

PERBANDINGAN KINERJA 6 ALGORITME KLASIFIKASI DATA MINING UNTUK PREDIKSI MASA STUDI MAHASISWA

Mariana Windarti, Agustinus Suradi

Program Studi Manajemen Informatika,
Fakultas Ilmu Komputer,
Universitas Widya Dharma Klaten
Email : mariana@unwidha.ac.id, simpati2000@mailcity.com

(Naskah masuk: 14 Januari 2019, diterima untuk diterbitkan: 28 Februari 2019)

ABSTRAK

Salah satu faktor yang memengaruhi kualitas sebuah perguruan tinggi adalah kinerja mahasiswa yang dapat diukur melalui lamanya masa studi. Perolehan pengetahuan dalam basis data (sejumlah data yang besar) biasa disebut dengan *data mining* atau penambangan data. Penelitian ini bertujuan untuk mengetahui kinerja keenam algoritme klasifikasi yang digunakan yaitu *Decision Tree (DT) C4.5*, *Bayesian Network (BN)*, *K-Nearest Neighbors (KNN)*, *Naïve Bayes (NB)*, *Neural Network (NN)* dan *SVM (Support Vector Machine)*. Kemudian menganalisa perbandingan kinerja keenam algoritme tersebut. Hasil penelitian menunjukkan bahwa *Bayesian Network* memiliki kinerja paling baik dengan nilai akurasi sebesar 80.615%, nilai presisi dan *recall* sebesar 0.785 dan 0.806, sedang untuk nilai AUC (*Area Under Curve*) termasuk dalam kategori baik yaitu 0.837. Sedangkan *DT C4.5* memiliki kinerja terendah dengan nilai akurasi sebesar 76.615%.

Kata Kunci: *data mining*, klasifikasi, masa studi, perbandingan

ABSTRACT

One of the factors that affect the quality of a college is student performance that can be measured through the length of the study period. The acquisition of knowledge in a database (large amounts of data) is commonly referred to as *data mining*. This study aims to determine the algorithms performance of six classification that used, namely *Decision Tree (DT) C4.5*, *Bayesian Network (BN)*, *K-Nearest Neighbors (KNN)*, *Naïve Bayes (NB)*, *Neural Network (NN)* and *SVM (Support Vector Machine)*. Then analyze the comparison of the performance of the six algorithms. The results showed that *Bayesian Network* had the best performance with an accuracy value of 80,615%, precision and recall values of 0.785 and 0.806, for the AUC (*Area Under Curve*) value was included in the good category, 0.837. Whereas *DT C4.5* has the lowest performance with an accuracy value of 76,615%.

Keywords: classification, comparison, data mining, study period

PENDAHULUAN

Penambahan data dibagi menjadi beberapa kelompok berdasarkan tugas yang dapat dilakukan yaitu model deskripsi, estimasi, prediksi, klasifikasi, pengklusteran dan asosiasi. Terdapat beberapa metode atau teknik klasifikasi seperti *Decision Tree* (pohon keputusan), *Naive Bayes Classifier*, Jaringan Syarat Tiruan, Analisis Statistik, *Rough Sets*, *K-Nearest Neighbor*, SVM, dan lain-lain.

Data mining dapat diimplementasikan pada beberapa area seperti *Customer Relationship Management (CRM)*, deteksi penipuan, evaluasi kredit, evaluasi risiko, perawatan medis dan deteksi penyakit. Penelitian ini dilakukan dengan menganalisa perbandingan algoritme seperti *Naive Bayes*, *Decision Tree*, *K-Nearest Neighbour*, *Neural Network* dan SVM dengan memerhatikan kriteria seperti keakuratan, kecepatan dan ketangguhan dalam penggunaan beberapa jenis data (Dogan & Tanrikulu, 2013). Dibidang kesehatan penerapan data mining dilakukan untuk memprediksi penyakit Diabetes. Teknik klasifikasi yang digunakan yaitu *Random Forest*, NB dan SVM (*Support Vector Machine*). *Random Forest* memiliki kinerja paling baik berdasarkan presisi dan waktu yang diperlukan, sedangkan untuk recall, SVM memiliki kinerja paling baik dari pada *Random Forest* dan NB (Tate, Rajpurohit, Pawar, & Gavhane, 2016).

Penelitian tahun 2017 dilakukan menggunakan beberapa algoritme klasifikasi yaitu *Naïve Bayes*, *Nearest Neighbour* dan *Decision Tree* J48 untuk pengambilan keputusan pemilihan pola pakaian. Kemudian membandingkan kinerja setiap algoritma tersebut berdasarkan tingkat akurasi. Hasil penelitian menunjukkan bahwa *Decision Tree* memiliki akurasi paling tinggi yaitu sebesar 75.6% (Sartika & Sensuse, 2017).

