Indonesian Journal of Applied Statistics

Journal homepage: https://jurnal.uns.ac.id/ijas/index
E-ISSN 2621–086X
Volume 5 No. 1, May 2022

Perbandingan *K-Nearest Neighbor* dan *Random Forest* dengan Seleksi Fitur *Information Gain* untuk Klasifikasi Lama Studi Mahasiswa

Isran K. Hasan[1], Resmawan[2*], dan Jefriyanto Ibrahim[3]
[1,2,3]Jurusan Matematika Fakultas MIPA Universitas Negeri Gorontalo

*resmawan@ung.ac.id

Abstract. Accreditation is a quality and feasibility assessment form in carrying out higher education. One of the factors that affect accreditation is the length of student study. In this study, the length of student study is classified by using the best attributes resulting from selecting information gain features. In optimizing the classification algorithm, we process the data by converting the original data into data that is ready to be mined. The next step is dividing the data into training and testing data so that the classification algorithm can be applied. This study gives the best four attributes, with *K*-nearest neighbor (K-NN) classification of 86.67% and random forest classification of 100%.

Keywords: length of study; information gain; *K*-nearest neighbor; random forest

1. Latar Belakang

Akreditasi menjadi salah satu bentuk evaluasi mutu dan kelayakan program studi di suatu perguruan tinggi. Ketepatan lama studi mahasiswa menjadi masalah yang signifikan karena ketepatan ini adalah alasan berhasilnya suatu perguruan tinggi [1]. Dalam menjalankan masa studi sarjana/S1 mahasiswa dikatakan tepat waktu jika menyelesaikan studinya maksimal 4 tahun ataupun kurang dari itu [2]. Setiap perguruan tinggi berusaha untuk membenahi menejemennya dalam meningkatkan mutu pendidikan dan meningkatkan akreditasi. Salah satu komponen penilaian pada perguruan tinggi yaitu tingkat kelulusan tepat waktu.

Ketiadaan data dan analisa yang didapat oleh Bidang Akademik menyebabkan sulitnya melakukan klasifikasi terhadap lama studi mahasiswa. Klasifikasi lama studi mahasiswa bisa membantu Bidang Akademik untuk membuat metode yang sesuai dalam memperpendek dan mempersingkat lama studi mahasiswa. Untuk itu, penting dilakukan suatu pengujian klasifikasi dalam memprediksi seorang mahasiswa disebut lulus tepat waktu atau tidak berlandaskan informasi atau data yang diperoleh dari mahasiswa itu sendiri.

Klasifikasi yakni suatu operasi yang melakukan evaluasi pada suatu objek data sehingga masuk pada suatu kelas tertentu dari beberapa kelas yang ada [3]. Metode *K-nearest neighbor* (K-NN) dan *random forest* termasuk dalam metode klasifikasi. Metode K-NN adalah salah satu algoritma yang termasuk dalam *supervised* [4]. Digunakannya K-NN karena metode K-NN itu sendiri mampu diaplikasikan terhadap sejumlah data training yang banyak maupun sedikit, dan juga dalam pengoperasiannya lebih mudah, efektif dan gampang untuk dipahami. *Random forest*

adalah algoritma yang digunakan untuk masalah klasifikasi dalam *machine learning* dan *data mining* [5].

Penelitian menggunakan algoritma K-NN diantaranya dilakukan oleh Badu [6] yang mengklasifikasikan dana desa dengan tingkat akurasi 78,95% dengan nilai $K = 2$. Pada penelitian Subrata, dkk. [7], K-NN digunakan untuk mengklasifikasikan penggunaan protokol komunikasi pada trafik jaringan dengan tingkat akurasi 99,14 %. Selanjutnya, penelitian tentang *random forest* diantaranya dilakukan oleh Ratnawati dan Sulistyaningrum [8] yang menerapkan *random forest* untuk mengukur tingkat keparahan penyakit pada daun apel. Dalam hal ini diperoleh hasil akurasi sebesar 75,32%. Lebih lanjut Hanun dan Zailani [9] menerapkan klasifikasi *random forest* dalam menentukan kelayakan pemberian kredit. Penelitian tersebut menganalisis debitur yang bermasalah dan debitur tidak bermasalah dengan tingkat akurasi sebesar 87,88%. Dari hasil penelitian yang telah disebutkan menunjukkan tingkat akurasi yang baik dari algoritma K-NN maupun *random forest*.

