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

by R Resmawan

Submission date: 29-May-2022 08:31PM (UTC+0900)
Submission ID: 1846337336
File name: Revisi_Jefriyanto_Ibrahim.docx (99.54K)
Word count: 3463
Character count: 20587

Indonesian Journal of Applied Statistics

Journal homepage: https://jurnal.uns.ac.id/ijas/index
ISSN 2621–086X
Volume X No. X, Month Year

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

Jefriyanto Ibra[1], Resmawan[2], dan Isran K. Hasan[3]
[1]Jurusan Matematika Fakultas MIPA Universitas Negeri Gorontalo
[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, the next stage is to process the data by converting the original data to data that is ready to be mined. After the data is ready, the next step is divided into training and test data so that the classification algorithm can be applied. This study obtained the best four attributes, with KNN classification results of 86.67% and a Random Forest of 100%.

Keywords: length of student study; information gain; knn; 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 dinyatakan 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 analisis 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 KNN 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 yaitu algoritma yang digunaka untuk masalah klasifikasi dalam machine learning dan data mining [5].

Adapun penelitian menggunakan algoritma K-NN diantaranya dilakukan oleh Badu (2016) yang mengklasifikasikan dana desa dengan tingkat akurasi 78,95% dengan nilai K = 2. Pada penelitian Subrata, dkk (2017) 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 (2019) yang menerapkan Random Forest untuk mengukur tingkat keparahan penyakit pada daun apel. Dalam hal ini diperoleh hasil akurasi sebesar 75,3191%. Lebih lanjut Hanun dan Zailani (2020) 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. Namun sejauh ini, belum dapat dikonfirmasi secara pasti, mana yang terbaik diantara kedua metode tersebut.

Pada penelitian ini, Alogoritma K-NN dan Random Forest digunakan untuk mengklasifikasikan lama studi mahasiswa, namun digunakan Information Gain untuk menyeleksi fitur-fitur yang tidak memiliki pengaruh. Hal ini sesuai dengan penelitian yang dilakukan Bimantoro dan Uyun (2017) yang menggunakan Information Gain dalam menyeseleksi fitur citra untuk menilai kesesuain lahan pada tanaman cingkeh. 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 dengan tabel matriks.

2. Landasan Teori

2.1 Klasifikasi. Klasifikasi adalah suatu siklus dalam mendapatkan suatu model atau manfaat yang menggambarkan dan mengenali informasi atau gagasan yang diharapkan dapat dimanfaatkan dalam menilai kelas item yang labelnya tidak diketahui [6]. Adapun macam-macam algoritma yang banyak dipakai dalam klasifikasi secara luas yaitu, Decision/classification trees, Bayesian Classifiers/Nave Bayes Classifiers, Neural networks, K-nearest neighbor, metode rule based, dan Support Vector Machines (SVM). Berikut langkah-langkah dalam metode klasifikasi, yaitu [4]:

1. Pembelajaran (learning): pelatihan (training) pada fase ini algoritma kllasifikasi. 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 kotor, sehingganya pada tahap ini data akan di bersihkan dan diharapkan dapat mempermudah dalam proses klasifikasi. Preprocessing dibutuhkan dalam mengoptimalkan kemampuan algoritma klasifikasi [7]. Biasanya ada empat langkah dalam preprocessing untuk dokumen teks, yakni case folding, tokenizing, stopwords removal dan stemming [8].

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, untuk keadaan ini merupakan cerminan dari sifat sifat yang akan dimanfaatkan [9]. 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 [10]. Penentuan fitur dengan Information Gain dilakukan dengan 3 tahap [11], 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 dibawah batas.
3. Dataset diperbaiki dengan menghilangkan atribut tidak relevan.

2.4 K-Nearest Neighbor (KNN). KNN adalah metode yang menjalankan klasifikasi berlandaskan pada kedekatan suatu jarak data dengan data lainnya [3]. Pada KNN nilai K berarti data yang paling dekat dari data uji. Karena sederhana dalam melakukan proses klasifikasi pada kelompok data, metode KNN menjadi salah satu metode pengenalan pola yang umum dan sering dimanfaatkan. Cara kerja KNN 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 [12]:

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

Ada beberapa hal yang dapat mempengaruhi hasil KNN, diantaranya yaitu menentukan nilai K. Jika K terlalu kecil maka akan berdampak pada hasil perkiraan atau prediksi yang diperoleh bisa sensitif pada adanya noise. Sedangkan 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 [13]. 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 [14].

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

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 [15]. 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

3.1 Ruang Lingkup 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.

3.2 Metode Analisis. Sebelum masuk pada proses klasifikasi untuk kedua model hal pertama yang dilakukan yakni mempersiapkan data, information gain digunakan untuk menyeleksi atribut yang tidak berpengaruh, selanjutnya data di konversi sehingga dapat digunakan pada kedua metode klasifikasi, kemudian data dibagi menjadi 2 bagian yaitu data training dan data testing. Setelah didapatkan pembagain data lanjut pada metode KNN dan Random forest.