Penelitian tahun 2015 untuk prediksi masa studi mahasiswa menggunakan *Bayesian Network* menghasilkan nilai akurasi sebesar 84,28%. Proses prediksi dilakukan berdasarkan variabel nilai Indeks Prestasi (IP) semester 1 & 2, nilai ujian nasional (UN), jurusan sekolah, jalur masuk perguruan tinggi dan hasil tes masuk (Windarti, Santoso, & Ernawati, 2015). Tahun 2017 dilakukan penelitian untuk menganalisis kinerja algoritme KNN, *Random Forest* dan *Naïve Bayesian*

dalam memprediksi Diabetes Melitus (DM). Hasil uji *10-fold cross validation* diperoleh akurasi tertinggi pada algoritme Naïve Bayesian yaitu 75,65%, sedang KNN 75,53 dan random forest 73,69%. Untuk *sensitivity* terbaik dihasilkan Naïve Bayes sebesar 62,69% dan *specificity* terbaik oleh *Random Forest* sebesar 86,4% (Setyawan & Suradi, 2017).

Terdapat beberapa *tools data mining* antara lain Rapid Miner, Orange, KNIME, Weka, keel dan R. WEKA sudah berbasis GUI (*Graphical User Interface*) sehingga meminimalkan penggunaan *coding* yang dapat memudahkan pengguna sistem (Rangra, 2014). Pada penelitian ini menggunakan tools WEKA. Data yang diolah pada WEKA memiliki format ARFF maupun CSV. Metode pengujian yang digunakan pada WEKA adalah *cross validation*.

Penelitian oleh Sakshi & Kare dilakukan untuk melakukan analisis perbandingan terhadap beberapa algoritme seperti *Decision Tree*, *Bayesian Network*, KNN dan *Artificial Neural Network* (ANN). Dari keempat algoritme tersebut *Decision Tree* memiliki akurasi tertinggi dari pada algoritme lainnya yaitu sebesar 96,3218%, kemudian disusul ANN 94,4828%, KNN 92,4138% dan terakhir BN 90,1149% (Sakshi & Kare, 2015). Kemudian Studi perbandingan untuk klasifikasi kinerja akademik mahasiswa ke dalam 4 kelas yaitu kelas A, B, C & F. Algoritme yang digunakan adalah DT, NB, *Naïve Bayes Tree*, *K-Nearest Neighbors* dan *Bayesian Network*. Metode *Bootstrap* digunakan untuk meningkatkan akurasi setiap algoritme. Akurasi tertinggi DT 95.49%, kemudian KNN 93.5% dan BN 87.98% (Taruna & Pandey, 2014).

Penelitian tahun 2018 untuk mengklasifikasi masa studi mahasiswa menggunakan 100 *record* dataset dan algoritme yang digunakan adalah BN dan NB. Metode pengujian yang digunakan *percentage split*. Hasil penelitian menunjukkan BN memiliki kinerja lebih baik dari pada NB pada *percentage split* 80 yaitu sebesar 75% (Windarti, 2018). Pada tahun yang sama terdapat penelitian untuk menganalisis perbandingan algoritme klasifikasi pada dataset blogger dengan Rapid Miner. Algoritme yang digunakan dalam mengklasifikasi dataset blogger yaitu *Decision Tree*, NB, KNN, ID3, dan CHAID. Akurasi tertinggi

adalah KNN sebesar 85%, sedang untuk nilai AUC algoritma CHAID memiliki nilai tertinggi sebesar 0.758 dan hasil uji t-test menyatakan ID3, CHAID dan NB merupakan algoritma dengan performa terbaik pada dataset blogger (Ardiyansyah, Rahayuningsih, & Maulana, 2018).

Algoritme yang memiliki kinerja paling baik antara satu penelitian atau kasus dengan kasus lain dapat berbeda-beda. Menurut Alghobiri, 2018 terdapat bermacam-macam algoritme klasifikasi yang digunakan pada beberapa dataset. Algoritme yang digunakan adalah NB, C4.5 dan SVM. Hasil penelitian Alghobiri menunjukkan bahwa SVM memiliki kinerja paling baik dari pada NB dan C4.5. Kinerja algoritme dapat diukur berdasarkan nilai presisi, akurasi, *F-measure*, Kappa, MSE (*Mean Absolute Error*), *Relative Absolute Error*, ROC, dan lain-lain. Pengukuran kinerja yang digunakan adalah akurasi, presisi dan *F-measure*. Peneliti memilih menggunakan parameter akurasi dan presisi (Alghobiri, 2018).

