PERBANDINGAN MINAT SISWA SMU PADA METODE KLASIFIKASI MENGGUNAKAN 5 ALGORITMA

Agus Purwanto[1], Eko Agus Darmadi[2]

[1,2] Politeknik Trimitra Karya Mandiri
Jl. By Pass Jomin – Blok Semper – Jomin Barat – Kotabaru – Cikampek - Karawang
E-mail : aguspurwanto44@yahoo.com[1], ekoagus.darmadi@gmail.com

ABSTRAK

Mendekati akhir tahun banyak siswa SMU khususnya yang sudah dikelas 13 atau akhir kesulitan untuk menentukan pilihan yang akan diambil selepas selesai sekolah. Hal ini membuat beberapa pihak mencoba memprediksi minat mana yang akan diambil oleh siswa-siswa tersebut guna kepentingan masing-masing pihak. Pada paper ini bertujuan untuk menangkap animo dari minat siswa-siswa tersebut dan membandingkannya dalam 5 algoritma yang akan diterapkan pada metode klasifikasi guna mengetahui akurasi mana yang terbaik dari kelima algoritma tersebut. Adapun algoritma yang akan diuji untuk dataset yang ada adalah algoritma K-Nearest Neighbor, Naïve Bayes, Pohon Keputusan (C4.5), Rule Induction dan Deep Learning. Hasil dari kelima algoritma tersebut akan diuji dalam Cross Validation T-Test guna mengetahui model mana yang lebih akurat.

.

Kata kunci : Algoritma, Metode Klasifikasi, Akurasi

1. PENDAHULUAN

Berakhirnya masa sekolah membuat siswa-siwa sekolah yang duduk dibangku SMU akan memikirkan masa depan yang akan dijalani oleh mereka, banyak dari siswa-siswa SMU tersebut mempunyai bermacam-macam keinginan ketika selesai sekolah.Hal ini membuat beberapa pihak yang mempunyai kepentingan dengan hal tersebut mencoba menangkap dan mencari tahu animo terbesar dari para siswa-siswa tersebut selepas mereka selesai Pendidikan SMU, dimana data tersebut akan berguna bagi beberapa pihak untuk menentukan langkah yang akan mereka ambil

Menyikapi hal tersebut penulis mencoba memasukan minat siswa-siswa tersebut yang didapat dari survey/poling yang dilakukan kedalam metode klasifikasi data mining. Dengan menggunakan 5 algoritma yaitu algoritma K-Nearest Neighbor, Naïve Bayes, Pohon Keputusan (C4.5), Rule Induction dan Deep Learning maka penulis mencoba membandingkan data minat siswa tersebut berdasarkan algoritma masing-masing. Adapun yang menjadi focus penulis untuk perbandingan adalah nilai akurasi dan nilai AUC dari masing-masing algoritma yang diuji.

Hasil dari percobaan pada masing-masing algoritma akan diuji menggunakan Cross Validation T-Test, dimana dari hasil T-Test tersebut diharapkan dapat diketahui algoritma mana yang mempunyai nilai akurasi paling baik, sehingga hasil yang dimasukan dalam table uji banding dapat dimanfaatkan guna kepentingan yang diinginkan

Data yang digunakan dalam paper ini adalah data poling siswa SMK TRIMITRA KARYA MANDIRI kelas 13 yang dilakukan pada tahun 2014. Paper ini menggunakan 220 data dari poling siswa SMK tersebut. Pada data tersebut mempunyai 4 jenis peminatan, yaitu Kuliah, Kerja, Kursus, Wiraswasta. Namun untuk kepentingan paper ini hanya digunakan 2 jenis peminatan yaitu Kuliah dan Kerja.