Pada penelitian ini, algoritma K-NN dan *random forest* digunakan untuk mengklasifikasikan lama studi mahasiswa, sedangkan *information gain* digunakan untuk menyeleksi fitur-fitur yang tidak memiliki pengaruh. Hal ini sesuai dengan penelitian yang dilakukan Bimantoro dan Uyun [10] yang menggunakan *information gain* dalam menyeleksi fitur citra untuk menilai kesesuain lahan pada tanaman cengkeh. Dengan demikian akurasi yang diperoleh pada pengguaan fitur tanpa proses seleksi hanya 50%, sedangkan fitur yang didapat pada hasil seleksi dengan menggunakan *information gain* dengan nilai threshold 0,7 naik menjadi 88%. Selanjutnya, untuk melihat evaluasi dari sebuah model yang dibangun akan digunakan *confusion matrix*.

2. Landasan Teori

2.1. Klasifikasi. Klasifikasi adalah suatu siklus dalam mendapatkan suatu model atau manfaat yang menggambarkan dan mengenali informasi atau gagasan yang dimanfaatkan dalam menilai kelas item yang labelnya tidak diketahui [11]. Adapun jenis algoritma yang banyak dipakai dalam klasifikasi yaitu, *decision/classification trees, Bayesian classifiers/Naive Bayes classifiers, neural networks, K-nearest neighbor,* metode *rule based*, dan *support vector machines* (SVM). Berikut langkah-langkah dalam metode klasifikasi, yaitu:

1. Pembelajaran (*learning*): pelatihan (*training*) pada fase ini dibuat untuk menganalisa data *training* kemudian dipresentasikan.
2. Klasifikasi: data yang dicobakan digunakan untuk memperoleh ketepatan dari metode klasifikasi. Apabila ketepatan diterima, maka metode bisa digunakan pada klasifikasi data *tuple* yang baru.

2.2. *Preprocessing.* *Preprocessing* merupakan suatu proses penting yaitu mengurangi atribut yang tidak berpengaruh pada proses klasifikasi. Dalam tahap ini data yang digunakan masih dalam keadaan belum siap diolah, sehingga pada tahap ini data akan disiapkan agar mempermudah dalam proses klasifikasi. *Preprocessing* dibutuhkan dalam mengoptimalkan kemampuan algoritma klasifikasi [12]. Biasanya ada empat langkah dalam *preprocessing* untuk dokumen teks, yakni *case folding, tokenizing, stopwords removal* dan *stemming* [13].

2.3 *Information Gain.* *Information gain* adalah suatu metode yang bertujuan sebagai pembatas yang akan digunakan untuk suatu karakter atau atribut yang tersedia, minimal 1 atau lebih atribut yang akan digunakan, dan merupakan cerminan dari sifat sifat yang akan dimanfaatkan [14]. *Information gain* membantu dalam mereduksi atau mengelola *noise* yang diakibatkan oleh fitur

immaterial. *Information gain* mampu mengidentifikasi fitur yang memiliki banyak data yang terkandung dalam suatu informasi dalam pandangan kelas tertentu. Dalam memilih atribut terbaik diselesaikan dengan menghitung nilai *entropy* terlebih dahulu. *Entropy* adalah jenis kerentanan kelas dengan memanfaatkan peluang kejadian atau sifat tertentu [15]. Penentuan fitur dengan *Information gain* dilakukan dengan 3 tahap [16], yakni:

1. Menghitung nilai *information gain* untuk setiap atribut pada dataset.
2. Memastikan garis batas (*threshold*) yang diperlukan. Ini akan memungkinkan atribut dengan bobot yang setara dengan garis batas atau lebih menonjol untuk ditahan pada atribut yang berada di bawah batas.
3. Dataset diperbaiki dengan menghilangkan atribut yang tidak relevan.

