PENERAPAN ALGORITMA KLASIFIKASI UNTUK PENENTUAN MAHASISWA BERPOTENSI *DROP OUT*

Ni Luh Ratniasih
Program Studi Sistem Informasi
Institut Teknologi Dan Bisnis STIKOM Bali, Denpasar, Bali
ratni@stikom-bali.ac.id

ABSTRACT

The use of data mining is very helpful in the classification of students who have the potential to drop out. There are several algorithms that can be applied in its implementation so that in this study a comparison of the two classification algorithms is carried out by comparing the level of accuracy produced by the two algorithms. The classification algorithms compared in this study are the K-Nearest Neighbor algorithm and the C4.5 algorithm. The data used in this study were students of the ITB STIKOM Bali class of 2014 using 6 attributes, namely gender, age, religion, class status, practical work, and GPA value. The results showed that the accuracy rate of the K-Nearest Neighbor algorithm was 81.50% while the accuracy level of the C4.5 algorithm was 80.54%.

Keywords: *KNN, C4.5, Student Drop Out*

ABSTRAK

Pemanfaatan data mining sangat membantu dalam klasifikasi mahasiswa yang berpotensi drop out. Terdapat beberapa algoritma yang dapat diterapkan dalam implementasinya sehingga dalam penelitian ini dilakukan komparasi dua algoritma klasifikasi dengan membandingkan tingkat akurasi yang dihasilkan kedua algoritma. Algoritma klasifikasi yang dibandingkan dalam penelitian ini adalah algoritma K-Nearest Neighbour dan algoritma C4.5. Data yang digunakan dalam penelitian ini adalah mahasiswa ITB STIKOM Bali angkatan tahun 2014 dengan menggunakan 6 atribut yaitu jenis kelamin, umur, agama, status kelas, kerja praktek, dan nilai IPK. Hasil penelitian menunjukkan bahwa tingkat akurasi algoritma K-Nearest Neighbour sebesar 81.50% sedangkan tingkat akurasi algoritma C4.5 sebesar 80.54%.

Kata kunci: *KNN, C4.5, Mahasiswa Drop Out*

PENDAHULUAN

Data mahasiswa *drop out* menjadi sesuatu hal yang penting untuk diteliti dan dianalisa, sehingga dapat diketahui bagaimana karakteristik mahasiswa yang berpotensi *drop out* sedini mungkin. Institut Teknologi dan Bisnis (ITB) STIKOM Bali merupakan salah satu perguruan tinggi di Bali di bidang ICT (*Information Communication Technology*) yang memiliki data sangat besar. Salah satu cara untuk melakukan pengolahan data yang besar adalah dengan metode penggalian data atau *data mining*. Penggalian data berdasarkan data pendidikan di perguruan tinggi dapat meningkatkan kualitas pembelajaran mahasiswa di Perguruan Tinggi [1]. Menurut Turban, data mining diartikan sebagai suatu istilah yang digunakan untuk menguraikan penemuan pengetahuan di dalam basis data [2]. Teknologi *data mining* dapat dimanfaatkan untuk menggali pengetahuan di basis data ITB STIKOM Bali untuk mengetahui model yang menggambarkan karakteristik mahasiswa yang berpotensi *drop out*.

Seiring dengan perkembangan teknologi, *data mining* dengan teknik klasifikasi data dikembangkan dengan metode – metode. Terdapat beberapa penelitian yang telah dilakukan terkait klasifikasi data menggunakan metode. Salah satu penelitian terkait adalah penelitian dalam menganalisis perbandingan *algoritma classification* untuk *authentication* uang kertas, dimana hasil persentase akurasi yang sangat tinggi adalah menggunakan metode *Decision Tree C4.5* dengan nilai akurasi sebesar 98,5 %,

sedangkan *Neural Network* sebesar 95%, dan *Navies Bayes* sebesar 85% [3]. Selanjutnya penelitian yang menerapkan teknik *data mining* menggunakan metode *Support Vector Machine* (SVM) untuk memprediksi siswa yang berpeluang *drop out* menghasilkan tingkat akurasi sebesar 43,33% [4]. Penelitian yang mengimplementasi pengolahan citra dan klasifikasi *K-Nearest Neighbour* untuk membangun aplikasi pembeda daging sapi dan daging babi berbasis web, penelitian tersebut berhasil melakukan klasifikasi perbedaan daging babi dan daging sapi berdasarkan extraksi ciri tekstur dengan akurasi 88,75%[5]. Penelitian terkait klasifikasi