2. METODOLOGI

Data Mining adalah kajian yang meliputi kegiatan pengumpulan, pembersihan, pemrosesan, dan analisa sekumpulan data sehingga dengan kegiatan tersebut dapat diperoleh pemahaman yang mendalam akan data (Charu C. Aggarwal, 2015). Selain itu data mining juga dikenal sebagai disiplin ilmu yang mempelajari metode untuk mengekstrak pengetahuan atau menemukan pola dari suatu data yang besar (Jiawei Han, 2006). Secara teknis, *data mining* dapat disebut sebagai proses untuk menemukan korelasi atau pola dari ratusan atau ribuat *field* dari sebuah relasional *database* yang besar. Data mining juga merupakan proses mencari pola atau informasi menarik dalam data terpilih dengan menggunakan teknik atau metode tertentu. Pemilihan metode yang tepat tergantung pada tujuan dan proses KDD (*Knowledge Discovery in Database*) secara keseluruhan. Data mining mampu menganalisis data yang besar menjadi informasi

berupa pola yang mempunyai arti bagi pendukung keputusan (Setiawan, 2017).

Dalam paper ini akan digunakan salah satu metode data mining yaitu metode klasifikasi, dimana metode Klasifikasi adalah jenis analisis data yang dapat membantu orang memprediksi label kelas sampel harus diklasifikasikan. Berbagai macam teknik klasifikasi telah diusulkan dalam bidang-bidang seperti pembelajaran mesin, sistem pakar dan statistik. Biasanya, model klasifikasi dilatih pertama pada dataset sejarah (yaitu, training set) dengan label kelas mereka sudah dikenal. Tujuan dari klasifikasi adalah untuk menemukan model dari training set yang membedakan atribut ke dalam kategori atau kelas yang sesuai, model tersebut kemudian digunakan untuk mengklasifikasikan atribut yang kelasnya belum diketahui sebelumnya (Jiawei Han, 2006).

3. LANDASAN TEORI

Dalam metode klasifikasi terdapat beberapa algoritma yang dapat dilakukan, untuk paper ini penulis menggunakan 5 algoritma guna dibandingankan hasilnya satu dengan lainnya. Adapun algoritma yang digunakan adalah:

K-Nearest Neighbor

Algoritma k-NN adalah suatu metode yang menggunakan algoritma supervised. Perbedaan antara supervised learning dengan unsupervised learning adalah pada supervised learning bertujuan untuk menemukan pola baru dalam data dengan menghubungkan pola data yang sudah ada dengan data yang baru. Sedangkan pada unsupervised learning, data belum memiliki pola apapun, dan tujuan unsupervised learning untuk menemukan pola dalam sebuah data. Tujuan dari algoritma k-NN adalah untuk mengklasifikasi objek baru berdasarkan atribut dan training samples. Dimana hasil dari sampel uji yang baru diklasifikasikan berdasarkan mayoritas dari kategori pada k-NN. Pada proses pengklasifikasian, algoritma ini tidak menggunakan model apapun untuk dicocokkan dan hanya berdasarkan pada memori. Algoritma k-NN menggunakan klasifikasi ketetanggaan sebagai nilai prediksi dari sampel uji yang baru. Jarak yang digunakan adalah jarak Euclidean Distance. Jarak Euclidean adalah jarak yang paling umum digunakan pada data numerik (Ashari, 2013).

Naïve Bayes

Naive Bayes merupakan sebuah pengklasifikasian probabilistik sederhana yang menghitung sekumpulan probabilitas dengan menjumlahkan frekuensi dan kombinasi nilai dari dataset yang diberikan. Algoritma mengunakan teorema Bayes dan mengasumsikan semua atribut independen atau tidak saling ketergantungan yang diberikan oleh nilai pada variabel kelas. Definisi lain mengatakan Naive Bayes merupakan pengklasifikasian dengan metode probabilitas dan statistik yang dikemukan oleh ilmuwan Inggris Thomas Bayes, yaitu memprediksi peluang di masa depan berdasarkan pengalaman di masa sebelumnya (Sivakumari, 2009).

Naive Bayes didasarkan pada asumsi penyederhanaan bahwa nilai atribut secara kondisional saling bebas jika diberikan nilai output. Dengan kata lain, diberikan nilai output, probabilitas mengamati secara bersama adalah produk dari probabilitas individu. Keuntungan penggunaan Naive Bayes adalah bahwa metode ini hanya membutuhkan jumlah data pelatihan (Training Data) yang kecil untuk menentukan estimasi paremeter yang diperlukan dalam proses pengklasifikasian. Naive Bayes sering bekerja jauh lebih baik dalam kebanyakan situasi dunia nyata yang kompleks dari pada yang diharapkan.