2.4. *K*-Nearest Neighbor (K-NN). K-NN adalah metode yang menjalankan klasifikasi berlandaskan pada kedekatan suatu jarak data dengan data lainnya [3]. Pada K-NN nilai *K* berarti data yang paling dekat dari data uji. Karena sederhana dalam melakukan proses klasifikasi pada kelompok data, metode K-NN menjadi salah satu metode pengenalan pola yang umum dan sering dimanfaatkan. Cara kerja K-NN itu sendiri yaitu dengan mencari jarak antara dua titik yakni titik pelatihan dan titik uji, yang selanjutnya dilakukan penilaian dengan *K* tetangga paling dekat dengan data latih. Pada penelitian ini akan menggunakan pengukuran jarak dengan *Euclidean distance.* Adapun rumus dari *Euclidean distance* dipresentasikan pada persamaan berikut [17]:

$$d_{(x_i x_j)} = \sqrt{\sum_{r=1}^{n} = (x_i - x_j)^2}$$

Ada beberapa hal yang dapat mempengaruhi hasil K-NN, diantaranya yaitu menentukan nilai *K*. Jika *K* terlalu kecil maka akan berdampak pada hasil perkiraan atau prediksi yang diperoleh bisa sensitif pada adanya noise. Apabila *K* terlalu besar, maka tetangga paling dekat yang dipilih terlalu banyak dari kelas lain yang tidak relevan karena jaraknya terlalu jauh. Pemilihan nilai *K* genap atau ganjil juga menjadi perhatian. Untuk *K* genap dengan jumlah klasifikasi genap akan ada kemungkinan voting dari kedua klasifikasi mendapat suara yang sama. Akan tetapi untuk *K* ganjil dengan jumlah klasifikasi genap akan memudahkan karena dijamin kedua kelas tidak akan mendapat suara yang sama [3].

2.5. *Random forest.* *Random forest* adalah sekelompok tree dimana tiap-tiap tree tergantung pada jumlah piksel untuk setiap vektor yang diambil dengan acak dan independent [18]. Metode *random forest* merupakan model klasifikasi yang dipakai dengan menembangkan beberapa pohon keputusan berdasarkan seleksi data dan variabel yang dilakukan dengan acak. Hasil *random forest* yaitu sekumpulan pohon acak. Kelas yang dihasilkan dari metode klasifikasi dipilih dari kelas dengan angka paling banyak yang dibuat oleh pohon acak yang ada [19].

Banyak pohon yang ditumbuhkan kemudian terbentuk hutan atau yang kenal dengan *forest*, selanjutnya kumpulan pohon tersebut dianalisis sehingga menjadi metode *random forest*. Pada sekelompok data yang tersusun dari *n* yang diamati dan *p* peubah penjelas, *random forest* dikerjakan dengan cara berikut:

1. Di grup data, lakukan pemeriksaan tidak teratur ukuran *n* dengan pemulihan. Tahapan ini dikenal tahapan *bootstrap*.
2. Memanfaatkan kasus bootstrap, pohon dibuat dengan ukuran paling besar (tanpa pemangkasan). Di setiap simpul, pemilihan diselesaikan dengan memilih *m* faktor penjelas dengan acak, di mana $m < p$. Pemilihan terbaik dipilih dari m faktor informatif.

Tahap ini merupakan penentuan komponen yang tidak beraturan. Tahapan ini merupakan random feature selection.

