Prosiding Seminar Nasional Lingkungan Lahan Basah
Volume 4 Nomor 3 Halaman 562-565 April 2019

p-ISSN 2623-1611
e-ISSN 2623-1980

PERBANDINGAN MODEL KLASIFIKASI *LINEAR DISCRIMINANT ANALYSIS* DAN *K-NEAREST NEIGHBOR* UNTUK DATA PENJURUSAN SISWA MADRASAH ALIYAH NEGERI SAMARINDA

Comparison of Classification Models Between Linear Discriminant Analysis and *k*-Nearest Neighbor for Students Majoring Data of Madrasah Aliyah Negeri in Samarinda

Nanda Arista Rizki *, Wasono, Yuki Novia Nasution
Laboratorium Matematika Komputasi, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Mulawarman,
Jalan Barong Tongkok No 4, Samarinda, Indonesia
*Penulis koresponden: nanda.arista.r@gmail.com

Abstract

Madrasah Aliyah Negeri (MAN) under Ministry of Religion used curriculum 2013 as well as others senior high schools. However, MAN subjects are more than others senior high schools. It's became particular concern to parents of students so that their children can compete with graduates of public schools when they continue to higher education. Curriculum 2013 provides that majoring in Ilmu Pengetahuan Alam (IPA), Ilmu Pengetahuan Sosial (IPS), Agama, and Bahasa have been implemented since the 10th grade. Picking the wrong major can be avoided by knowing the characteristics of students. The process of finding data patterns for student majors can be done using data mining. The purpose of this research was to classify the data of placement MAN's student majors using Linear Discriminant Analysis (LDA) model as linear model and *k*-Nearest Neighbor (*k*-NN) model as non-linear model. The data were resampled using Bootstrap with *n*=1000 and *n*=5000. The data were also divided into training data and testing data with the probability of each data being drawn was 60:40, 70:30, 80:20, and 90:10. Based on the results, the accuracy of LDA models for distribution of training data and testing the data at 60:40, 70:30, 80:20, and 90:10 and bootstrap *n*=5000 respectively were 0.838, 0.834, 0.833, 0.832. Meanwhile the accuracy of *k*-NN models with *k*=3 for distribution of training data and testing the data at 60:40, 70:30, 80:20, and 90:10 and bootstrap *n*=5000 respectively were 0.837, 0.846, 0.854, 0.861. Therefore, the *k*-NN model with *k*=3 became the best model.

Kata Kunci: classification model, curriculum, *k*-Nearest Neighbor, Linear Discriminant Analysis,

1. PENDAHULUAN

Berdasarkan publikasi BPS (2017), mayoritas penduduk (90,80%) Kota Samarinda menganut agama Islam. Masyarakat umum telah mengenal madrasah sebagai lembaga pendidikan keagamaan yang disediakan untuk mendalami agama Islam. Untuk mengikuti perkembangan zaman, madrasah telah mengubah paradigma sehingga seperti layaknya sekolah umum. Kurikulumnya, dalam hal ini Madrasah Aliyah (MA) sesuai dengan jurusan yang terdiri atas 4 jurusan, yaitu Ilmu Pengetahuan Alam (IPA), Ilmu Pengetahuan Sosial (IPS), Bahasa, dan Keagamaan. Kurikulum 2013 mengatur agar penjurusan untuk siswa dimulai pada kelas X. Pada sisi lain, kesiapan dan pemahaman yang kurang mengenai penjurusan akan berdampak pada peran pada siswa dan pihak MA.

MA mempersiapkan alumni untuk menjadi calon mahasiswa di berbagai perguruan tinggi. Jurusan yang telah diambil siswa di MA akan berdampak pada pemilihan jurusan saat kuliah di universitas. Diharapkan tidak ada kesalahan klasifikasi dalam penjurusan siswa kelas X. Artinya, siswa yang memang berpotensi dengan jurusan tertentu akan ditempatkan di jurusan tersebut.