4. Hasil dan Pembahasan

4.1 Profile Data. 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.

4.2 Information Gain. Jumlah seleksi fitur menggunakan asumsi seperti pada penelitian sebelumnya yang menjelaskan bahwa jumlah seleksi fitur yang disarankan adalah $\log_2 n$ dengan nilai n adalah jumlah seluruh fitur [16]. Jadi pada penelitian ini akan digunakan sebanyak 5 atribut yang memiliki nilai pengaruh terbesar, karena $\log_2 15 = 3{,}9 \approx 4$ yang di tunjukan pada Tabel 2.

Tabel 2. Hasil 5 atribut yang berpengaruh

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

4.3 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 dimining. Adapun hasil konversi data yang diperoleh dari tahap preprocessing data yakni sebagai berikut:

1. Lama studi (Y) Pada lama studi tidak terjadi perubahan dikaranakan, 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.

4.4 Pembagian Data. Pembagian data bertujuan untuk memperoleh 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 dengan nilai 80% untuk data training dan 20% data testing.

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

Berdasarkan Tabel 3. dapat dipahami 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 KNN 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 KNN dengan menggunakan Confusion Matrix. Pada siklus penyusunan dengan teknik KNN, jumlah objek yang benar dan salah diklasifikasikan pada tiap-tiap kelompok bisa dilihat pada Tabel 5. Tanda (*) pada angka menunjukkan jumlah objek kelompok tertentu yang salah diklasifikasikan dengan menjalankan metode KNN.

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

Dengan akurasi sebagai berikut:

$$\text{Akurasi} = \frac{13}{15} \times 100 = 86{,}67\%$$

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. Sedangkan 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. Dengan tingkat akurasi sebesar 86,67% untuk K=1. Selanjutnya dalam memudahakan untuk menghitung nilai K lainnya akan 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, 53,33 untuk K=19, dan 60 untuk K=21.

4.6 Random Forest. Random forest merupakam metode klasifikasi yang tersusun dari sejumlah pohon keputusan diberbagai 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.

4.6.1 Proses Pelatihan dan Pembentukan Model Random Forest. Syntax Model Random Forest selanjutnya 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. Model random forest

Syntax 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 digunakn pada tiap iterasinya sebanyak 2 dengan perkiraan tingkat kesalahan OOB pada data training yang digunakan 0%. Dari gambar diatas juga dapat dilihat semua data berhasil dimodelkan dengan benar, sehingga class error yang dihasilkan sebesar 0%.

4.6.2 Pengujian Akurasi Model Random Forest. Setelah model terbentuk pada data training, tahap selanjutnya adalah menguji data uji 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.

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

Berdasarkan pada Tabel 4.6 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. Dan juga dari 5 orang yang lewat batas waktu, tidak terdapat kesalahan. Sehingga dapat dilihat tingkat akurasi dengan algoritma random forest sebesar 100%.

4.7 Perbandingan Tingkat Akurasi. Pengukuran tingkat akurasi baik pada algoritma KNN maupun Random forest diselesaikan dengan memastikan peluang kesalahan klasifikasi. Dalam siklus 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 terhadap algoritma K-NN dengan akurasi model sebesar 86,67%. Hal ini menunjukkan bahwa

algoritma random forest bekerja lebih baik dibandingkan dengan 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 memiliki tingkat akurasi yang lebih besar dari algoritma KNN yakni untuk random forest sebesar 100% dan KNN sebesar 86,67%. Sehingga dalam penelitian ini algoritma random forest bekerja lebih baik dibandingkan dengan algoritma K-NN dalam mengklasifikasikan Lama Studi Mahasiswa..

DAFTAR PUSTAKA

[1] Nasrullah, A.H. Penerapan Metode C4.5 untuk Klasifikasi Mahasiswa Berpotensi Drop Out. *ILKOM Jurnal Ilmiah*. 10(2): 244-250. 2018. doi: https://doi.org/10.33096/ilkom.v10i2.300.244-250

[2] PERMENDIKNAS. Keputusan Menteri Pendidikan Nasional Republik Indonesia No.232. Depdiknas. Jakarta. 2000.

[3] Prasetyo, E. *Data Mining Konsep dan Aplikasi Menggunakan Matlab*. Andi. Yogyakarta. 2012.

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

[5] Larose, D. T., & Larose, C. D. *Discovering knowledge in data: an introduction to data mining*. John Wiley & Sons. New York. 2014.

[6] Gorunescu, F. *Data Mining: Concepts, models and techniques*. Springer Science & Business Media. 2011.

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

[8] Crone, S. F., Lessmann, S., & Stahlbock, R. The impact of preprocessing on data mining: An evaluation of classifier sensitivity in direct marketing. *Eur. J. Oper. Res.*, 173 (3): 781–800, Sep. 2006. doi: 10.1016/j.ejor.2005.07.023.

[9] Budiman, A. S., & Parandani, X. A. 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.*, 9(1): 565–578. Apr. 2018. doi: 10.24176/simet.v9i1.2021.