Berdasarkan uraian di atas terdapat banyak algoritme dan tidak semua algoritme memiliki kinerja yang baik, karena setiap algoritme memiliki kelebihan dan kekurangan. Sehingga peneliti bermaksud menganalisis perbandingan kinerja beberapa algoritme klasifikasi yaitu *DT C4.5*, *Bayesian Network*, *K-Nearest Neighbors*, *Naïve Bayes*, *NN* dan SVM. Peneliti memilih enam algoritme ini karena secara umum memiliki kinerja baik pada beberapa kasus atau dataset. Penelitian ini menggunakan WEKA dengan metode pengujian *cross validation*. Tujuan penelitian adalah untuk menganalisis perbandingan kinerja keenam algoritme diatas. Pengukuran kinerja menggunakan parameter akurasi, *precision*, *recall* dan nilai AUC (*Area Under Curve*). Penelitian ini juga diharapkan mampu memberikan manfaat khususnya bagi para peneliti yaitu untuk dijadikan bahan referensi dan studi perbandingan dalam melakukan penelitian selanjutnya.

METODE PENELITIAN

1. Pengumpulan Data

Pengumpulan data dilakukan langsung di lapangan yaitu data lulusan program studi Teknik Informatika UNWIDHA Klaten dengan tahun lulus 2012-

2017 yang diperoleh dari Bagian Akademik. Sebagian data belum tersimpan di komputer sehingga pengumpulan beberapa data dilakukan secara manual. Variabel yang digunakan antara lain IP (Indeks Prestasi) semester 1, IP semester 2, IP semester 3, penjurusan sekolah, jalur masuk perguruan tinggi dan masa studi yang diperoleh dari tanggal masuk dan tahun lulus. Peneliti memilih beberapa variabel diatas karena peneliti mendapatkan data dari sisi akademis yaitu Indeks Prestasi dan latar belakang sebelum masuk perguruan tinggi yaitu SMA/SMK. Sebenarnya ada beberapa faktor lain selain akademis yang mempengaruhi masa studi. Dalam hal ini peneliti tidak dapat memperoleh data terkait tersebut. Sehingga dapat digunakan sebagai saran untuk penelitian berikutnya.

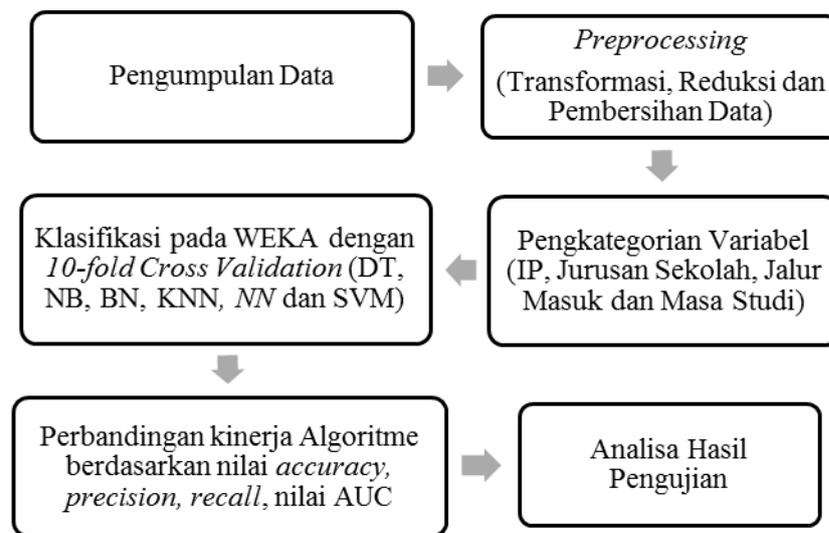
2. Alat Dan Bahan

Bahan yang digunakan pada penelitian ini antara lain:

1. Ms. Excel untuk mengolah data awal atau data mentah
2. *Tool* WEKA dengan metode pengujian *Cross Validation* menggunakan *10-fold*.
3. Pengklasifikasi *tree classifier* J48 yang merupakan implementasi algoritme C4.5, *Bayes Network*, IBk (*K-Nearest Neighbors*), *Naïve Bayes*, *Multilayer Perceptron (Neural Network)* dan SMO (SVM).

3. Tahapan Penelitian

Ada beberapa tahapan yang dilakukan dalam penelitian ini, yang dapat dilihat pada Gambar 1 berikut ini.



Gambar 1. Tahapan Penelitian

Tahapan tersebut dapat dirincikan sebagai berikut:

1. Pengumpulan data yang disimpan dalam Microsoft Excel
2. Kemudian melakukan tahap persiapan data dengan *preprocessing*. Setelah itu data telah siap untuk diproses yaitu untuk pelatihan dan pengujian,
3. Kemudian mengelompokkan atau mengkategorikan setiap nilai variabel ke dalam kelas yang sudah ditentukan.
4. Klasifikasi diimplementasikan menggunakan WEKA dengan metode pengujian *10-fold cross validation* menggunakan DT, BN, NB, KNN, *Neural Network* dan SVM.
5. Tahapan terakhir adalah membuat perbandingan kinerja masing-masing algoritme lalu menganalisa hasil pengujian.