Proses menemukan pola dari data proses penjurusan dapat dituntaskan menggunakan data mining. Beberapa teknik data mining yang dikenal adalah klasifikasi, *clustering*, asosiasi dan prediksi. Penelitian ini menggunakan teknik klasifikasi untuk menemukan model penjurusan siswa dengan cara memprediksi jurusan siswa berdasarkan beberapa variabel bebas. Karena kesederhanaan dan kecepatan proses klasifikasi, model yang digunakan adalah model Analisis Diskriminan Linier (ADL) dan model *k*-Nearest Neighbor (*k*-NN).

Model ADL dan model *k*-NN merupakan dua model berbeda. Metode ADL bertujuan untuk mencari kombinasi linier dari karakteristik antar kelas. Metode ini bersifat homogen dalam satu kelas yang sama, namun heterogen antar kelas yang

p-ISSN 2623-1611
e-ISSN 2623-1980

berbeda. Dengan menggunakan metode ini, hasil yang diperoleh dapat memisahkan karakteristik antar jurusan yang melekat pada siswa kelas X. Model *k*-NN mengklasifikasi data nilai siswa berdasarkan data *training* yang jaraknya paling dekat dengan nilai siswa. Hasil prediksi kelas diklasifikasi dengan cara dipilih hasil mayoritas dari kategori pada model *k*-NN.

Berdasarkan latar belakang itu tim penulis melakukan penelitian berjudul "Perbandingan Model Klasifikasi Analisis Diskriminan Linier dan *k-Nearest Neighbor* untuk Data Penjurusan Siswa Madrasah Aliyah Negeri Samarinda".

2. METODE

2.1. Model Klasifikasi Analisis Diskriminan Linier

Analisis diskriminan biasa digunakan untuk mengetahui hubungan kebergantungan antara satu variabel respon dengan dua atau lebih variabel bebas. Namun berbeda dengan analisis variansi dan analisis regresi, bahwa variabel respon analisis diskriminan berbentuk kategorik yang telah dikelompokkan dalam beberapa kelas. Analisis diskriminan bertujuan untuk mengklasifikasikan suatu pengamatan ke dalam kelas. Hasil klasifikasi bersifat homogen untuk setiap pengamatan yang berada dalam satu kelas, namun bersifat heterogen antar kelas yang terbentuk.

Dalam prosesnya, analisis diskriminan membuat batas pemisah kelas dan mengelompokkan amatan ke dalam kelompok yang sesuai. Analisis diskriminan melibatkan kombinasi dari dua atau lebih variabel bebas yang akan membentuk pemisah terbaik di antara kelas awal.

Bentuk linier dari analisis diskriminan menghasilkan interpretasi hasil analisis yang mudah dipahami. Oleh karena itu, penelitian yang dilakukan menggunakan Analisis Diskriminan Linier (ADL). Analisis ini cukup sensitif dengan kehadiran pencilan. Model ADL memproyeksikan masing-masing data kelas ke suatu garis penentu, sehingga hasil proyeksi dari masing-masing kelas memiliki selisih rata-rata yang besar dengan variansi yang terkecil untuk masing-masing kelas. Garis pemisah tersebut merupakan kombinasi linier dari variabel-variabel pilihan yang dapat memisahkan target kelas klasifikasi.

Misalkan variabel $\boldsymbol{Y}$ yang terdiri dari k kelas, akan diklasifikasikan berdasarkan *d* variabel bebas dengan *n* pengamatan. Lalu matriks $\boldsymbol{W}$ mentransformasi matriks $\boldsymbol{X}$ menjadi matriks $\tilde{\boldsymbol{Y}}$ yang memiliki *k* kemungkinan kombinasi linier dari variabel $\boldsymbol{X}$ sehingga seperti pada Persamaan (1).

