ISSN: 2339-2541
JURNAL GAUSSIAN, Volume 3, Nomor 4, Tahun 2014, Halaman 831 - 838
Online di: http://ejournal-s1.undip.ac.id/index.php/gaussian

PERBANDINGAN METODE KLASIFIKASI NAÏVE BAYES DAN K-NEAREST NEIGHBOR PADA ANALISIS DATA STATUS KERJA DI KABUPATEN DEMAK TAHUN 2012

Riyan Eko Putri[1], Suparti[2], Rita Rahmawati[3]
[1]Mahasiswa Jurusan Statistika FSM Universitas Diponegoro
[2,3]Staf Pengajar Jurusan Statistika FSM Universitas Diponegoro

ABSTRACT

Large population in Indonesia is closely related to the working status of the population which is unemployed or employed. It can lead to the high unemployment when the avaliable jobs arent balance with the population. Used two methods to perform the classification of employment status on the number of residents in the labor force in Demak for 2012 which is Naïve Bayes and K-Nearest Neighbor. Naïve Bayes is a classification method based on a simple probability calculation, while the K-Nearest Neighbor is a classification method based on the calculation of proximity. Variables used in determining whether a person's employment status is idle or not are gender, status in the household, marital status, education, and age. Employment status of the data processing methods of Naïve Bayes with the accuracy obtained is equal to 94.09% and the K-Nearest Neighbor method obtained is equal to 96.06% accuracy. To evaluate the results of the classification used calculations Press's Q and APER. Based on the analysis, the Press's Q values obtained indicate that both methods are already well in the classification of employment status data in Demak. Based on the calculation of APER, the classification of data in the employment status of Demak using the K-Nearest Neighbor method has an error rate smaller than the Naïve Bayes method. From this analysis it can be concluded that the K-Nearest Neighbor method works better compared with the Naïve Bayes for employment status data in the case of Demak for 2012.

Keywords : Classification, Naïve Bayes, K-Nearest Neighbor (K-NN), Classification evaluation

1. PENDAHULUAN

Indonesia merupakan negara yang luas dengan beribu pulau di dalamnya menyebabkan negara ini memiliki jumlah penduduk yang besar dengan karakteristik masyarakat yang bermacam-macam. Jumlah penduduk yang besar ini erat kaitannya dengan status kerja penduduknya apakah menganggur atau tidak menganggur (bekerja) dimana ketika tidak diimbangi dengan lapangan kerja yang tersedia dapat menyebabkan tingkat pengangguran yang tinggi.

Naïve Bayes dan K-Nearest Neighbor merupakan metode pengklasifikasi yang terkenal dengan tingkat keakuratan yang baik. Banyak penelitian telah dilakukan berkaitan dengan metode klasifikasi tersebut. Kebanyakan dari penelitian tersebut berbasiskan pada ilmu komputer atau informatika sehingga pada pembahasannya lebih ditekankan pada hasil pemrograman serta tema yang diambil berkaitan dengan hal-hal yang bersifat elektronik. Selain itu berbeda dengan metode pengklasifikasian dengan regresi logistik ordinal maupun nominal, pada metode Naïve Bayes dan K-Nearest Neighbor pengklasifikasian tidak diperlukan adanya permodelan maupun uji statistik seperti uji signifikansi.

Naïve Bayes merupakan metode pengklasifikasian peluang sederhana dengan asumsi antar variabel penjelas saling bebas (independen). K-Nearest

Neighbor atau dapat disingkat dengan K-NN adalah salah satu metode non parametrik yang digunakan dalam pengklasifikasian. Pada penulisan tugas akhir kali ini akan diaplikasikan kedua metode tersebut pada bidang statistika dengan permasalahan yang diangkat adalah kependudukan serta membandingkan keoptimalan dua metode tersebut dalam mengklasifikasi data status kerja di Kabupaten Demak pada tahun 2012.