3.1 Decision Tree C4.5

Decision Tree atau pohon keputusan merupakan metode klasifikasi dan prediksi yang sangat kuat dan terkenal. Fakta yang ada akan diubah menjadi pohon keputusan yang merepresentasikan aturan. Pohon keputusan juga dapat digunakan untuk mengeksplorasi data, menemukan hubungan tersembunyi antara sejumlah variabel input dengan target (Kusrini & Luthfi, 2009). Terdapat

beberapa algoritme dalam pembentukan pohon keputusan antara lain ID3, CART dan C4.5. C4.5 merupakan pengembangan dari algoritme ID3. Secara umum algoritme C4.5 untuk membangun pohon keputusan adalah sebagai berikut:

- 1) Pilih atribut sebagai akar
- 2) Buat cabang untuk tiap nilai
- 3) Bagi kasus dalam cabang
- 4) Ulangi proses setiap cabang sampai semua kasus memiliki kelas yang sama

Untuk memilih atribut akar didasarkan pada gain ratio tertinggi dari atribut yang ada. Proses perhitungan *gain* dilakukan dengan persamaan (1) berikut ini.

$$Gain(S,A) = Entropy(S) - \sum_{i=1}^n \frac{S_i}{S} * Entropy(S_i) \quad (1)$$

Keterangan:

S : himpunan kasus

A : atribut

n : jumlah partisi atribut A

|Si| : jumlah kasus pada partisi ke-i

|S| : jumlah kasus dalam S

Sedangkan proses perhitungan nilai entropi dilakukan dengan persamaan (2) berikut ini.

$$Entropy(S) = \sum_{i=1}^n - p_i * \log_2 p_i \quad (2)$$

Keterangan:

S : himpunan kasus

A : fitur

N : jumlah partisi S

pi : proporsi dari Si terhadap S

3.2 Bayesian Network (BN) & Naïve Bayes (NB)

Bayesian Network merupakan suatu metode pemodelan data berbasis probabilitas yang merepresentasikan suatu himpunan variabel dan atribut yang saling berkorespondensi atau berhubungan melalui DAG (*Directed Acyclic*

Graph). *Bayesian Network* memiliki dua tugas yaitu pembelajaran melalui DAG dan struktur dari *Bayesian Network* berupa jaringan (Friedman, Geiger, & Goldszmidt, 1997).

Bayesian Network didasarkan pada Teorema Bayes (*Naïve Bayes*) yaitu *conditional probability* (peluang bersyarat). Pada *Bayesian Network* antar variabel atau atribut bisa saling *dependent* atau berhubungan, sedangkan *Naive Bayes* mengabaikan hubungan antar atribut atau variabel. Proses perhitungan untuk mencari peluang pada Teorema Bayes dilakukan dengan persamaan (3) berikut ini.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad (3)$$

Keterangan:

$P(A|B)$ = disebut juga *posterior probability*, yaitu peluang A terjadi setelah B terjadi.

$P(B \cap A)$ = peluang B dan A terjadi bersamaan

$P(B|A)$ = disebut juga *likelihood*, yaitu peluang B terjadi setelah A terjadi.

$P(A)$ = disebut juga *prior*, yaitu peluang kejadian A

$P(B)$ = peluang kejadian B

Bayesian Network digambarkan seperti graf yang terdiri dari simpul (*node*) dan busur (*arc*). *Node* menunjukkan variabel atau atribut beserta nilai probabilitasnya dan busur menunjukkan hubungan antar simpul. Langkah-langkah untuk menerapkan *Bayesian Network* yaitu (Meigaran, Setiawan, & Riza, 2012) :

1. Membangun struktur *Bayesian Network*
2. Menentukan parameter
3. Membuat *Conditional Probability Table* (CPT)
4. Membuat *Joint Probability Distribution* (JPD), untuk menghitung *Joint Probability Distribution* adalah mengalikan nilai *Conditional Probability* dengan *Prior Probability*.
5. Menghitung *Posterior Probabilistik*, didapatkan dari hasil JPD yang telah diperoleh.

6. Inferensi Probabilistik yaitu penelusuran yang dilakukan berdasarkan variabel input yang diberikan pengguna sehingga menghasilkan suatu nilai probabilitas.