$$\tilde{\boldsymbol{Y}} = \boldsymbol{W}'\boldsymbol{X} \quad (1)$$

Garis penentu yang digunakan sebagai kombinasi linier untuk mengklasifikasi adalah satu kolom dari matriks $\boldsymbol{W}$, sedemikian sehingga

$$y_j = (\boldsymbol{W}^*)'\boldsymbol{X} = w_1^* x_{1j} + w_2^* x_{2j} + \cdots + w_d^* x_{dj}, \quad (2)$$

untuk $j=1,2,\ldots,n$. Garis penentu dalam Persamaan (2) diukur menggunakan fungsi skor berikut

$$\lambda = \frac{W' S_B W}{W' S_W W},$$

dengan

$$S_B = \sum_{i=1}^{k} d\left(x_i - \bar{x}\right)\left(x_i - \bar{x}\right)'$$

$$S_W = \sum_{i=1}^{k}\sum_{j=1}^{n}\left(x_{ij} - \bar{x}_i\right)\left(x_{ij} - \bar{x}_i\right)'.$$

2.2 Model Klasifikasi *k-Nearest Neighbor*

Model klasifikasi *k-Nearest Neighbor* (*k*-NN) bertujuan untuk mengklasifikasikan obyek baru berdasarkan variabel bebas dan data *training*. Berbeda dengan ADL, kelas model *k*-NN tidak harus dapat dipisahkan secara linear. Klasifikasi model *k*-NN akan memprediksi kategori sampel data *testing* dari *k* sampel data *training* yang karakternya terdekat. Jauh tidaknya karakter dari variabel bebas, diukur berdasarkan jarak terpendek melalui formula Euclidean berikut.

$$D(train, test) = \sqrt{\sum_{l=1}^{d}\left(x_{l(train)} - x_{l(test)}\right)^2}, \quad (3)$$

untuk $test=1,2,\ldots,n$. Setelah diambil *k* tetangga terdekatnya, maka keputusan hasil klasifikasinya adalah modus kelas dari *k* hasil tersebut. Oleh karena itu, parameter dari model ini adalah nilai *k* dan jarak Euclidean-nya.

Walaupun implementasi model *k*-NN sangat sederhana, namun nilai *k* perlu ditentukan dengan waktu komputasi yang cukup lama. Hal ini dikarenakan bahwa banyaknya perhitungan jarak dari tiap *k* berbeda untuk suatu data *testing* pada keseluruhan data *training*.

2.3 Penjurusan di Madrasah Aliyah

Kurikulum 2013 di MA mengatur bahwa siswa kelas X MA yang naik ke kelas XI MA akan mengalami pemilihan jurusan. Di MA Jurusan yang disediakan lebih banyak dari pada jurusan yang ada di sekolah reguler seperti Sekolah Menengah Atas (SMA). Penjurusan yang tersedia di MA meliputi ilmu alam,

p-ISSN 2623-1611
e-ISSN 2623-1980

ilmu sosial, bahasa, dan keagamaan. Karena kompleksitasnya, penjurusan di MA menarik diteliti.

Proses penjurusan di MA dapat merujuk nilai Ujian Nasional (UN) dan Ujian Sekolah Berstandar Nasional (USBN) yang diperoleh siswa MA ketika lulus dari Sekolah Menengah Pertama (SMP). Mata pelajaran yang diujikan dalam UN adalah Bahasa Indonesia (INA), Bahasa Inggris (ING), Matematika (MTK), dan Ilmu Pengetahuan Alam (IPA), sedangkan yang diujikan dalam USBN adalah Pendidikan Agama Islam (PAI), Pendidikan Kewarganegaraan (PKN), Bahasa Indonesia (INA.1), Bahasa Inggris (ING.1), Matematika (MAT), dan Ilmu Pengetahuan Alam (IPA.1), dan Ilmu Pengetahuan Sosial (IPS). Nilai-nilai UN dan USBN dipercaya dapat mengukur capaian kompetensi siswa untuk memperoleh pengakuan atas prestasi belajarnya sesuai dengan Standar Kompetensi Lulusan (SKL). Baik nilai UN maupun nilai USBN dipercaya dapat mengukur pendidikan di suatu daerah, khususnya Samarinda, sehingga dijadikan acuan dalam perbaikan mutu pembelajaran.

Peminatan, kecakapan, dan kemampuan bakat siswa terhadap suatu jurusan tertentu akan mempengaruhi kualitas pencapaian hasil belajarnya. Diharapkan pemilihan ini cocok dengan karakteristik siswa, karena keberhasilan proses belajar siswa dapat diukur dari ketepatan dalam memilih jurusan.