2. TINJAUAN PUSTAKA

2.1 Status Kerja

Status kerja dibedakan menjadi dua yaitu bekerja dan menganggur. Sedangkan pengangguran sendiri terbagi menjadi dua macam yaitu pengangguran terbuka dan setengah pengangguran (BPS, 2012), namun pada penelitian ini digunakan data pada pengangguran terbuka. Pengangguran terbuka adalah mereka yang tidak bekerja dan saat ini sedang aktif mencari pekerjaan, termasuk juga mereka yang pernah bekerja atau sekarang sedang dibebastugaskan sehingga menganggur dan sedang mencari pekerjaan. Pengangguran umumnya disebabkan karena jumlah angkatan kerja atau para pencari kerja tidak sebanding dengan jumlah lapangan kerja (BPS, 2012).

2.2 Konsep Klasifikasi

Menurut Prasetyo (2012), klasifikasi merupakan suatu pekerjaan menilai objek data untuk memasukkannya ke dalam kelas tertentu dari sejumlah kelas yang tersedia. Dalam klasifikasi terdapat dua proses yang dilakukan yaitu dengan membangun model untuk disimpan sebagai memori dan menggunakan model tersebut untuk melakukan pengenalan atau klasifikasi atau prediksi pada suatu data lain supaya diketahui di kelas mana objek data tersebut dimasukkan berdasarkan model yang telah disimpan dalam memori.

Sistem dalam klasifikasi diharapkan mampu melakukan klasifikasi semua set data dengan benar, namun tidak dapat dipungkiri bahwa kesalahan akan terjadi dalam proses pengklasifikasian tersebut sehingga perlunya dilakukan pengukuran kinerja dari sistem klasifikasi tersebut. Umumnya, pengukuran kinerja klasifikasi dilakukan dengan matriks konfusi *(confusion matrix)*. Matriks konfusi merupakan tabel pencatat hasil kerja klasifikasi. Contoh dari matriks konfusi untuk dua kelas (biner) dapat dilihat pada Tabel 1.

Tabel 1. Matriks Konfusi untuk Klasifikasi Dua Kelas

f_{ij}		Kelas hasil prediksi (j)	
		Kelas = 1	Kelas = 2
Kelas asli (i)	Kelas = 1	f_{11}	f_{12}
	Kelas = 2	f_{21}	f_{22}

Setiap sel f_{ij} dalam matriks menyatakan jumlah rekord atau data dari kelas i yang hasil prediksinya masuk ke kelas j. Dari matriks konfusi dapat diketahui

jumlah data pemetaan yang diprediksi benar dengan cara menjumlahkan nilai f_{11} dan f_{22} ($f_{11} + f_{22}$) dan jumlah data pemetaan yang diprediksi salah dengan menjumlahkan nilai f_{21} dan f_{12} ($f_{21} + f_{12}$). Akurasi hasil prediksi dapat dihitung ketika jumlah data yang diklasifikasi secara benar maupun salah telah diketahui. Untuk menghitung akurasi digunakan formula:

$$\text{Akurasi} = \frac{\text{jumlah data yang diprediksi secara benar}}{\text{jumlah prediksi yang dilakukan}} = \frac{f_{11}+f_{22}}{f_{11}+f_{12}+f_{21}+f_{22}}$$

2.3 Klasifikasi Naïve Bayes

Naïve Bayes merupakan sebuah metode penggolongan berdasarkan probabilitas sederhana dan dirancang untuk dipergunakan dengan asumsi bahwa antar satu kelas dengan kelas yang lain tidak saling tergantung (independen). Pada klasifikasi Naïve Bayes, proses pembelajaran lebih ditekankan pada mengestimasi probabilitas. Keuntungan dari pendekatan ini yaitu pengklasifikasian akan mendapatkan nilai error yang lebih kecil ketika data set berjumlah besar (Berry, 2006). Selain itu menurut Han and Kamber (2006) klasifikasi Naïve Bayes terbukti memiliki akurasi dan kecepatan yang tinggi saat diaplikasikan ke dalam basis data dengan jumlah yang besar.