3.3 K-Nearest Neighbors (K-NN)

Algoritme *K-Nearest Neighbors* merupakan pendekatan untuk mencari kasus dengan menghitung kedekatan antara kasus yang baru dengan kasus lama berdasarkan pencocokan bobot dari sejumlah fitur yang ada. Kedekatan biasa berada pada nilai 0 s.d 1, nilai 0 artinya kedua kasus mutlak tidak mirip, sedang nilai 1 kasus mirip mutlak (Kusrini & Luthfi, 2009). Proses perhitungan kedekatan antara dua kasus dilakukan dengan persamaan (4) berikut ini.

$$\text{similarity}(T, S) = \frac{\sum_{i=1}^n f(T_i, S_i) \cdot W_i}{W_i} \quad (4)$$

Keterangan:

T : kasus baru

S : kasus yang ada dalam penyimpanan

n : jumlah atribut dalam setiap kasus

i : atribut individu antara 1 s.d. n

f : fungsi similarity atribut i antara kasus T dan kasus S

w : bobot yang diberikan pada atribut ke-i

3.4 Artificial Neural Network (ANN)

Artificial Neural Network atau Jaringan Syaraf Tiruan (JST) merupakan suatu konsep rekayasa pengetahuan dalam bidang kecerdasan buatan yang didesain dengan mengadopsi sistem saraf manusia, yang pemrosesan utamanya ada di otak. JST dianggap sebagai model yang membutuhkan proses pelatihan agar dapat melakukan prediksi kelas suatu data uji baru yang ditemukan.

JST ditentukan oleh 3 hal yaitu (Siang, 2005) :

- a. Pola hubungan antar neuron (arsitektur jaringan)
- b. Metode untuk menentukan bobot penghubung (metode training/learning/ algoritma).
- c. Fungsi Aktivasi

3.5 Support Vector Machine (SVM)

SVM merupakan algoritme yang bekerja menggunakan pemetaan nonlinier untuk mengubah data pelatihan asli ke dimensi yang lebih tinggi. Dimensi baru akan mencari *hyperplane* untuk memisahkan secara linier dengan pemetaan *nonlinear* ke dimensi yang cukup tinggi. Persoalan nonlinear merupakan persoalan dengan data yang tidak dapat dipisahkan secara linear yaitu tidak ada sebuah garis yang dapat dibuat sebagai pemisah antar kelas data. SVM menemukan *hyperplane* menggunakan *support vector* dan *margin* (Widodo, Handayanto, & Herlawati, 2013). Waktu pelatihan SVM kebanyakan lambat, tetapi metode ini sangat akurat karena kemampuannya menangani model *nonlinear* yang kompleks.

4. PENGOLAHAN AWAL DATA

Data awal dibuat dalam format .xls atau .xlsx (Microsoft Excel). Pada data mentah sering ditemukan nilai yang hilang, sampling data tidak cukup bagus, data tidak lengkap dan sebagainya. Maka perlu dilakukan penyiapan data terlebih dahulu sebelum data siap diolah. Tahapan penyiapan data dalam *data mining* disebut juga tahapan *preprocessing*, yang meliputi proses *cleaning* yaitu membuang duplikasi data dan memperbaiki kesalahan pada data. Selain itu juga dilakukan transformasi data yaitu mengubah nilai variabel ke format yang sesuai, mereduksi data yaitu data yang hanya mempunyai atribut yang berhubungan saja yang akan digunakan.

5. PENGUKURAN KINERJA KLASIFIKASI

Teknik yang digunakan untuk mengestimasi kinerja berdasarkan model pelatihan yang telah terbentuk adalah dengan *k-fold cross validation* (Jiawei & Kamber, 2001). Metode ini akan membagi data pelatihan dan data pengujian sebanyak *k* bagian data. Data yang digunakan dibagi secara acak ke dalam *k* subset dengan ukuran yang sama. Proses pelatihan dan pengujian akan dilakukan

sebanyak k kali secara berulang-ulang. Pada penelitian ini menggunakan *fold* sebanyak 10 karena merupakan jumlah *fold* standar yang terdapat pada WEKA.

Pengukuran kinerja klasifikasi dilakukan dengan mengevaluasi hasil pengujian menggunakan matriks konfusi atau *confusion matrix*. *Confusion Matrix* merupakan metode untuk mengevaluasi model klasifikasi untuk memperkirakan objek yang benar atau salah (Gorunescu, 2011). Pada Tabel 1 berikut diberikan matriks konfusi dua kelas:

Tabel 1. Tabel *Confusion Matrix* Dua Kelas

<i>Classification</i>	<i>Predicted Class</i>	
	<i>Class=Yes</i>	<i>Class=No</i>
Class=Yes	a (True Positive)	b (False Negative)
Class=No	c (False Positive)	d (True Negative)