Algoritma C4.5

Algoritma C4.5 dan pohon keputusan merupakan dua model yang tak terpisahkan, karena untuk membangun sebuah pohon keputusan dibutuhkan algoritma C4.5. Algoritma C4.5 merupakan pengembangan dari algoritma ID3. Algoritma C4.5 dan ID3 diciptakan oleh seorang peneliti di bidang kecerdasan buatan bernama J. Rose Quinlan pada akhir tahun 1970-an. Algoritma C4.5 membuat pohon keputusan dari atas ke bawah, di mana atribut paling atas merupakan akar *(root)*, dan yang paling bawah dinamakan daun *(leaf)* (Swastina, 2013).

Secara umum alur proses algoritma C4.5 untuk membangun pohon keputusan dalam *data mining* adalah [5]:

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

Rule Induction

Rumus Rule Induction sebagai berikut:

1. Untuk mendapatkan nilai support untuk sebuah item A
   Support = jumlah transaksi yang mengandung item A / Total transaksi
2. Untuk mencari nilai support dari 2-item
   Support (A,B) = P (A ∩ B)
   P (A ∩ B) = Jumlah transaksi yang mengandung A dan B / Total Transaksi
3. Mencari nilai confidance
   Confidance (A □ B) = P (A | B)
   P (A | B) = Jumlah transaksi yang mengandung A dan B / Jumlah transaksi yang mengandung item A
   Pada metode rule induction dalam proses perhitungannya menggunakan algoritma information gain.

Deep Learning

Deep Learning (deep machine learning, or deep structured learning, or hierarchical learning, or sometimes DL) adalah cabang dari machine learning berdasarkan satu set algoritma yang digunakan untuk model abstraksi tingkat tinggi pada data dengan menggunakan beberapa lapisan implementasi dan menggunakan struktur yang kompleks atau sebaliknya, terdiri dari beberapa transformasi non-linear (Charu C. Aggarwal, 2015).

Rapid Miner

Rapid Miner adalah sebuah software untuk pengolahan data mining. Rapid Miner adalah sebuah solusi untuk melakukan analisis terhadap data mining, text mining dan analisis prediksi. RapidMiner menggunakan berbagai teknik deskriptif dan prediksi dalam memberikan wawasan kepada pengguna sehingga dapat membuat keputusan yang paling baik (Setiawan, 2017). Rapid Miner memiliki kurang lebih 500 operator data mining, termasuk input, output, data preprocessing dan visualisasi. RapidMiner merupakan software yang berdiri sendiri untuk analisis data dan sebagai mesin data mining yang diintegrasikan pada produknya sendiri. RapidMiner ditulis menggunakan bahasa java sehingga dapat bekerja disemua sistem operasi.

Beberapa fitur dari Rapid Miner sebagai berikut :

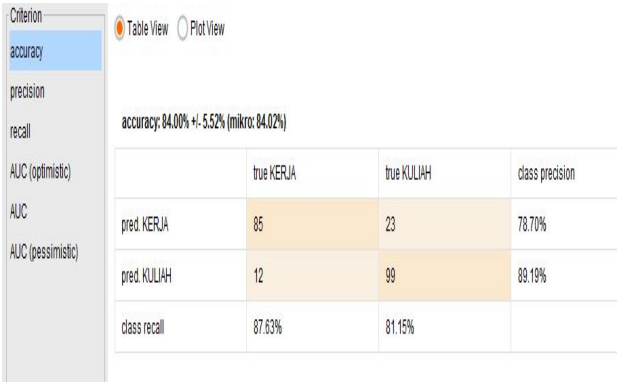
a. Banyaknya algoritma data mining, seperti decision tree.
b. Bentuk grafis yang canggih, seperti tumpang tindih diagram histogram, tree chart dan 3D scatter plots.
c. Banyaknya variasi plugin, seperti plugin untuk melakukan analisis teks.
d. Menyediakan prosedur data mining dan machine learning termasuk ETL (Extraction, Transformasi, Loading), data preprocessing, visualisasi, modeling dan evaluasi.
e. Proses data mining tersusun atas operator-operator yang nestable, dideskripsikan dengan XML dan dibuat dengan GUI.
f. Mengintegrasikan proyek data mining weka dan statistika R.