Berdasarkan uraian latar belakang di atas maka dalam penelitian ini akan dilakukan penelitian untuk klasifikasi mehasiswa *drop out* dengan menerapkan algoritma *K-Nearest Neighbour* untuk mengetahui tingkat akurasi dari metode tersebut. Data mahasiswa yang digunakan dalam penelitian ini adalah mahasiswa program studi Sistem Informasi angkatan tahun 2014 dengan menggunakan atribut yaitu jenis kelamin, umur, agama, status kelas, kerja praktek, dan nilai IPK.

LANDASAN TEORI

Definisi *Data Mining*

Menurut Gartner Group, *data mining* adalah suatu proses menemukan hubungan yang berarti, pola, dan kecenderungan dengan memeriksa dalam sekumpulan besar data yang tersimpan dalam penyimpanan dengan menggunakan teknik pengenalan pola seperti teknik statistik dan matematika [7]. *Data mining* bukanlah suatu bidang yang sama sekali baru. Salah satu kesulitan untuk mendefinisikan *data mining* adalah kenyataan bahwa *data mining* mewarisi banyak aspek dan teknik dari bidang-bidang ilmu yang sudah mapan terlebih dulu.

Berawal dari beberapa disiplin ilmu, *data mining* bertujuan untuk memperbaiki teknik tradisional sehingga bisa menangani:

1. Jumlah data yang sangat besar
2. Dimensi data yang tinggi
3. Data yang heterogen dan berbeda bersifat

Metode *K-Nearest Neighbour*

Algoritma K-*Nearest Neighbor* merupakan metode klasifikasi yang mengelompokkan data baru berdasarkan jarak data baru itu kedalam beberapa data tetangga (*negihbord*) terdekat. Teknik *K-Nearest Neighbor* dengan melakukan langkah-langkah yaitu, *input* : data training, label data training, k, data testing [8].

mahasiswa *drop out* telah dilakukan pada tahun 2018 dengan judul perancangan sistem klasifikasi mahasiswa *drop out* menggunakan algoritma *K-Nearest Neighbor* dengan tempat studi kasus STMIK STIKOM Bali. Pada penelitian ini komponen atribut yang digunakan adalah Indeks Prestasi Komulatif (IPK), jurusan, penghasilan orang tua, semester dengan menggunakan nilai k yang berbeda. Hasil penelitian ini adalah sebuah sistem yang dapat digunakan untuk mengklasifikasikan mahasiswa *drop out* namun tidak membahas tingkat akurasi dari sistem tersebut [6].

Proses yang dilakukan dalam *K-NN* untuk mendapatkan clas kategori adalah dengan menghitung kemiripan antara data baru dengan tiap – tiap data yang sebelumnya telah dikategotrikan. Dengan kata lain algoritma ini bekerja berdasarkan dari data baru terhadap data yang sudah ada atau latih. Data latih tersebut kemudian diurutkan mulai dari data yang memiliki nilai kemiripan paling besar dengan data yang akan dikategorikan. Data dipilih sebanyak *k* data dengan nilai kemiripan terbesar, kemudian memprediksikan data yang baru ke dalam kategori data tersebut.Adapun rumus untuk melakukan perhitungan kedekatan antara dua kasus dapat dilihat pada persamaan 3 sebagai berikut :

Similarity (T,S) = X = $\frac{\sum_{i=1}^{n} f(Ti,Si) * wi}{\text{wi}}$

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* kedekatan biasanya berada pada nilai antara 0 s.d. 1. Nilai 0 artinya kedua kasus mutlak tidak mirip, sebaliknya untuk nilai 1 kasus mirip dengan mutlak.

Decision Tree (Pohon Keputusan)

Decision tree (pohon keputusan) adalah sebuah diagram alir yang mirip dengan struktur pohon, dimana setiap internal node menotasikan atribut yang diuji, setiap cabangnya merepresentasikan hasil dari atribut tes tersebut, dan leaf node merepresentasikan kelas-kelas tertentu atau distribusi dari kelas-kelas [2].