Pada tabel di atas, *true positive* (TP) adalah jumlah *record* positif yang diklasifikasikan sebagai positif, *false positive* (FP) adalah jumlah *record* negatif yang diklasifikasikan sebagai positif, *false negatives* (FN) adalah jumlah *record* positif yang diklasifikasikan sebagai negatif, *true negatives* (TN) adalah jumlah *record* negatif yang diklasifikasikan sebagai negatif. Semakin tinggi nilai TP dan TN semakin baik pula tingkat klasifikasi dari akurasi, presisi dan *recall*. Akurasi merupakan tingkat kedekatan antara nilai prediksi dengan nilai sebenarnya. Presisi menunjukkan tingkat ketepatan atau ketelitian dalam pengklasifikasian. Sedangkan *recall* berfungsi untuk mengukur proporsi positif aktual yang benar diidentifikasi. Selain itu parameter pengukuran kinerja klasifikasi menggunakan nilai AUC (*Area Under Curve*), yaitu luas daerah dibawah kurva. Luas AUC berada antara nilai 0 – 1. Jadi pengukuran kinerja menggunakan nilai akurasi, presisi, *recall* dan AUC karena parameter ini sering digunakan pada beberapa penelitian untuk mengukur kinerja algoritme yang digunakan.

HASIL DAN PEMBAHASAN

1. Analisis Data

Data yang digunakan pada proses pelatihan dan pengujian telah melewati tahapan *preprocessing* sehingga data telah siap untuk diolah. Data yang digunakan sebagai proses pelatihan dan pengujian terdiri dari atribut IP (Indeks Prestasi) semester 1, IP semester 2, IP semester 3, jurusan sekolah, jalur penerimaan masuk perguruan tinggi dan masa studi. Berikut pengelompokan nilai setiap variabel kedalam kelas-kelas tertentu :

- 1) Indeks Prestasi : <2 , $\geq 2 - 2.5$, $\geq 2.5 - 3$, $\geq 3 - 3.5$ dan ≥ 3.5 .
- 2) Jurusan Sekolah : SMA-IPA, SMA-IPS, SMK-IPA dan SMK-IPS, Bahasa
- 3) Jalur masuk : raport, test & undi
- 4) Masa studi : < 4 tahun, $\geq 4 - < 5$ tahun, $\geq 5 - < 6$ tahun dan ≥ 6 tahun.

Pengelompokan diatas dilakukan berdasarkan data yang ada yaitu nilai setiap variabel. Pada Tabel 2 di bawah ini terdapat sampel data yang digunakan dalam proses pelatihan dan pengujian setelah dilakukan tahap *preprocessing*.

Tabel 2. Sampel Data Pelatihan dan Pengujian

<i>ip1</i>	<i>ip2</i>	<i>ip3</i>	<i>jur_sek</i>	<i>jalur_msk</i>	<i>ms_studi</i>
≥ 3.5	$\geq 3 - 3.5$	$\geq 3 - 3.5$	SMK_IPA	und	<4
$\geq 2.5 - 3$	$\geq 2 - 2.5$	$\geq 2 - 2.5$	SMK_IPA	raport	≥ 6
$\geq 3 - 3.5$	$\geq 3 - 3.5$	$\geq 3 - 3.5$	SMA_IPS	test	$\geq 4 - 5$
$\geq 2.5 - 3$	$\geq 3 - 3.5$	$\geq 3 - 3.5$	SMA_IPA	raport	$\geq 4 - 5$
$\geq 2.5 - 3$	$\geq 3 - 3.5$	$\geq 3 - 3.5$	SMA_IPS	test	$\geq 4 - 5$
$\geq 3 - 3.5$	$\geq 3 - 3.5$	$\geq 3 - 3.5$	SMK_IPA	test	$\geq 4 - 5$
$\geq 3 - 3.5$	$\geq 2.5 - 3$	$\geq 2.5 - 3$	SMA_IPS	test	$\geq 4 - 5$
$\geq 3 - 3.5$	$\geq 2.5 - 3$	$\geq 2 - 2.5$	SMK_IPA	test	$\geq 4 - 5$
$\geq 3 - 3.5$	$\geq 2.5 - 3$	$\geq 2.5 - 3$	SMK_IPA	raport	$\geq 4 - 5$
$\geq 3 - 3.5$	$\geq 2.5 - 3$	$\geq 2.5 - 3$	SMK_IPS	raport	<4