[10] Shaltout, N. A., El-Hefnawi, M., Rafea, A., & Moustafa, A. 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*, 1(1), pp. 625-631. 2014.

[11] Maulana, M. R., & Al Karomi, M. A. Information Gain untuk Mengetahui Pengaruh Atribut Terhadap Klasifikasi Persetujuan Kredit. *Jurnal Litbang Kota Pekalongan*, 9: 113-123. 2015.

[12] Lestari, M. E. I. Penerapan algoritma klasifikasi Nearest Neighbor (K-NN) untuk mendeteksi penyakit jantung. *Faktor Exacta*, 7(4), 366-371. 2014

[13] Breiman, L. Randon Forests. Machine Learning 202. Pbworks. Com, 135. 1999.

[14] Biau, G. Analysis of a Random Forests Model. *Journal of Machine Learning Research*., 13(2012): 1063-1095. 2012.

[15] Adnyana, I. M. B. Prediksi Lama Studi Mahasiswa Dengan Metode Random Forest (Studi Kasus: STIKOM Bali). *Computer Science Research and Its Development Journal (CSRID)*., 8(3): 201-208. 2016.

[16] Roihan, A. *Seleksi fitur menggunakan Symmetrical Uncertainty pada prediksi cacat perangkat lunak*. Doctoral dissertation, Universitas Islam Negeri Maulana Malik Ibrahim, Malang. 2018.

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

ORIGINALITY REPORT

17%	14%	8%	4%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	journal.umpalangkaraya.ac.id Internet Source	2%
2	123dok.com Internet Source	1%
3	ejournal.uin-suka.ac.id Internet Source	1%
4	text-id.123dok.com Internet Source	1%
5	journals.upi-yai.ac.id Internet Source	1%
6	Jefry Antonius Karlia, Wawan Nurmansyah. "Application of C4.5 Algorithm for Late Payment Classification of Insurance Premiums", Tekinfo: Jurnal Ilmiah Teknik Industri dan Informasi, 2021 Publication	1%
7	jurnal.unimed.ac.id Internet Source	1%

8	repository.ub.ac.id Internet Source	1 %
9	etheses.uin-malang.ac.id Internet Source	1 %
10	eprints.uny.ac.id Internet Source	1 %
11	Sri Winiarti, Desy Widayanti, Ulaya Ahdiani, Taufiq Ismail. "Klasifikasi Jenis Buku Berdasarkan Cover dan Judul Buku Menggunakan Metode Support Vector Machine dan Cosine Similarity", Sainteks, 2022 Publication	1 %
12	Submitted to Universitas Islam Lamongan Student Paper	<1 %
13	journal.universitassuryadarma.ac.id Internet Source	<1 %
14	Submitted to Universitas Amikom Student Paper	<1 %
15	docplayer.info Internet Source	<1 %
16	Submitted to Universitas Brawijaya Student Paper	<1 %
17	fmforever-fm-fm.blogspot.com Internet Source	<1 %

18	Rina Novita, Supratman Zakir, Agus Nur Khomarudin, Efmi Maiyana, Hamimah Hasyim. "Use of the C4.5 Algorithm in Determining Scholarship Recipients", Journal of Physics: Conference Series, 2021 Publication	<1 %
19	e-journals.unmul.ac.id Internet Source	<1 %
20	Fadhila Tangguh Admojo, Ahsanawati. "Klasifikasi Aroma Alkohol Menggunakan Metode KNN", Indonesian Journal of Data and Science, 2020 Publication	<1 %
21	diskominfotiksan.pekanbaru.go.id Internet Source	<1 %
22	garuda.kemdikbud.go.id Internet Source	<1 %
23	journal.umg.ac.id Internet Source	<1 %
24	www.coursehero.com Internet Source	<1 %
25	www.scribd.com Internet Source	<1 %
26	Zhongfu Ye. "", Antennas and Wireless Propagation Letters, 12/2007 Publication	<1 %

27	id.123dok.com Internet Source	<1%
28	studentjournal.petra.ac.id Internet Source	<1%
29	Noivia Cyta Hari, Hanny Komalig, Yohanes Langi. "Analisis Survival Dalam Menentukan Faktor-faktor Yang Mempengaruhi Lama Studi Mahasiswa Matematika Di Jurusan Matematika FMIPA Universitas Sam Ratulangi Manado", d'CARTESIAN, 2018 Publication	<1%
30	jurnal.pekalongankota.go.id Internet Source	<1%
31	Bambang Hermanto, Azhari SN Azhari SN, Fajri Profesio Putra. "Analisis Kinerja Decision Tree C4.5 dalam Prediksi Potensi Pelunasan Kredit Calon Debitur", INOVTEK Polbeng - Seri Informatika, 2017 Publication	<1%
32	doku.pub Internet Source	<1%
33	jurnal.umt.ac.id Internet Source	<1%

Exclude quotes	On	Exclude matches	Off

Exclude bibliography On