3. HASIL DAN PEMBAHASAN

3.1 Analisis Deskriptif

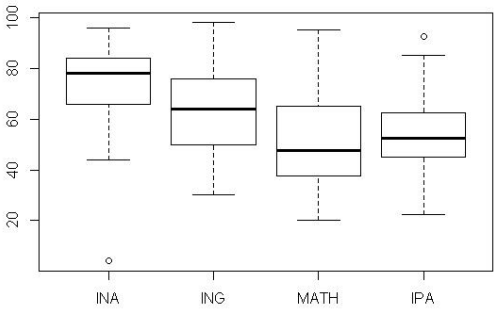
Data dalam penelitian adalah data siswa Madrasah Aliyah Negeri (MAN) di Samarinda kelas X tahun ajaran 2017/2018. Dengan rumus Slovin pada tingkat kepercayaan 95% dalam penentuan sampel, diperoleh sampel untuk kelas IPA, IPS, Agama, dan Bahasa masing-masing 118, 54, 50, dan 16.

Gambar 1. Diagram *pie* asal sekolah

Sebelum menganalisis model ADL dan *k*-NN, dilihat deskripsi umum data siswa MAN. Berdasarkan asal sekolah, siswa MAN mayoritas berasal dari sekolah keagamaan bernama Madrasah Tsanawiyah (MTs). Siswa yang asal sekolahnya dari MTs sebesar 70.59% dari sampel yang diambil (Gambar 1). Banyaknya sampel untuk kelas IPA, IPS, Agama, dan Bahasa masing-masing adalah 43.10%, 28.08%, 21.43% and 7.39% (Gambar 2).

Gambar 2. Diagram *pie* jurusan siswa MAN

Pola penyebaran data nilai UN disajikan pada Gambar 3. Ujian matematika dan IPA masih dianggap momok oleh siswa MAN. Nilai-nilainya lebih rendah dari keempat mata pelajaran UN; nilai mediannya kurang dari 60. Karakteristik nilai Bahasa Indonesia cenderung lebih baik.



Gambar 3. Diagram *boxplot* untuk nilai UN

Berdasarkan data nilai Ujian Sekolah Berstandar Nasional (USBN) (Gambar 4), kemampuan untuk setiap mata pelajaran hampir merata. Nilai untuk setiap mata pelajaran USBN berkisar 80 - 90, karena nilai media untuk setiap pelajaran USBN berada dalam interval 80-90.

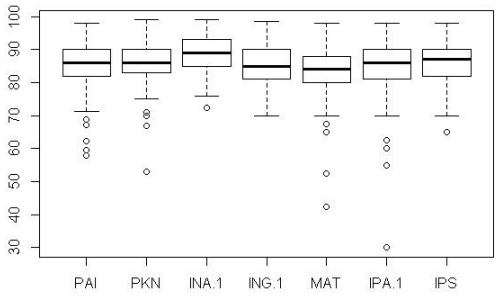
3.2 Hasil Klasifikasi Analisis Diskriminan Linier

Analisis Diskriminan Linier (ADL) dilakukan dengan menggunakan *open source software* bernama R versi 3.1.3. Persamaan garis penentu diskriminan yang diperoleh sebagai berikut.

ADL(***X***)= 0.0832 INA + 0.0112 ING + 0.0154 MTK - 0.0001 IPA + 0.0075 PAI + 0.0071 PKN +

p-ISSN 2623-1611
e-ISSN 2623-1980

0.0155 INA.1 - 0.0543 ING.1 + 0.1028 MAT + 0.0267 IPA + 0.0201 IPS.



Gambar 4. Diagram *boxplot* untuk nilai USBN

Dengan *resampling bootstrap* untuk *n*=1000 dan *n*=5000, serta pembagian data *training* dan data *testing* diperoleh hasil ketepatan klasifikasi model ADL sekitar 83% (Tabel 1).