3. Kemudian mengulang tahap 1 dan 2 sejumlah *k* kali, kemudian tercipta dari hutan yang tersusun atas *k* pohon.

2.6. *Confusion Matrix*. *Confusion matrix* salah satu metode yang sering dipakai dalam melakukan pengujian akurasi pada konsepsi data mining. Sistem yang menjalankan klasifikasi diharapkan mampu melakukan klasifikasi semua dataset dengan tepat, namun tidak bisa dipungkiri juga untuk hasil kerja dari sistem belum mampu bisa 100% tepat, maka sebuah sistem ini harus diukur kinerjanya [3]. Pengujian *confusion matrix* mengutarakan hasil penilaian model dengan memanfaatkan table matrix. Jika dataset tersusun atas dua kelas, maka kelas pertama dikatakan positif dan kelas kedua dikatakan negatif. Tabel *confusion matrix* disajikan pada Tabel 1.

Tabel 1. Tabel *Confusion matrix*

	Classified as	
Correct Classification	*Predicted "+"*	*Predicted "-"*
Actual "+"	*True Positives*	*False Negatives*
Actual "-"	*False Negatives*	*True Positives*

2.7. Lama Studi Mahasiswa. Salah satu bentuk evaluasi dari akreditasi perguruan tinggi yaitu lama studi [20]. Lama studi merupakan waktu yang diperlukan mahasiswa dalam menyelesaikan pendidikan yang ditunjukkan oleh tiap-tiap tingkatan, umumnya untuk tingkat sarjana adalah 4 tahun. Disamping itu, kelulusan tepat waktu menjadi masalah penting mengingat tingkat kelulusan menjadi alasan efektifnya perguruan tinggi [1].

Adapun variabel-variabel yang berhubungan dengan lama studi untuk seorang mahasiswa dalam penelitian ini yaitu, tempat lahir, jenis kelamin, predikat, seleksi, dosen penasehat akademik, Jenis sekolah, pekerjaan orang tua, pendapatan orang tua, indeks prestasi kumulatif (IPK), sistem kredit semester (SKS), jumlah cuti/nonaktif, jumlah mata kuliah nilai bagus, jumlah mata kuliah nilai buruk dan waktu studi. Penelitian yang akan diteliti yakni Program Studi Pendidikan Matematika, Jurusan Matematika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Negeri Gorontalo angkatan 2013 yang lulus di tahun 2017 hingga 2019.

3. Metode Penelitian

Data pada penelitian merupakan data sekunder berupa data mahasiswa Program Studi Pendidikan Matematika, Jurusan Matematika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Negeri Gorontalo angkatan 2013 yang lulus ditahun 2017 hingga tahun 2019. Data tersebut didapat dari Tata Usaha (TU) Jurusan Matematika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Negeri Gorontalo.

Sebelum masuk pada proses klasifikasi untuk kedua model, hal pertama yang dilakukan yakni mempersiapkan data, menghitung *information gain* yang digunakan untuk menyeleksi atribut yang tidak berpengaruh, selanjutnya data dikonversi sehingga dapat digunakan pada kedua metode klasifikasi. Data dibagi menjadi dua bagian yaitu data *training* dan data *testing*, masing-masing sebesar 80% dan 20%.

4. Hasil dan Pembahasan

Jumlah seluruh sampel data pada penelitian ini yakni 75 sehingga untuk nilai *s*=75, dengan 42 orang tepat waktu dan 33 orang yang lewat batas waktu dan terdapat atribut dan salah satunya yakni atribut target yang akan dilihat pengaruhnya dalam suatu klasifikasi. Jumlah seleksi fitur menggunakan asumsi penelitian sebelumnya yang menjelaskan bahwa jumlah seleksi fitur yang disarankan adalah $\log_2 n$ dengan n adalah jumlah seluruh fitur [21]. Jadi pada penelitian ini akan digunakan 5 atribut yang memiliki nilai pengaruh terbesar, karena $\log_2 15 = 3{,}9 \approx 4$ yang ditunjukkan pada Tabel 2.