4. HASIL DAN PEMBAHASAN

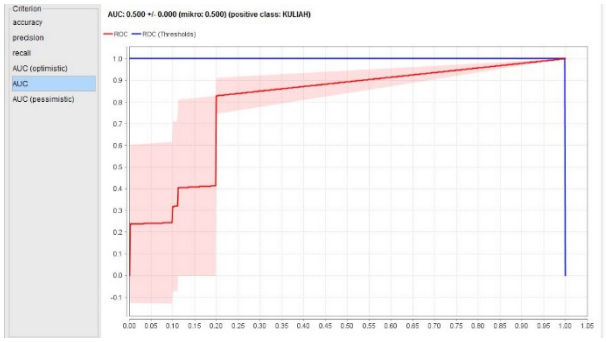
Penelitian ini bertujuan membandingkan minat siswa-siswa SMU khususnya SMK TRIMITRA KARYA MANDIRI yang duduk dikelas 13 pada saat mereka lulus sekolah SMU menggunakan 5 Algoritma yang berbeda pada metode klasifikasi dengan menggunakan variabel-variabel yang sudah ditentuan.

1. Hasil Implementasi Algoritma K-Nearest Neighbor

Pada implementasi *data mining* ini untuk menganalisis minat siswa SMU kelas 13 menggunakan aplikasi *Rapid Miner*. Dalam analisa K-Nearest Neighbor tersebut di dapatkan hasil berupa nilai *Accuracy* dan nilai AUC yang nantikan akan menjadi pembanding terhadap motede algoritma yang lainnya, berikut adalah hasilnya:



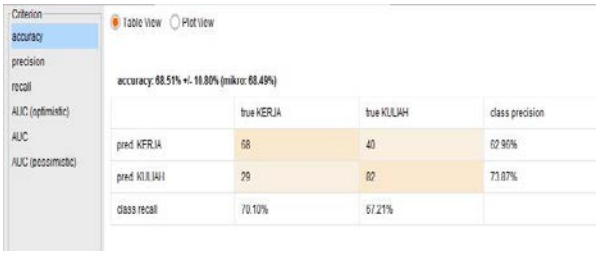
Gambar 3.1. Akurasi Algoritma K-Nearest Neighbor



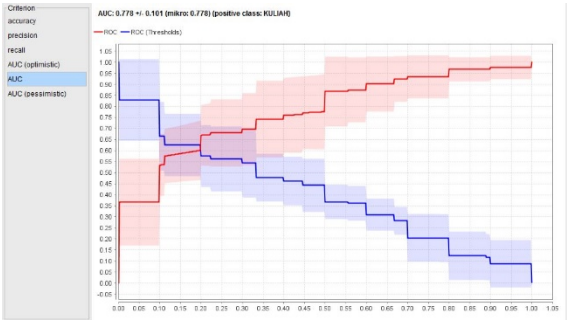
Gambar 3.2. AUC Algoritma K-Nearest Neighbor

2. Hasil Implementasi Algoritma Naive Bayes

Pada implementasi *data mining* ini untuk menganalisis minat siswa SMU kelas 13 menggunakan aplikasi *Rapid Miner*. Dalam analisa Naive Bayes tersebut di dapatkan hasil berupa nilai *Accuracy* dan nilai AUC yang nantikan akan menjadi pembanding terhadap motede algoritma yang lainnya, berikut adalah hasilnya:



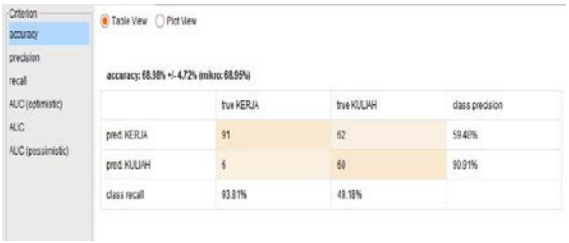
Gambar 3.3. Akurasi Algoritma Naive Bayes