2. Analisis Hasil Klasifikasi

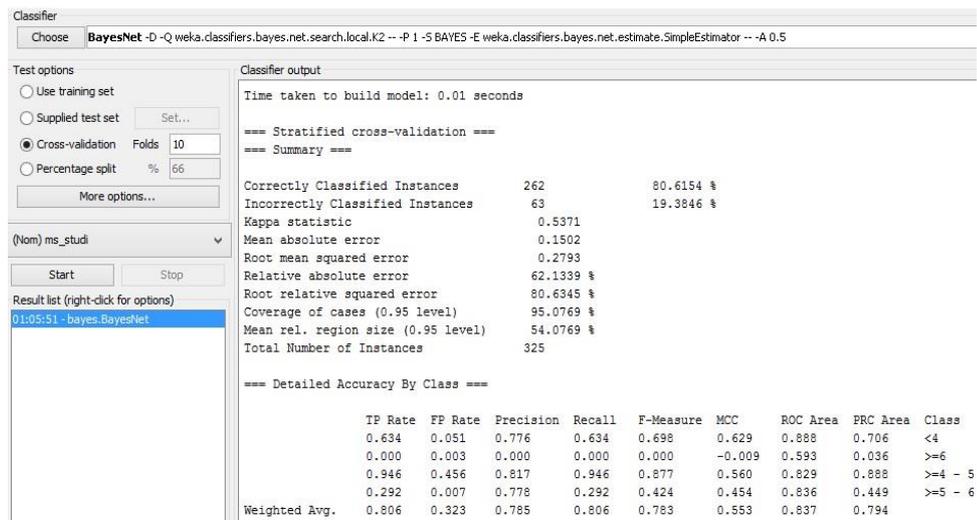
Pengujian dilakukan dengan menggunakan *confusion matrix* yaitu berdasarkan parameter nilai akurasi, presisi, *recall* dan nilai AUC. Metode pengujian yang digunakan pada WEKA adalah *10-fold cross validation*. *Dataset* yang digunakan baik untuk proses pelatihan dan pengujian sebanyak 325 record

data yang berasal dari data lulusan Program Studi Teknik Informatika UNWIDHA Klaten. Data yang berasal dari Ms.Excel (.xls atau .xlsx) diubah terlebih dahulu menjadi format .arff, karena WEKA mampu mengolah data yang memiliki format .arff atau .csv. Setelah mengimplementasikan keenam algoritme klasifikasi di atas pada *tool* WEKA menggunakan metode *10-fold cross validation*, diperoleh hasil pengukuran kinerja masing-masing pengklasifikasi pada Tabel 3 berikut ini.

Tabel 3. Hasil Perbandingan Kinerja Klasifikasi pada WEKA

<i>Algoritme (Pengklasifikasi)</i>	<i>Akurasi (%)</i>	<i>Presisi</i>	<i>Recall</i>	<i>Nilai AUC</i>
<i>Bayesian Network</i>	80.615	0.785	0.806	0.837
<i>Naïve Bayes</i>	80	0.778	0.8	0.836
<i>Neural Network (Multilayer Perceptron)</i>	80.308	0.779	0.803	0.797
<i>K-NN (IBk)</i>	79.077	0.758	0.791	0.79
<i>Decision Tree C4.5 (J.48)</i>	76.615	0.724	0.766	0.709
<i>SVM (SMO)</i>	79.077	0.756	0.791	0.765

Berdasarkan hasil perbandingan kinerja pada Tabel 3 di atas, *Bayesian Network* (BN) memiliki kinerja yang paling baik atau berada pada posisi teratas. Hal ini ditunjukkan dengan nilai akurasi BN yaitu 80.615%. Selain itu juga BN memiliki nilai presisi dan *recall* yang paling tinggi yaitu 0.785 & 0.806. Nilai AUC pada algoritme BN juga paling tinggi yaitu 0.837 dan termasuk dalam kategori pengklasifikasi yang baik (*good*). Posisi berikutnya adalah NB dengan nilai akurasi sebesar 80%, nilai presisi dan *recall* sebesar 0.778 dan 0.8. Sedangkan untuk posisi terendah atau pengklasifikasi yang memiliki kinerja kurang baik adalah *DT* C4.5 dengan nilai akurasi 76.615%. Salah satu hasil pengujian pada WEKA yang diimplemetasikan menggunakan *Bayesian Network* dapat dilihat pada Gambar 2 berikut ini.

Gambar 2. Hasil Pengujian *Bayesian Network* Menggunakan WEKA

KESIMPULAN DAN SARAN

Kesimpulan dari penelitian yang telah dilakukan adalah algoritme klasifikasi BN, NB, NN, K-NN, DT C4.5 dan SVM mampu melakukan prediksi masa studi mahasiswa UNWIDHA Klaten ke dalam empat kelas yaitu masa studi < 4 tahun, $\geq 4 - < 5$ tahun, $\geq 5 - < 6$ tahun dan ≥ 6 tahun. Algoritme *Bayesian Network* memiliki kinerja paling baik dibanding lainnya baik berdasarkan nilai akurasi, presisi, *recall* dan nilai AUC. Nilai akurasi BN sebesar 80.615%, nilai presisi dan *recall* sebesar 0.785 & 0.806, sedang nilai AUC sebesar 0.837. Sedangkan algoritme dengan kinerja paling buruk atau berada di posisi terendah adalah *Decision Tree C4.5*.