Tabel 2**.** Atribut yang berpengaruh

Atribut	Nilai pengaruh
Dosen PA	0,510
Tempat Lahir	0,736
Asal Sekolah	0,763
Waktu Studi	0,990

4.1. *Preprocessing* Data. Preprocessing dibutuhkan dalam memaksimalkan kinerja algoritma klasifikasi [7]. Data yang digunakan dalam sistem *mining* umumnya tidak dalam kondisi optimal untuk ditangani. Jadi dalam penelitian ini hasil atribut yang telah diperoleh pada proses *information gain* selanjutnya akan diproses dengan mengkonversi data mengubah data dari bentuk asalnya menjadi data yang siap untuk dianalisis. Adapun hasil konversi data yang diperoleh dari tahap preprocessing data sebagai berikut:

1. Lama studi (*Y*). Pada lama studi tidak terjadi perubahan karena lama studi merupakan atribut target dan tidak akan mempengaruhi proses.
2. Waktu studi (*X1*) Berhubung untuk waktu studi mahasiswa datanya berbentuk numerik maka data ini tidak terjadi perubahan.
3. Asal sekolah (*X2*) Untuk asal sekolah dikelompokan berdasarkan kabupaten yang ada di provinsi Gorontalo. Adapun untuk yang dari luar Gorontalo akan digabung menjadi satu kelompok.
4. Tempat lahir (*X3*) Sama halnya denga nasal sekolah, untuk tempat lahir dikelompokan berdasarkan kabupaten yang ada di provinsi Gorontalo. Adapun untuk yang dari luar Gorontalo akan digabung menjadi satu kelompok.
5. Dosen PA (*X4*) Untuk Dosen PA dikelompokan berdasarkan dosen-dosen yang berada di lingkungan jurusan itu sendiri.

Selanjutnya data dibagi menjadi data *training* dan data *testing*. Pada penelitian ini data *training* digunakan untuk direpresentasikan dalam bentuk aturan klasifikasi, selanjutnya data *testing* digunakan dalam memprediksi akurasi dari aturan klasifikasi. Pembagian data ini dilakukan dengan pemilihan secara random, yaitu 80% untuk data *training* dan 20% data *testing*.

4.2. Klasifikasi K-NN. Algoritma K-NN adalah teknik yang menjalankan algoritma *supervised*, yang bermaksud untuk mengklasifikasi objek baru berlandaskan atribut dan data sampel. Pada proses klasifikasi dengan K-NN terdapat tiga alur yakni pemilihan parameter *K*, menghitung jarak Euclid antara data *training* dan data *testing*, dan menentukan rangking dari hasil perhitungan jarak. Hasil Perhitungan jarak dan telah diurutkan ditampilkan pada Tabel 3.

Berdasarkan Tabel 3 dan Tabel 4 dapat dilihat bahwa dengan menggunakan 1-NN, pada data *testing* pertama yaitu mahasiswa dengan lama studi 3,9 tahun, berasal dari sekolah dan lahir

di Kab. Gorontalo Utara (4), dan menjadi Bimbingan dosen PA dari Bapak Drs. Sumarno Ismail, M.Si (4) diklasifikasikan lulus tepat waktu. Hasil prediksi di tampilkan pada Tabel 4.

Tabel 3**.** Ranking Jarak Euclid

Y	*X1*	*X2*	*X3*	*X4*	Jarak Setelah Diurutkan	Rank	Klasifikasi *K*=1
1	4	2	2	5	3.00	1	**ya**
1	4	5	5	1	3.32	2	tidak
2	5.6	3	7	2	4.11	3	tidak
1	4	7	2	2	4.12	4	tidak
1	4	6	6	1	4.12	5	tidak
1	3.9	7	7	3	4.36	6	tidak
1	4	7	7	3	4.36	7	tidak
…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…
2	5.6	1	10	19	16.52	59	tidak
2	5.6	1	1	20	16.64	60	tidak

Tabel 4. Tabel hasil klasifikasi untuk *k* = 1.