Formulasi Naïve Bayes untuk klasifikasi menurut Prasetyo (2012) adalah sebagai berikut:

$$P(Y|X) = \frac{P(Y)\prod_{i=1}^{q} P(X_i|Y)}{P(X)}$$

Dimana:

P(Y|X) = probabilitas data dengan vektor X pada kelas Y.

P(Y) = probabilitas awal kelas Y (*prior probability*).

$\prod_{i=1}^{q} P(X_i|Y)$ = probabilitas independen kelas Y dari semua fitur dalam vektor X.

Nilai P(X) = probabilitas dari X.

Probabilitas P(X) selalu tetap sehingga dalam perhitungan prediksi nantinya dapat diabaikan dan hanya menghitung bagian $P(Y)\prod_{i=1}^{q} P(X_i|Y)$ saja dengan memilih nilai yang terbesar sebagai kelas hasil prediksi atau yang biasa dikenal dengan sebutan *Maximum A Posteriori* (MAP) dimana MAP ini dapat dinotasikan dengan:

$$hMAP = \arg\ (\max P(Y)\prod_{i=1}^{q} P(X_i|Y))$$

Sementara probabilitas independensi $\prod_{i=1}^{q} P(X_i|Y)$ merupakan pengaruh semua fitur dari data terhadap setiap kelas Y, yang dinotasikan dengan:

$$P(X|Y=y) = \prod_{i=1}^{q} P(X|Y = y)$$

Setiap set fitur $X = [X_1, X_2, X_3, \dots, X_q]$ terdiri atas *q* atribut.

2.4 *Laplace Estimator*

Untuk menyiasati supaya hasil probabilitas pada perhitungan Naïve Bayes tidak bernilai nol dikarenakan tidak adanya data untuk suatu kategori tertentu dalam kelasnya dapat digunakan teknik estimasi yang biasa disebut *Laplace estimator* atau *Laplacian correction* (Han and Kamber, 2006). Dalam teknik ini digunakan penambahan nilai 1 pada data untuk masing-masing kategori ketika ada kategori yang memiliki nilai 0 sehingga untuk sebanyak k kategori dimana

j=1,2,...,k dan N= $N = \sum_{j=1}^{k} n_j$ jika masing masing kategori dalam kelasnya bernilai n_i maka $P(X=i) = \frac{n_i+1}{N+\text{bnyk\_kategori}}$

2.5 Klasifikasi K-Nearest Neighbor

Menurut Prasetyo (2012), algoritma Nearest Neighbor (kadang disebut K-Nearest Neighbor atau K-NN) merupakan algoritma yang melakukan klasifikasi berdasarkan kedekatan lokasi (jarak) suatu data dengan data yang lain. Dekat atau jauhnya lokasi (jarak) biasanya dihitung berdasarkan jarak Euclidean dengan rumus sebagai berikut (Han and Kamber, 2006):

$$d(x_i,x_j) = \sqrt{\sum_{l=1}^{N}(\text{diff}(x_{il}, x_{jl}))^2} \quad (1)$$

dengan :
x_{il} = data testing ke-i pada variabel ke-l
x_{jl} = data training ke-i pada variabel ke-l
$d(x_i,x_j)$ = jarak
N = dimensi data variabel bebas
$\text{diff}(x_{il},x_{jl})$ = *difference* atau ketidaksamaan

Penghitungan nilai *difference* atau ketidaksamaan pada persamaan (1) tergantung pada tipe data yang digunakan. Menurut Prasetyo (2012), penghitungan nilai ketidaksamaan berdasarkan tipe data untuk tiap variabel dapat diringkas seperti pada Tabel 2.