Adapun saran yang diusulkan penulis untuk penelitian selanjutnya adalah:

- 1) Penelitian ini hanya sebatas pada analisis statistik atau berdasarkan data yang ada di lapangan, maka diperlukan faktor eksternal yang dapat dijadikan variabel tambahan seperti kondisi psikologis, keluarga, status mahasiswa apakah bekerja atau tidak, dan lain-lain.
- 2) Perlu menambah *dataset* sehingga dapat diperoleh akurasi yang lebih baik.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada Program Studi Teknik Informatika UNWIDHA Klaten yang telah membantu dalam hal pengumpulan data. Kemudian ucapan terima kasih kepada Kementerian Riset, Teknologi dan Pendidikan Tinggi (Kemenristekdikti), Karena penulisan paper ini dibuat sebagai luaran wajib dalam skema Penelitian Dosen Pemula (PDP) tahun anggaran 2018.

DAFTAR PUSTAKA

- Alghobiri, M. (2018). A Comparative Analysis of Classification Algorithms on Diverse Datasets. *Engineering, Technology & Applied Science Research*, 8(2), 2790–2795.
- Ardiyansyah, Rahayuningsih, P. A., & Maulana, R. (2018). Analisis Perbandingan Algoritma Klasifikasi Data Mining Untuk Dataset Blogger Dengan Rapid Miner. *Jurnal Khatulistiwa Informatika*, VI(1), 20–28.
- Dogan, N., & Tanrikulu, Z. (2013). A comparative analysis of classification algorithms in data mining for accuracy, speed and robustness. *Information Technology and Management*, 14(2), 105–124.
- Friedman, N., Geiger, D., & Goldszmidt, M. (1997). Bayesian Network Classifiers. *Machine Learning*, 29, 131–163.
- Gorunescu, F. (2011). *Data Mining Concept, Models and Techniques*. Berlin Heidelberg: Springer.
- Jiawei, H., & Kamber, M. (2001). *Data mining: concepts and techniques*. San Francisco, CA, itd: Morgan Kaufmann.
- Kusrini, & Luthfi, E. T. (2009). *Algoritma Data Mining*. Yogyakarta: C.V ANDI OFFSET (Penerbit ANDI).
- Meigaran, I., Setiawan, W., & Riza, L. S. (2012). PENGGUNAAN METODE BAYESIAN NETWORK DALAM SISTEM PAKAR UNTUK DIAGNOSIS PENYAKIT LEUKEMIA. *Uma Ética Para Quantos?*,

XXXIII(2), 81–87.

- Rangra, K. (2014). Comparative Study of Data Mining Tools. *International Journal of Advanced Research in Computer Science and Software Engineering*, 4(6), 216–223.
- Sakshi, & Kare, S. (2015). A Comparative Analysis of Classification Techniques on Categorical Data in Data Mining. *International Journal on Recent and Innovation Trends In Computing and Communication (IJRITCC)*, 3(8), 5142–5147.
- Sartika, D., & Sensuse, D. I. (2017). Perbandingan Algoritma Klasifikasi Naive Bayes , Nearest Neighbour , dan Decision Tree pada Studi Kasus Pengambilan Keputusan Pemilihan Pola Pakaian. *Jurnal Teknik Informatika Dan Sistem Informasi (JATISI)*, 1(2), 151–161.
- Setyawan, D., & Suradi, A. (2017). Implementasi web service dan analisis kinerja algoritma klasifikasi data mining untuk memprediksi diabetes mellitus. *Jurnal SIMETRIS*, 8(2), 701–710.
- Siang, J. J. (2005). *Jaringan Syaraf Tiruan dan Pemrogramannya Menggunakan Matlab*. Yogyakarta: Penerbit Andi.
- Taruna, S., & Pandey, M. (2014). An empirical analysis of classification techniques for predicting academic performance. In *International Advance Computing Conference (IACC)* (pp. 523–528).
- Tate, A., Rajpurohit, B., Pawar, J., & Gavhane, U. (2016). Comparative Analysis of Classification Algorithms Used for Disease Prediction in Data Mining. *International Journal of Engineering and Techniques*, 2(6), 204–211.
- Widodo, P. P., Handayanto, R. T., & Herlawati. (2013). *Penerapan Data Mining Dengan Matlab*. Rekayasa Sains.
- Windarti, M. (2018). PERBANDINGAN KINERJA ALGORITMA NAIVE BAYES DAN BAYESIAN NETWORK DALAM KLASIFIKASI

MASA STUDI MAHASISWA, (September).

Windarti, M., Santoso, A. J., & Ernawati. (2015). Analisis Prediksi Masa Studi Mahasiswa Menggunakan Algoritma Bayesian Network. In *Seminar Nasional Teknologi Terapan (SNTT) SV UGM*.