Y	*X1*	*X2*	*X3*	*X4*	Hasil klasifikasi K-NN k = 1
1	3.9	4	4	4	1
1	3.9	1	1	5	1
1	4	7	7	7	1
1	4	9	9	9	1
1	4	7	7	13	1
1	4	2	1	13	1
1	4	2	1	1	1
1	4	7	7	1	1
1	4	7	7	14	1
1	4	3	3	14	1
2	5.6	2	3	20	2
2	5.6	8	8	4	2
2	6	1	1	6	1
2	6.7	7	8	6	1
2	6.7	10	10	19	2

Setelah mendapatkan hasil prediksi dari seluruh data *testing* menggunakan *K*=1, dapat dilakukan evaluasi hasil klasifikasi metode K-NN dengan menggunakan *confusion matrix*. Pada siklus penyusunan dengan teknik K-NN, jumlah objek yang benar dan salah diklasifikasikan pada tiap-tiap kelompok (Tabel 5). Tanda (*) pada angka menunjukkan jumlah objek kelompok tertentu yang salah diklasifikasikan dengan menjalankan metode K-NN.

Berdasarkan Tabel 5 dapat diketahui bahwa Lama Studi Mahasiswa Pendidikan Matematika angktan 2013 dengan menggunakan algritma K-NN diperoleh hasil yakni dari 10 orang dengan kelas tepat waktu dan tidak terdapat kesalahan dalam klasifikasi. Dari 5 orang yang lewat batas waktu, terdapat 3 orang yang dapat diklasifikasikan dengan benar dan 2 orang lainya

di klasifikasikan tepat waktu dengan demikian terdapat 2 orang yang tidak dapat diklasifikasikan dengan benar. Tingkat akurasi untuk *K*=1 sebesar 86,67%. Selanjutnya dalam memudahkan menghitung nilai K lainnya digunakan bantuan aplikasi R, dan didapatkan hasil 86,67% untuk *K*=3, 73,33% untuk *K*=5, 66,66% untuk *K*=7, 66,66% untuk *K*=9, 66,66% untuk *K*=11, 66,66% untuk *K*=13, 60% untuk *K*=15, 60% untuk *K*=17, 53,3%3 untuk *K*=19, dan 60% untuk *K*=21.

Tabel 5. Tabel hasil klasifikasi K-NN untuk *K*=1

Klasifikasi lama studi	Prediksi klasifikasi		Total
	Tepat waktu	Lewat batas waktu	
Tepat waktu	10	0*	10
Lewat batas waktu	2*	3	5
Total	12	3	15

4.3. Klasifikasi *Random Forest*. *Random forest* merupakan metode klasifikasi yang tersusun dari sejumlah pohon keputusan berbagai subset dari dataset dan mengambil rata-rata dalam meningkatkan akurasi prediksi dari dataset tersebut. *Random forest* mengambil prediksi dari setiap pohon berdasarkan pada suara mayoritas. Sintak model *random forest* disajikan pada Gambar 1.

```
Call:
 randomForest(formula = as.factor(Y) ~ ., data = latih)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 2
        OOB estimate of  error rate: 0%
Confusion matrix:
  1  2 class.error
1 32  0         0
```

Gambar 1. Sintak model *random forest*

Sintak yang diberikan pada Gambar 1 menunjukkan bahwa jenis *random forest* yang terbentuk merupakan klasifikasi dengan jumlah pohon yang dibuat sebanyak 500 dan banyaknya variabel yang digunakan pada tiap iterasinya sebanyak 2 dengan perkiraan tingkat kesalahan OOB pada data *training* yang digunakan 0%. Dari Gambar 1 juga dapat dilihat semua data berhasil dimodelkan dengan benar, sehingga *class error* yang dihasilkan sebesar 0%.