Tabel 2. Ketidakmiripan Dua Data dengan Satu Atribut

Tipe Atribut	Formula Jarak
Nominal	$\text{diff}_{(x_{il},x_{jl})} = \begin{cases} 0 & \text{Jika } x_{il} = x_{jl} \\ 1 & \text{Jika } x_{il} \neq x_{jl} \end{cases}$
Ordinal	$\text{diff}_{(x_{il},x_{lj})} = \lvert x_{il} - x_{jl} \rvert/(n-1)$ n adalah banyaknya pengkategorian dalam x
Interval atau Rasio	$\text{diff}_{(x_{il},x_{jl})} = \lvert x_{il} - x_{jl} \rvert$

2.6 Teknik Validasi

Cross-validasi (*cross validation*) atau yang sering disebut dengan estimasi rotasi merupakan teknik validasi model untuk menilai keoptimalan hasil analisis, selain itu cross-validasi juga merupakan teknik komposisi dalam penentuan banyaknya data training dan data testing yang akan digunakan. Ada beberapa metode dalam cross-validasi diantaranya yang pertama metode k-fold. Dalam metode k-fold, data disegmentasi secara random ke dalam k partisi yang

berukuran sama. Selama proses, salah satu dari partisi dipilih untuk menjadi data testing, sedangkan sisanya digunakan untuk data training.

Metode cross-validasi yang kedua yaitu metode holdout. Dalam metode holdout, data awal yang diberi label dipartisi ke dalam dua himpunan secara random yang dinamakan data training dan data testing. Proporsi data yang dicadangkan untuk data training dan data testing tergantung pada analisis misalnya 50%-50% atau 2/3 untuk training dan 1/3 untuk testing, namun menurut Witten (2005) serta Han and Kamber (2006) pada umumnya perbandingan yang digunakan yaitu 2:1 untuk data training berbanding data testing.

2.7 Evaluasi Ketepatan Hasil Klasifikasi

Untuk mengetahui apakah hasil klasifikasi yang didapatkan sudah akurat atau belum maka perlu dilakukan uji ketepatan hasil evaluasi. Untuk mengetahui ketepatan tersebut dapat dilakukan dengan menghitung nilai Press's Q dan APER. Press's Q merupakan pengujian untuk mengukur apakah klasifikasi yang dilakukan sudah akurat atau belum. Formulasi untuk menghitung Press's Q menurut Hair (2006) adalah:

Press's Q = $\frac{[N-(nK)]^2}{N(K-1)}$

dimana:

N = ukuran total sampel

n = banyak kasus yang diklasifikasi secara tepat

K = banyak grup

Pengklasifikasian dikatakan akurat apabila nilai Press's Q lebih besar dari pada nilai kritis yang diambil dari table *Chi-Square* dengan derajat bebas bernilai satu dan tingkat keyakinan sesuai yang diinginkan.

APER (*Apparent Error Rate*) atau yang disebut laju error merupakan ukuran evaluasi yang digunakan untuk melihat peluang kesalahan klasifikasi yang dihasilkan oleh suatu fungsi klasifikasi. Semakin kecil nilai APER maka hasil pengklasifikasian semakin baik (Prasetyo,2012).

Formulasi untuk menghitung APER (Johnson and Wichern, 2007) yaitu:

APER = $\frac{f_{12}+f_{21}}{f_{11}+f_{12}+f_{21}+f_{22}}$ x 100%

dimana:

f_{11} = banyak data dalam kelas 1 yang secara benar dipetakan ke kelas 1

f_{22} = banyak data dalam kelas 2 yang secara benar dipetakan ke kelas 2

f_{12} = banyak data dalam kelas 1 yang dipetakan secara salah ke kelas 2

f_{21} = banyak data dalam kelas 2 yang dipetakan secara salah ke kelas 1

3. METODOLOGI

Data yang digunakan dalam penelitian ini adalah data Status Kerja di Kabupaten Demak tahun 2012. Data bersumber dari Survei Angkatan Kerja Nasional (SAKERNAS) yang berjumlah 1375.

Pada penelitian tugas akhir ini digunakan beberapa atribut atau variabel di antaranya status dalam rumah tangga, jenis kelamin, umur, status perkawinan dan pendidikan. Sedangkan variabel yang akan diklasifikasikan adalah status kerja yang terdiri dari pengangguran dan bukan pengangguran.