Klasifier pohon keputusan merupakan teknik klasifikasi yang sederhana yang banyak

digunakan. Bagian ini membahas bagaimana pohon keputusan bekerja dan bagaimana pohon keputusan dibangun. Seringkali untuk mengklasifikasikan obyek, kita ajukan urutan pertanyaan sebelum bisa kita tentukan kelompoknya.

Walaupun banyak variasi model decision tree dengan tingkat kemampuan dan syarat yang berbeda, pada umumnya beberapa ciri kasus cocok untuk diterapkan *decision tree* [3] :

1. Data dinyatakan dengan pasangan atribut dan nilainya. Misalnya atribut satu data adalah temperatur dan nilainya adalah dingin. Biasanya untuk satu data nilai dari satu atribut tidak terlalu banyak jenisnya. Dalam contoh atribut warna buah ada beberapa nilai yang mungkin yaitu hijau, kuning, merah.
2. Label/output data biasanya bernilai diskrit. Output ini bisa bernilai ya atau tidak, sakit atau tidak sakit, diterima atau ditolak. Dalam beberapa kasus mungkin saja outputnya tidak hanya dua kelas, tetapi penerapan decision tree lebih banyak untuk kasus binary.
3. Data mempunyai missing value. Misalkan untuk beberapa data, nilai dari suatu atributnya tidak diketahui. Dalam keadaan seperti ini decision tree masih mampu memberi solusi yang baik.

Algoritma C4.5

Algoritma C4.5 adalah salah satu algoritma untuk mengubah fakta yang besar menjadi pohon keputusan (*decision tree*) yang merepresentasikan aturan (*rule*). Tujuan dari pembentukan pohon keputusan dalam algoritma C4.5 adalah untuk mempermudah dalam penyelesaian permasalahan.

Dalam menggunakan algoritma C4.5 terdapat beberapa tahapan yang umum yaitu pertama mengubah bentuk data dalam tabel menjadi model pohon kemudian mengubah model pohon menjadi aturan (*rule*) dan terakhir menyederhanakan rule [4].

Secara umum, algoritma C4.5 untuk membangun sebuah pohon keputusan adalah sebagai berikut:

a. Hitung jumlah data, jumlah data berdasarkan anggota atribut hasil dengan syarat tertentu. Untuk proses pertama syaratnya masih kosong.
b. Pilih atribut sebagai *Node*.
c. Buat cabang untuk tiap-tiap anggota dari *Node*.
d. Periksa apakah nilai *entropy* dari anggota Node ada yang bernilai nol. Jika ada, tentukan daun yang terbentuk. Jika seluruh nilai *entropy* anggota Node adalah nol, maka proses pun berhenti.
e. Jika ada anggota Node yang memiliki nilai entropy lebih besar dari nol, ulangi lagi proses dari awal dengan *Node* sebagai syarat sampai semua anggota dari *Node* bernilai nol.

Node adalah atribut yang mempunyai nilai gain tertinggi dari atribut-aribut yang ada. Untuk menghitung nilai gain suatu atribut digunakan rumus seperti yang tertera dalam persamaan berikut:

Berikut adalah bentuk umum dari C4.5 :

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

Keterangan:
S : himpunan kasus
A : atribut
n : jumlah partisi atribut A
|Si| : jumlah kasus pada partisi ke-i
|S| : jumlah kasus dalam S

penghitungan nilai entropi dapat dilihat pada persamaan berikut.

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

Keterangan:
S : himpunan kasus
A : fitur
n : jumlah partisi S
pi : proporsi dari Si terhadap S

Rapid Miner

Rapid miner adalah aplikasi data mining yang berbasis *open source*. *Open source* rapid miner berlisensi AGPL (*GNU Affero General Public License*) versi 3. Penelitian mengenai *tools* ini dimulai sejak tahun 2001 oleh Ralf Klinkenberg, Ingo Mierswa, dan Simon Fischer di Artificial Intelligence Unit dari *University of Dortmund* yang kemudian diambil alih oleh SourceForge sejak tahun 2004. Rapid miner memperoleh peringkat satu sebagai *tools* data mining untuk proyek nyata pada poll oleh KDnuggets, sebuah koran data-mining pada 2010-2011.

Dalam penerapannya, rapid miner menyediakan prosedur *data mining* dan *machine* learning termasuk : ETL (*extraction, transformation, loading*), data preprocessing, visualisasi, modelling dan evaluasi. Proses data mining tersusun atas operator-operator yang nestable, dideskripsikan dengan XML, dan dibuat dengan GUI. *Tools* rapid miner ditulis dalam bahasar pemrograman Java dan

juga mengintegrasikan proyek data mining Weka dan statistika [5].