Tabel 1. Hasil ketepatan klasifikasi model ADL dengan *resampling bootstrap n*=1000 dan *n*=5000 serta variasi pembagian data *training* dan data *testing*

Bootstrap	train:test	ADL
n = 1000	60:40	0.8377
	70:30	0.8341
	80:20	0.8327
	90:10	0.8319
n = 5000	60:40	0.8379
	70:30	0.8346
	80:20	0.8332
	90:10	0.8322

3.3 Hasil Klasifikasi Model *k*-NN

Analisis selanjutnya membandingkan dengan model non linier *k*-NN. Parameter yang dicobakan adalah *k*=2, *k*=3, *k*=4, dan *k*=5. *Software* yang digunakan adalah R 3.1.3. Setelah pembagian data *training* dan data *testing* 60:40, 70:30, 80:20, dan 90:10, serta *resampling bootstrap* untuk *n*=1000 dan *n*=5000 yang dilanjutkan dengan parameter *k* berbeda diperoleh ketepatan klasifikasi yang pada model *k*-NN optimumnya pada parameter *k*=3. Hasil klasifikasinya tertinggi dibanding parameter lain (Tabel 2).

4. SIMPULAN

Model terbaik adalah *k*-NN dengan nilai *k* = 3, karena ketepatan klasifikasi untuk *k*-NN tertinggi untuk setiap *sampling bootstrap* dan pembagian data *training-testing* berbeda. Hal ini berarti bahwa pengklasifikasian jurusan data seorang siswa MAN, dipertimbangkan berdasarkan 3 siswa lainnya dengan karakteristik nilai ujian sekolah dan ujian nasional yang mirip.

Tabel 2. Hasil ketepatan klasifikasi model analisis *k*-NN dengan *resampling bootstrap n*=1000 dan *n*=5000 serta variasi pembagian data *training* dan data *testing*

Bootstrap	train:test	*k*-NN			
		***k* = 2**	***k* = 3**	***k* = 4**	***k* = 5**
n = 1000	60:40	0.8268	**0.8371**	0.8029	0.7829
	70:30	0.8311	**0.8461**	0.8138	0.7908
	80:20	0.8328	**0.8538**	0.8215	0.7984
	90:10	0.8327	**0.8608**	0.8286	0.8060
n = 5000	60:40	0.8262	**0.8378**	0.8030	0.7826
	70:30	0.8307	**0.8460**	0.8119	0.7905
	80:20	0.8318	**0.8539**	0.8211	0.7977
	90:10	0.8337	**0.8611**	0.8291	0.8055

5. UCAPAN TERIMA KASIH

Tim peneliti mengucapkan terima kasih kepada Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Mulawarman atas bantuan dana pada skim penelitian dan pengabdian masyarakat pendanaan BOPTN tahun anggaran 2018. Tidak lupa kami ucapkan terima kasih kepada MAN 1 dan 2 di Samarinda yang berkenan memberi data.

6. DAFTAR PUSTAKA

BPS [Badan Pusat Statistik]. 2017. *Kota Samarinda dalam Angka*. Badan Pusat Statistik, Samarinda.

Crawley MJ. 2015. *Statistics: An Introduction Using R*, Second Edition. John Wiley& Sons, New York.

Dewi YS, Wati PI, Hadi AF. 2013. Analisis diskriminan linier dan kuadratik. *Majalah Ilmiah Matematika dan Statistika, Universitas Jember*, 13(1):1-10.

Kim KS, Choi HH, Moon CS, Mun CW. 2011. Comparison of k-nearest neighbor, quadratic discriminant and linear discriminant analysis in classification of electromyogram signals based on the wrist-motion directions. *Current Applied Physics* 11(3): 740-745. doi:10.1016/j.cap.2010.11.051.

Larose DT, Larose CD. 2014. *Discovering Knowledge in Data: An Introduction to Data Mining*, Second Edition. John Wiley, New Jersey.

McLachlan GJ. 2004. *Discriminant Analysis and Statistical Pattern Recognition*. John Wiley& Sons, New York.

Venables WN, Ripley BD. 2002. *Modern Applied Statistics with S*, Fourth Edition. Springer, New York.