Langkah-langkah yang dilakukan pada metode Naïve Bayes adalah sebagai berikut:

1. Membagi data menjadi 2 yaitu data testing dan data training.

2. Menghitung probabilitas prior (P(Y)) dari data testing berdasarkan data training.
3. Menghitung probabilitas atribut terhadap masing-masing kelas ($P(X_i|Y)$) pada data testing berdasarkan data training.
4. Menghitung perkalian probabilitas dengan probabilitas atribut pada masing-masing kelas ($P(Y)\ P(X_i|Y)$).
5. Mencari nilai maksimal dari $\frac{(P(Y)\ P(X_i|Y))}{P(X)}$ pada kedua kelas.
6. Nilai terbesar dari penghitungan merupakan hasil prediksi.
7. Mengevalusi hasil klasifikasi dengan menghitung nilai press's Q dan APER.

Langkah-langkah yang dilakukan dalam menganalisis data menggunakan metode K-Nearest Neighbor adalah sebagai berikut:

1. Menentukan nilai K dengan bilangan ganjil dan pada penelitian ini ditentukan dari 3, 5, dan 7.
2. Menghitung jarak sesuai dengan tipe datanya.
3. Menghitung jumlah data yang mengikuti kelas yang ada.
4. Menentukan hasil klasifikasi berdasarkan kelas yang memiliki anggota terbanyak.
5. Mengevaluasi hasil klasifikasi dengan menghitung nilai press's Q dan APER.

Setelah diperoleh nilai Press's Q dan APER dengan menggunakan kedua metode tersebut, kemudian nilai Press's Q dan APER tersebut dibandingkan dan diambil nilai yang terbesar untuk ditarik kesimpulan.

.

4. ANALISIS DAN PEMBAHASAN

4.1 Pengklasifikasian dengan Metode Naïve Bayes

Pengolahan data dengan teknik validasi holdout dimana digunakan data sebanyak 457 sebagai testing dan 918 data sisanya sebagai training yang diulang sebanyak 25 kali menghasilkan akurasi yang terbaik yaitu 94.09% dimana pengklasifikasiannya dapat dilihat pada Tabel 3.

Dari Tabel 3 diketahui bahwa keakurasian pada pengklasifikasian data status kerja di Kabupaten Demak tahun 2012 sebesar 0.9406 atau 94.06% dengan laju error sebesar 0.0591 atau 5.91%.

Tabel 3. Matriks Konfusi Naïve Bayes

f_{ij}		Kelas hasil prediksi	
		Kelas = 1	Kelas = 2
Kelas asli	Kelas = 1	7	11
	Kelas = 2	16	423

4.2 Pengklasifikasian dengan Metode K-Nearest Neighbor

Pengolahan data dengan teknik validasi holdout dimana digunakan data sebanyak 457 sebagai testing dan 918 data sisanya sebagai training serta

digunakan jumlah tetangga terdekat sebanyak 3, 5, 7 yang diulang sebanyak 25 kali menghasilkan laju error seperti pada Tabel 4.

Tabel 4. Laju Error Pada K-NN untuk Berbagai Nilai K

Jumlah K	Laju Error
3	0.0678
5	0.0394
7	0.0394

Dari Tabel 4 diketahui bahwa laju error terkecil dimulai pada K= 5 dan bernilai konstan untuk K>5 sehingga pada kasus ini diperoleh hasil yang maksimum dengan metode K-Nearest Neighbor untuk jumlah tetangga terdekat yaitu 5 dengan matriks konfusi seperti pada Tabel 5.

Tabel 5. Matriks Konfusi KNN

f_{ij}		Kelas hasil prediksi	
		Kelas = 1	Kelas = 2
Kelas asli	Kelas = 1	0	18
	Kelas = 2	0	439

Dari Tabel 5 diketahui bahwa keakurasian pada pengklasifikasian data status kerja di Kabupaten Demak tahun 2012 yaitu sebesar 0.9606 atau 96.06% dengan laju error sebesar 0.0394 atau 3.94%.