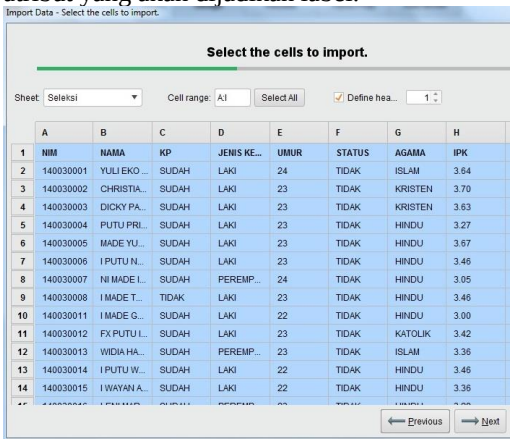
Beberapa solusi yang diusung oleh rapid miner antara lain :

a. Integrasi data, Analitis ETL, Data Analisis, dan pelaporan dalam satu suite tunggal.
b. Powerfull tetapi memiliki antarmuka pengguna grafis yang intuitif untuk desain anakisis proses.
c. Repositori untuk prose, data dan penanganan meta data.
d. Hanya solusi dengan transformasi meta data: lupakan trail and arror dan memeriksa hasil yang telah diinspeksi selama desain.
e. Hanya solusi yang mendukung *on-the-fly* kesalahan dan dapat melakukan perbaikan dengan cepat. Lengkap dan fleksibel: ratusan *loading* data, transformasi data, pemodelan data dan metode visualisasi data.

HASIL DAN PEMBAHASAN

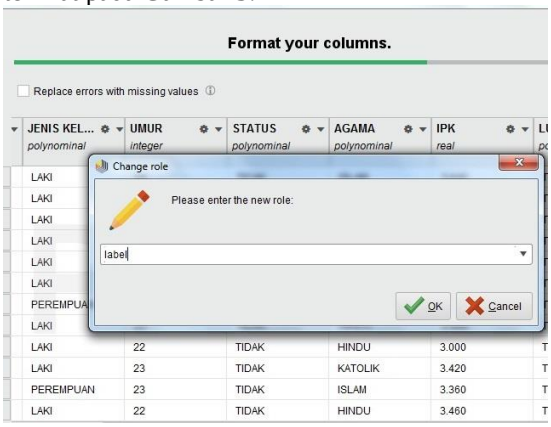
Pada tahap ini dilakukan pengujian algoritma *K-Nearest Neighbour* untuk mengetahui tingkat akurasi metode tersebut. Perhitungan tingkat akurasi dilakukan menggunakan aplikasi *RapidMiner Studio*. Dari tingkat akurasi yang dihasilkan dapat menentukan tinggi rendahnya kesalahan prediksi pada adat baru. semakin tinggi tingkat akurasi maka semakin rendah kesalahan prediksi pada data baru demikian sebaliknya. Proses perhitungan tingkat akurasi dilakukan dengan metode *Confussion Matrix*.

Tahap awal pengujian adalah *import* data yang akan diproses seperti Gambar 1, kemudian pemilihan cell range dari cell C sampai I selanjutnya menentukan kategori dan atribut yang akan dijadikan label.

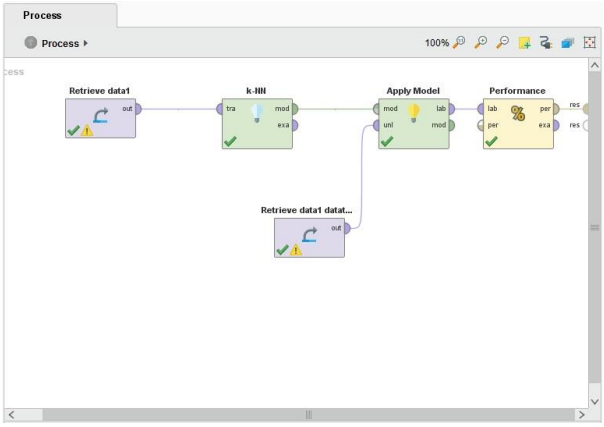


Gambar 1. Import Data

Atribut yang dijadikan sebagai label adalah atribut “LULUS” yang berada di cell I seperti Gambar 2. Tahap selanjutnya adalah melakukan pengujian dengan algoritma *K-Nearest Neighbour* sehingga terbentuk pemodelan *K-Nearest Neighbour* seperti yang terlihat pada Gambar 3.