Setelah model terbentuk pada data *training*, tahap selanjutnya adalah menguji data *testing* untuk melihat ketepatan model yang didapat. Hasil prediksi dari seluruh data *testing* dengan menggunakan model *random forest*, selanjutkan dapat dilakukan evaluasi hasil klasifikasi metode *random forest* dengan menggunakan *confusion matrix* ditampilkan pada Tabel 6. Berdasarkan pada Tabel 6 dapat diketahui bahwa Lama Studi Mahasiswa Pendidikan Matematika angktan 2013 dengan menggunakan algoritma *random forest* diperoleh hasil yakni dari 10 orang dengan kelas tepat waktu dan tidak terdapat kesalahan dalam klasifikasi. Selain itu, dari 5 orang yang lewat batas waktu, tidak terdapat kesalahan, sehingga dapat dilihat bahwa tingkat akurasi dengan algoritma *random forest* sebesar 100%.

Tabel 6. Tabel hasil klasifikasi

Klasifikasi lama studi	Prediksi klasifikasi		Total
	Tepat waktu	Lewat batas waktu	
Tepat waktu	10	0*	10
Lewat batas waktu	0*	5	5
Total	10	5	15

4.4. Perbandingan Tingkat Akurasi. Pengukuran tingkat akurasi baik pada algoritma K-NN maupun *Random forest* diselesaikan dengan memastikan peluang kesalahan klasifikasi. Dalam proses klasifikasi diharapkan melakukan klasifikasi pada semua obyek dengan benar, sehingga semakin kecil kesalahan klasifikasi membuktikan bahwa semakin baik hasil klasifikasi yang diperoleh. Pada penelitian ini didapatkan bahwa algoritma yang memiliki tingkat akurasi model terbesar adalah algoritma *random forest* dengan akurasi model sebesar 100%, lebih unggul daripada algoritma K-NN dengan akurasi model sebesar 86,67%. Hal ini menunjukkan bahwa algoritma *random forest* bekerja lebih baik daripada algoritma K-NN dalam mengklasifikasikan Lama Studi Mahasiswa.

5. Kesimpulan

Pembahasan hasil menunjukkan bahwa dari 15 atribut dan salah satu diantaranya adalah atribut target, didapatkan 4 atribut terbaik berdasarkan seleksi fitur *information gain*. Algoritma *random forest* bekerja lebih baik dibandingkan dengan algoritma K-NN dalam mengklasifikasikan lama studi mahasiswa.

Daftar Pustaka

[1] A. H. Nasrullah, "Penerapan Metode C4.5 untuk Klasifikasi Mahasiswa Berpotensi Drop Out," Ilk. J. Ilm., vol. 10, no. 2, pp. 244–250, Sep. 2018, doi: 10.33096/ilkom.v10i2.300. 244-250.

[2] Keputusan Menteri Pendidikan Nasional, Keputusan Menteri Pendidikan Nasional Republik Indonesia Nomor 232/U/2000 Tentang Pedoman Penyusunan Kurikulum Perguruan Tinggi, 2000.

[3] E. Prasetyo, Data Mining: konsep dan aplikasi menggunakan MATLAB, 1st ed. Yogyakarta: Andi, 2012.

[4] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques. New York: Morgan Kaufmann, 2012.

[5] D. T. Larose and C. D. Larose, Discovering Knowledge in Data: An Introduction to Data Mining, Vol 4. New York: John Wiley & Sons, Inc, 2014.

[6] Z. S. Badu, "Penerapan Algoritma K-Nearest Neighbor Untuk Klasifikasi Dana Desa," J. Inform., 2016.

[7] K. K. A. Subrata, I. M. O. Widyantara, dan L. Linawati, "Klasifikasi Penggunaan Protokol Komunikasi Pada Trafik Jaringan Menggunakan Algoritma K-Nearest Neighbor," Maj. Ilm. Teknol. Elektro, vol. 16, no. 1, p. 67, Jul. 2016, doi: 10.24843/MITE.1601.10.