4.3 Evaluasi Ketepatan Hasil Prediksi

Untuk mengevaluasi ketepatan hasil prediksi digunakan nilai Press's Q dan APER, maka dalam hal ini masing-masing metode menghasilkan ketepatan yaitu nilai Press's Q pada hasil klasifikasi menggunakan metode Naïve Bayes sebesar 355.381. Nilai Press's Q pada hasil klasifikasi menggunakan metode K-Nearest Neighbor sebesar 421.709. Pengklasifikasian dengan menggunakan metode Naïve Bayes maupun K-Nearest Neighbor tersebut dikatakan akurat karena nilai Press's Q lebih besar dari nilai *Chi-Square* dengan derajat bebas bernilai satu dan tingkat kepercayaan 5% yaitu 3.841.

Nilai APER pada hasil klasifikasi menggunakan metode Naïve Bayes yaitu 5.91%. Nilai APER pada hasil klasifikasi menggunakan metode K-Nearest Neighbor yaitu 3.94%. Hasil tersebut menunjukkan bahwa hasil pengklasifikasian tersebut memiliki nilai APER yang rendah sehingga dapat disimpulkan bahwa hasil pengklasifikasian dengan metode Naïve Bayes maupun K-Nearest Neighbor tergolong baik.

4.4 Perbandingan Keakuratan

Keakuratan Metode Naïve Bayes dan K-Nearest Neighbor dapat dilihat dalam Tabel 6.

Tabel 6. Perbandingan Keakuratan

Metode	Akurasi	Laju Error	Press's Q
Naïve Bayes	0.9409	0.0591	355.381
KNN	0.9606	0.0394	421.709

Dari hasil pengklasifikasian pada Tabel 6 dapat disimpulkan bahwa metode K-Nearest Neighbor bekerja lebih baik dibandingkan dengan Naïve Bayes untuk kasus data status kerja di Demak dilihat dari tingginya akurasi pengklasifikasian. Selain itu laju error pada K-Nearest Neighbor cenderung lebih rendah daripada Naïve Bayes serta nilai Press's Q yang lebih tinggi.

5. KESIMPULAN

Berdasarkan analisis dan pembahasan yang telah dipaparkan pada bab sebelumya, maka dapat diambil kesimpulan sebagai berikut:

1. Berdasarkan hasil perhitungan *Press's Q* dapat dikatakan bahwa pengklasifikasian status kerja di Kabupaten Demak tahun 2012 dengan metode Naïve Bayes dan metode K-Nearest Neighbor sudah baik atau sudah akurat.
2. Berdasarkan perhitungan APER diperoleh nilai laju error untuk metode Naïve Bayes yaitu sebesar 0.0591 dan metode K-Nearest Neighbor sebesar 0.0394. Dari kedua nilai tersebut dapat dikatakan bahwa baik metode Naïve Bayes ataupun metode K-Nearest Neighbor memiliki peluang yang kecil untuk melakukan kesalahan dalam pengklasifikasian.
3. Dilihat dari nilai APER dan akurasi dapat disimpulkan bahwa pengklasifikasian menggunakan metode K-Nearest Neighbor lebih baik dibandingkan dengan metode Naïve Bayes dalam mengklasifikasikan status kerja di Kabupaten Demak tahun 2012.

6. DAFTAR PUSTAKA

Berry, I. H. and Browne, M. 2006. *Lecture Notes in DATA MINING.* USA: World Scientific.

BPS. 2012. Profil Ketenagakerjaan Kabupaten Demak.

Hair, J. F., Black. W. C., Babin. B. J. and Anderson R. E. 2006. *Multivariate Data Analysis. Seventh Edition.* Pearson Education Prentice Hall. Inc

Han, J and Kamber, M. 2006. *Data Mining Concepts and Techniques, second edition*. California: Morgan Kaufman.

Johnson, R. A. and Wichern. D. W. 2007. *Applied Multivariate Statistical Analysis. Sixth Edition*. New Jersey: Prentice Hall International. Inc

Prasetyo, E. 2012. *Data Mining Konsep dan Aplikasi Menggunakan MATLAB.* Yogyakarta: ANDI Yogyakarta

Witten, I. H. and Frank, E. 2005. *DATA MINING Practical Machine Learning Tools and Teachniques. Second Edition.* California: Morgan Kaufman.