Gambar 2. Menentukan Label



Gambar 3. Pemodelan Algoritma *K-Nearest Neighbour*

Pada Gambar 3 terdapat parameter dan operator yang digunakan pada model algoritma *K-Nearest Neighbour* sebagai berikut :

a. *Retrieve data* dan *retrieve data dataset* adalah operator yang digunakan untuk meng-*import dataset* yang akan digunakan, pada tahap penelitian ini data di-*import* dari file excel.
b. *K-Nearest Neighbour* adalah metode klasifikasi yang digunakan dalam tahap penelitian ini.
c. *Apply Model* digunakan agar dapat membaca data yang akan diestimasi berdasarkan data yang telah dibaca sebelumnya.

d. *Performance* adalah operator yang digunakan untuk mengukur performa akurasi dari model *K-Nearest Neighbour*.

Berdasarkan hasil *performance* algoritma *K-Nearest Neighbour*, diperoleh tingkat akurasi (*accuracy*) sebesar 81.50%.

accuracy: 81.50%

	true TIDAK	true TEPAT	class precision
pred. TIDAK	99	32	75.57%
pred. TEPAT	64	324	83.51%
class recall	60.74%	91.01%	

Gambar 4. Akurasi Algoritma K-NN

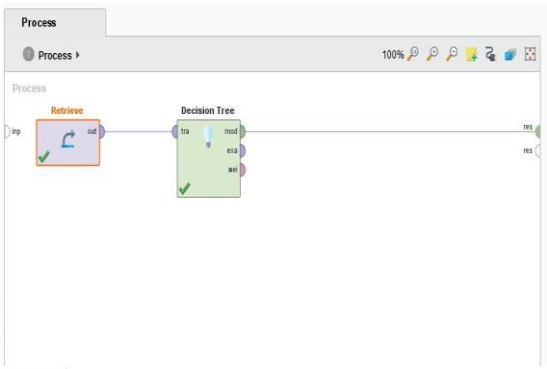
b. Pengujian Algoritma *C4.5*

Pengujian algoritma C4.5 dilakukan dengan menggunakan data yang sama dengan pengujian algoritma sebelumnya yaitu sebanyak 519 *record* data. Terdapat 6 atribut yang digunakan dalam penelitian ini dengan pengelompokan kategori sebagai berikut :

Tabel 1. Kelompok Kategori

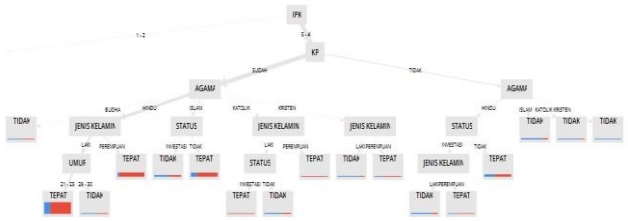
ATRIBUT	KATEGORI
Jenis Kelamin	Laki - laki
	Perempuan
Kerja Praktek	Sudah Selesai
	Tidak Selesai
Umur	21 - 25
	26 - 30
	31 - 35
	36 - 40
	41 - 45
Status Kelas	Investasi
	Tidak Investasi
Agama	Hindu
	Islam
	Kristen
	Katolik
	Budha
IPK	1 - 2
	3 - 4

Pengujian metode untuk menghitung tingkat akurasi dilakukan menggunakan aplikasi *RapidMiner Studio*. Hasil design dalam implementasi metode C4.5 dapat digambarkan seperti Gambar 5.



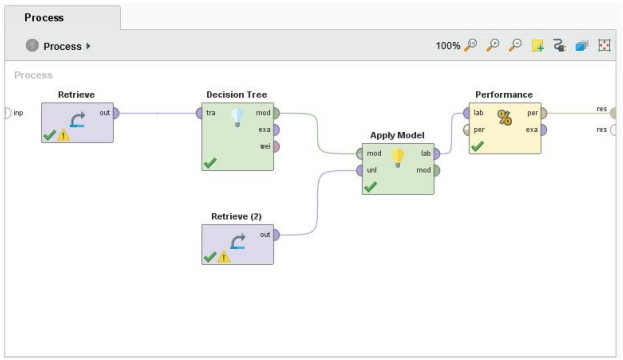
Gambar 5. Design Algoritma C4.5

Pohon keputusan yang terbentuk dari hasil implementasi algoritma C4.5 dapat dilihat pada Gambar 6.