Gambar 3.4. AUC Algoritma Naive Bayes

3. Hasil Implementasi Algoritma Pohon Keputusan C4.5

Pada implementasi *data mining* ini untuk menganalisis minat siswa SMU kelas 13 menggunakan aplikasi *Rapid Miner*. Dalam analisa Pohon Keputusan C4.5 tersebut di dapatkan hasil berupa nilai *Accuracy* dan nilai AUC yang nantikan akan menjadi pembanding terhadap motede algoritma yang lainnya, berikut adalah hasilnya:



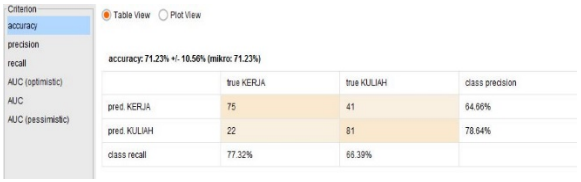
	true KERJA	true KULIAH	class precision
pred. KERJA	91	62	59.46%
pred. KULIAH	6	69	90.91%
class recall	93.81%	49.18%	

Gambar 3.5. Akurasi Algoritma Pohon Keputusan C4.5

Gambar 3.6. Akurasi Algoritma Pohon Keputusan C4.5

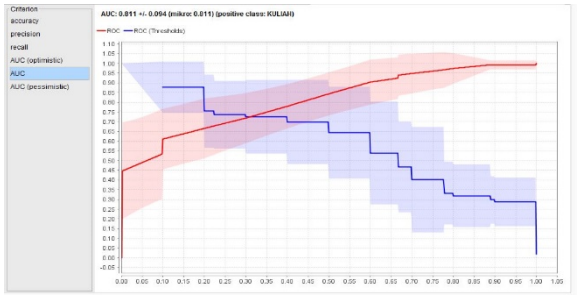
4. Hasil Implementasi Algoritma Rule Induction

Pada implementasi *data mining* ini untuk menganalisis minat siswa SMU kelas 13 menggunakan aplikasi *Rapid Miner*. Dalam analisa Rule Induction tersebut di dapatkan hasil berupa nilai *Accuracy* dan nilai AUC yang nantikan akan menjadi pembanding terhadap motede algoritma yang lainnya, berikut adalah hasilnya:



	true KERJA	true KULIAH	class precision
pred. KERJA	75	41	64.66%
pred. KULIAH	22	81	78.64%
class recall	77.32%	66.39%	

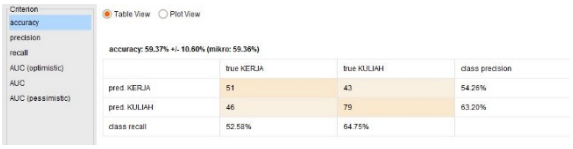
Gambar 3.7. Akurasi Algoritma Rule Induction



Gambar 3.8. Akurasi Algoritma Rule Induction

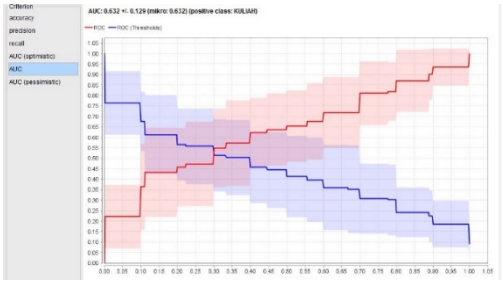
5. Hasil Implementasi Algoritma Deep Learning

Pada implementasi *data mining* ini untuk menganalisis minat siswa SMU kelas 13 menggunakan aplikasi *Rapid Miner*. Dalam analisa Deep Learning tersebut di dapatkan hasil berupa nilai *Accuracy* dan nilai AUC yang nantikan akan menjadi pembanding terhadap motede algoritma yang lainnya, berikut adalah hasilnya:



	true KERJA	true KULIAH	class precision
pred. KERJA	51	43	54.26%
pred. KULIAH	46	79	63.20%
class recall	52.58%	64.75%	

