karena terkonfirmasi positif virus corona. Berikut adalah saran yang dapat digunakan sebagai acuan penelitian selanjutnya.

1. Perlu dilakukan validasi akurasi model yang telah dibuat menggunakan teknik k-fold cross validation.
2. Perlu dilihat juga perbandingan distribusi kelas pada dataset sebelum diimplementasikan algoritma klasifikasi, supaya klasifikasi tidak cenderung pada kelas mayoritas.

Daftar Rujukan

[1] R. B. Akindola, “Towards a definition of poverty: Poor people’s perspectives and implications for poverty reduction,” Journal of Developing Societies, vol. 25, no. 2, pp. 121–150, Apr 2009, doi: 10.1177/0169796X0902500201.

[2] “Badan Pusat Statistik.” https://www.bps.go.id/pressrelease/2020/01/ 15/1743 /persentase-penduduk-miskin-september- 2019- turun-menjadi-9-22-persen.html (accessed May 02,2020).

[3] A. Mahdi, A. Razali, and A. Alwakil, “Comparison of Fuzzy Diagnosis with K- Nearest Neighbor and Naïve Bayes Classifiers in Disease Diagnosis.”

[4] Manav Rachna International University. Faculty of Engineering and Technology. Department of Computer Science and Engineering and Institute of Electrical and Electronics Engineers, Proceedings of the 2014 International Conference on Reliability, Optimization & Information Technology : ICROIT 2014 : 6-8 February 2014.

[5] S. D. Jadhav and H. P. Channe, “Comparative Study of K-NN, Naive Bayes and Decision Tree Classification Techniques,” 2013. [Online]. Available: www.ijsr.net.

[6] W. C. Org and S. S. Nikam, “): Pgs. 13-19 An International Open Free Access,”Peer Reviewed Research Journal Published By, vol. 8, no. 1, 2015, [Online]. Available: www.computerscijournal.org.

[7] J. Han, M. Kamber, and J. Pei, “Data Mining. Concepts and Techniques, 3rd Edition (The Morgan Kaufmann Series in Data Management Systems),” 2011.

[8] T. R. Patil and M. S. S. Sherekar, “Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification,” International Journal Of Computer Science And Applications, vol.6, no. 2,2013,[Online]. Available: http://www.cs.bme.hu/~kiskat/adatb/bank- data-.

[9] R. Arian, A. Hariri, A. Mehridehnavi, A. Fassihi, and F. Ghasemi, “Protein Kinase Inhibitors’ Classification Using K-Nearest Neighbor Algorithm,” Computational Biology and Chemistry, p. 107269, Apr. 2020, doi: 10.1016/j.compbiolchem.2020.107269.

[10] Z. E. Rasjid and R. Setiawan, “Performance Comparison and Optimization of Text Document Classification using k-NN and Naïve Bayes Classification Techniques,” in Procedia Computer Science, 2017, vol. 116, pp.107–112,doi: 10.1016/j.procs.2017.10.017.

[11] G. Gunadi, D. I. Sensuse “ Penerapan Metode Data Mining Market Basket Analysis Terhadap Data Penjualan Produk Buku Dengan Menggunakan Algoritma Apriori dan Frequent Patern Growth (FP- Growth): Studi Kasus Percetakan PT.Gramedia “, 2012.

[12] A. A. Fajrin, A. Maulana “ Penerapan Data Mining Untuk Analisis Pola Pembelian Konsumen Dengan Algoritma FP-Growth Pada Data Transaksi Penjualan Spare Part Motor “, 2018.

[13] W. D. Septiani “ Komparasi Metode Klasifikasi Data Mining Algoritma C4.5 Dan Naïve Bayes Untuk Prediksi Penyakit Hepatitis “, 2017.

[14] Maharani, N. A. Hasibuan, N. Silalahi, S. D. Nasution, Mesran, Suginam, D. U. Sutiksno,

H. Nurdiyanto, E. Buulolo, Yuhandri “ Implementasi Data Mining Untuk Pengaturan Layout Minimarket Dengan Menerapkan Association Rule “, 2017.

[15] Mrs. R. Sumithra, MCA, M.Phil., (Ph.D.), Dr (Mrs). S. Paul MCA, M.Phil, Ph.D. “Using distributed apriori association rule and classical apriori mining algorithms for grid based knowledge discovery “, 2010.

[16] Shashi Kant Shankar, Amritpal Kaur “Constraint Data Mining using Apriori Algorithm with And Operation “, 2016.