Gambar 6. Pohon Keputusan

Tahap selanjutnya melakukan pengujian tingkat akurasi model dengan menggunakan framework RapidMiner. Gambar hasil pengujian dapat dilihat pada Gambar 7.



Gambar 7. Pengujian Algoritma C4.5

Berdasarkan hasil *performance* algoritma *C4.5*, diperoleh tingkat akurasi (*accuracy*) sebesar 80.54%.

accuracy: 80.54%

	true TIDAK	true TEPAT	class precision
pred. TIDAK	71	9	88.75%
pred. TEPAT	92	347	79.04%
class recall	43.56%	97.47%	

Gambar 8. Akurasi Algoritma C4.5

Pengujian dilakukan untuk mengetahui tingkat akurasi dari algoritma K-NN dan C4.5 dengan dataset mahasiswa sebanyak 519. Hasil pengujian menujukkan bahwa tingkat akurasi algoritma *K-Nearest Neighbour* sebesar 81.50% sedangkan algoritma C4.5 sebesar 80.54%.

Tabel 5. 2 Hasil Uji Algoritma

Algoritma	Tingkat Akurasi
K-Nearest Neighbour	81.50 %
C4.5	80.54 %

SIMPULAN

Berdasarkan hasil penelitian dapat diambil kesimpulan bahwa algoitma K-Nearest Neighbour dan C4.5 dapat diterapkan dalam klasifikasi mahasiswa drop out dengan menggunakan 6 atribut yaitu jenis kelamin, umur, agama, status kelas, kerja praktek, dan nilai IPK, dimana hasil pengujian menujukkan bahwa tingkat akurasi algoritma K-Nearest Neighbour sebesar 81.50% sedangkan algoritma C4.5 sebesar 80.54%.

DAFTAR PUSTAKA

[1].Tair, M. Mohammed Abu dan El-halees, Alaa M., "Mining Educational Data to Improve Students Performance: A Case Study", International Journal of Information and Communication Technology Research, Vol. 2, No. 2, 140-146, 2012.

[2].Turban, Efraim, Aronson, Jay E. dan Liang, Ting-Peng, "Decision Support System and Intelligent System", Prentice-Hall of India: New Delhi, 7th ed, 2007.

[3].Khairul Sani, Wing Wahyu Winarno, Silmi Fauziati, 2016, Analisis Perbandingan Algoritma Classification Untuk Authentication Uang Kertas (Studi Kasus: Banknote Authentication), Jurnal Informatika Vol. 10, No. 1.

[4].Fiska, Ryci Rahmatil, 2017, Penerapan Teknik Data Mining dengan Metode Support Vector Machine (SVM) untuk Memprediksi Siswa yang Berpeluang Drop Out (Studi Kasudi SMKN 1 Sutera), Jaringan Sistem Informasi Robotik Vol. 1, No. 01.

[5].Dea Alverina, A. R. Chrismanto, and R. G. Santosa, 2018, Perbandingan Akurasi Algoritma C4.5 dan CART dalam Memprediksi Kategori Indeks Prestasi Mahasiswa, Jurnal Teknologi dan Sistem Komputer, vol. 6, no. 2, Apr. 2018. doi: 10.14710/jtsiskom.6.2.2018.76-83, ISSN:2338-0403.

[6].Ramayasa, I Putu. 2018. "Perancangan Sistem Klasifikasi Mahasiswa Drop Out Menggunakan Algoritma K-Nearest Neighbor". Seminar Nasional Sistem Informasi dan Teknologi Informasi 2018. STMIK Pontianak. 585 – 589.

[7].Larose, D.T, 2006. Discovering Knowledge in Data: An Introduction to Data mining. John Willey & Sons, Inc.

[8].Santosa B. 2007. Data Mining : Teknik Pemanfaatan Data untuk Keperluan Bisnis, Teori dan Aplikasi. Graha Ilmu Yogyakarta.

[9].Rapid-I GmbH. (2008). Rapidminer-4.2-tutorial. Germany: Rapid-I