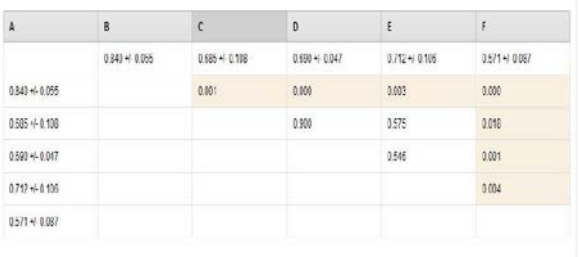
Gambar 3.9. Akurasi Algoritma Deep Learning



Gambar 3.10. AUC Algoritma Deep Learning

Dari hasil pengujian data minat siswa SMK TRIMITRA KARYA MANDIRI diatas menggunakan algoritma K-Nearest Neighbor, Naïve Bayes, Pohon Keputusan (C4.5), Rule Induction dan Deep Learning maka dilakukan uji banding Cross Validation dengan metode T-Test dengan mengukur tingkat Accuracy dan AUC.

Hasil uji tes Cross Validation dengan menggunakan T-test didapat hasil sebagai berikut:



A	B	C	D	E	F
	0.840 +/- 0.055	0.685 +/- 0.108	0.690 +/- 0.047	0.712 +/- 0.106	0.571 +/- 0.087
0.840 +/- 0.055		0.001	0.000	0.003	0.000
0.685 +/- 0.108			0.900	0.575	0.010
0.690 +/- 0.047				0.546	0.001
0.712 +/- 0.106					0.004
0.571 +/- 0.087					

Gambar 3.11. Hasil Cross Validation T-Test

Dari hasil T-test tersebut dimasukan dalam bentuk tabel perbandingan nilai Accuracy dan AUC dari setiap teknik klasifikasi model yang digunakan:

Tabel 3.1. Data Perbandingan Uji Algoritma

	KNN	NB	C4.5	RULE	DEEP
ACCURACY	84.00 %	68.50 %	69.50 %	71.20 %	57.10 %
AUC	5.50	10.80	4.70	10.60	8.70

Metode Klasifikasi

Melihat tabel data uji banding di atas dapat disimpulkan bahwa teknik model klasifikasi yang terbaik yaitu algoritma K-Nearest Neighbor , kemudian algoritma Rule Induction, algoritma C4.5, algoritma Naive Baiyes , dan terakhir adalah algoritma Deep Learning.

5. KESIMPULAN

Berdasarkan hasil penelitian ini maka dapat ditarik kesimpulan bahwa:

1. Berdasarkan hasil uji banding dari ke lima algoritma yaitu algoritma K-Nearest Neighbor, Naïve Bayes, Pohon Keputusan (C4.5), Rule Induction dan Deep Learning variable yang berpengaruh terhadap hasil adalah *minat siswa*.
2. Berdasarkan dari nilai *precision* dan *accuracy*, metode *Decision Tree* (C4.5) memiliki nilai lebih tinggi daripada algoritma yang lainnya dengan nilai *precision* sebesar 89,09 % dan nili *accuracy* sebesar 84,00 %.

DAFTAR PUSTAKA

Ashari, A. (2013). Performance Comparison between Naïve Bayes , Decision Tree and k-Nearest Neighbor in Searching Alternative Design in an Energy Simulation Tool, *4*(11), 33–39.

Charu C. Aggarwal, C. Z. (2015). Mining Text Data.

Jiawei Han, M. K. (2006). *Data Mining, Concept And Technique*.

Setiawan, D. (2017). Optimalisasi Data Mining Mengunakan Rapid Miner Untuk Prediksi Kelulusan Mahasiswa.

Sivakumari, S. (2009). Accuracy Evaluation Of C4 . 5 And Naive Bayes Classifier Using Attribute Ranking Method Received : 15-05-2008 Revised : 21-01-2009, *2*(1), 60–68.

Swastina, L. (2013). Penerapan Algoritma C4 . 5 Untuk Penentuan Jurusan Mahasiswa, *2*(1).