[8] L. Ratnawati and D. R. Sulistyaningrum, "Penerapan Random Forest untuk Mengukur Tingkat Keparahan Penyakit pada Daun Apel," J. Sains dan Seni ITS, vol. 8, no. 2, Jan. 2020, doi: 10.12962/j23373520.v8i2.48517.

[9] A. U. Zailani and N. L. Hanun, "Penerapan Algoritma Klasifikasi Random Forest Untuk Penentuan Kelayakan Pemberian Kredit Di Koperasi Mitra Sejahtera," Infotech J. Technol. Inf., vol. 6, no. 1, pp. 7–14, Jun. 2020, doi: 10.37365/jti.v6i1.61.

[10] D. A. Bimantoro and S. Uyun, "Pengaruh Penggunaan Information Gain untuk Seleksi

Fitur Citra Tanah dalam Rangka Menilai Kesesuaian Lahan pada Tanaman Cengkeh," JISKA (Jurnal Inform. Sunan Kalijaga), vol. 2, no. 1, pp. 42–52, Aug. 2017, doi: 10.14421/jiska.2017.21-06.

[11] F. Gorunescu, Data Mining: Concepts, models and techniques. New York: Springer, 2011.

[12] W. Sun, Z. Cai, Y. Li, F. Liu, S. Fang, and G. Wang, "Data Processing and Text Mining Technologies on Electronic Medical Records: A Review," J. Healthc. Eng., vol. 2018, pp. 1–9, 2018, doi: 10.1155/2018/4302425.

[13] S. F. Crone, S. Lessmann, and R. Stahlbock, "The impact of preprocessing on data mining: An evaluation of classifier sensitivity in direct marketing," Eur. J. Oper. Res., vol. 173, no. 3, pp. 781–800, Sep. 2006, doi: 10.1016/j.ejor.2005.07.023.

[14] A. S. Budiman dan X. A. Parandani, "Uji Akurasi Klasifikasi Dan Validasi Data Pada Penggunaan Metode Membership Function Dan Algoritma C4.5 Dalam Penilaian Penerima Beasiswa," Simetris J. Tek. Mesin, Elektro dan Ilmu Komput., vol. 9, no. 1, pp. 565–578, Apr. 2018, doi: 10.24176/simet.v9i1.2021.

[15] N. A. Shaltout, M. El-Hefnawi, A. Rafea, and A. Moustafa, "Information Gain as a Feature Selection Method for the Efficient Classification of Influenza Based on Viral Hosts," in Proceedings of the World Congress on Engineering, 2014, pp. 625–631.

[16] M. R. Maulana and M. A. Al-Karomi, "Information Gain untuk Mengetahui Pengaruh Atribut Terhadap Klasifikasi Persetujuan Kredit," J. LITBANG KOTA PEKALONGAN, vol. 9, pp. 113–123, 2015.

[17] M. Lestari, "Penerapan Algoritma Klasifikasi Nearest Neighbor (K-Nn) Untuk Mendeteksi Penyakit Jantung," Fakt. Exacta, vol. 7, no. 4, pp. 366–371, 2014.

[18] L. Breiman, "Random Forest," Mach. Learn., vol. 45, pp. 5–32, 2001, doi: https://doi.org/10.1023/A:1010933404324.

[19] G. Biau, "Analysis of a Random Forests Model," J. Mach. Learn. Res., vol. 13, pp. 1063–1095, 2012, doi: https://doi.org/10.48550/arXiv.1005.0208.

[20] I. M. Budi Adnyana, "Prediksi Lama Studi Mahasiswa Dengan Metode Random Forest (Studi Kasus : Stikom Bali)," CSRID (Computer Sci. Res. Its Dev. Journal), vol. 8, no. 3, pp. 201–208, Oct. 2016, doi: 10.22303/csrid.8.3.2016.201-208.

[21] A. Roihan, Seleksi fitur menggunakan Symmetrical Uncertainty pada Prediksi Cacat Perangkat Lunak, Universitas Islam Negeri Maulana Malik Ibrahim, 2018.