

PREDIKSI KELULUSAN MAHASISWA TEPAT WAKTU BERDASARKAN USIA, JENIS KELAMIN, DAN INDEKS PRESTASI MENGUNAKAN ALGORITMA *DECISION TREE*

Agus Romadhona¹, Suprapedi², H. Himawan³

¹²³Pascasarjana Teknik Informatika Universitas Dian Nuswantoro

ABSTRACT

Prediction of the study period in college is needed to determine the accuracy of the students' study period according to the specified time so that wisdom of prevention related to the study period is not lost. This research aims to find patterns to predict the timely graduation of students using data mining techniques and models to predict long period of study was Decision tree algorithm C4.5 to compare with ID3 and CHAID algorithms using test data to determine the percentage of precision, recall and accuracy is obtained that the algorithm Decision Tree C4.5 has a better performance compared with other algorithms. From this research it was found that the prediction of the students' study period are affected by incoming students' age, gender, GPA semesters 1 through 4 semesters GPA and the most influential is the 4th semester GPA of students graduate on time with a value of 0.340 gain of all attributes. Decision tree algorithm C4.5 reaches the highest accuracy on the amount of data 389 with 91.51% accuracy values for k-fold=3, 90.75 for k-fold = 5 and 90.77 with k-fold = 10, While ID3 and CHAID algorithms achieving a low accuracy value. So thus the value accuracy of Decision Tree algorithm C4.5 is better than the ID3 and CHAID algorithm. In this research, training data are used as much as 389. To see better performance in the accuracy of the results of each algorithm, thus for furthermore research the number of data records used training process should be improved.

Keywords: Decision Tree C.45, ID3, CHAID, Prediction, Student.

1. PENDAHULUAN

Perguruan tinggi merupakan penyelenggara pendidikan akademik bagi mahasiswa [1]. Pada setiap tahun akademik perguruan tinggi menyelenggarakan proses penerimaan mahasiswa baru. Dilihat dari banyaknya peminat dari tiap tahun akademik perguruan tinggi dapat memperhatikan berbagai faktor yang berpengaruh terhadap pengelolaan daya tampung mahasiswa, salah satunya ketepatan dari masa studi mahasiswa sesuai dengan waktu yang telah ditentukan.

Perguruan tinggi perlu mendeteksi perilaku mahasiswa sehingga dapat diketahui faktor yang menjadi penyebab kegagalan mahasiswa sehingga tidak lulus atau lulus sesuai dengan masa studi yang telah ditetapkan, diantaranya adalah rendahnya kemampuan akademik, usia masuk perguruan tinggi, indeks prestasi ataupun faktor lainnya. Salah satu penyelenggara pendidikan di Sulawesi Tengah adalah STMIK Adhiguna pada tiap tahun akademik banyak peminatnya dari dua program studi.

Dalam tiap tahun akademik terdapat beberapa hal yang tidak seimbang antara mahasiswa masuk dan keluar dalam menyelesaikan studinya. Mahasiswa yang masuk dalam jumlah besar tetapi mahasiswa yang lulus tepat waktu sesuai dengan ketentuan yaitu 3 (tiga) tahun 6 (enam) bulan jauh sangat kecil dibandingkan masuknya. Sehingga terjadi penumpukan mahasiswa dalam jumlah tinggi disetiap periode kelulusan sehingga proses akademik tidak berjalan maksimal, lama studi kelulusan mahasiswa lulus tepat waktu dan lulus tidak tepat waktu dapat dilihat pada lampiran 4.

Dari data penelitian yang diperoleh perlunya mengetahui sejak dini mahasiswa yang diidentifikasi akan gagal atau lulus tidak tepat waktu, sehingga kepada mahasiswa tersebut dapat diberikan beberapa alternatif pilihan atau peningkatan kemampuan melalui bimbingan yang sifatnya khusus. Ada beberapa penelitian yang membahas atau memprediksi keberhasilan mahasiswa dalam menempuh studi.

Marselina Silvia Suhartinah melakukan penelitian tentang prediksi kelulusan mahasiswa sesuai

dengan waktu studi dengan menggunakan algoritma *Naïve Bayes* dan C4.5 teknik klustering, analisis diskriminan, *Teorema Bayes*, *Decision Tree*, *Artificial Neural Networks*, *Support Vector Machine*, *Regresi Linear*, *Support Vector Regresi* [2]. Metode yang digunakan untuk mengolah data yang sifatnya besar untuk menemukan pola yang terdapat di dalamnya diantaranya teknik klustering, analisis diskriminan, *Teorema Bayes*, *Decision Tree* *Artificial Neural Networks*, *Support Vector Machine*, *Regresi Linear*, *Support Vector Regresi* yang dapat digunakan untuk memproses data.

Muhammad Hanief Meinanda, dkk.. melakukan penelitian menggunakan metode minimasi *Sum Square Error* (SSE) sedangkan model untuk lama studi menggunakan *Artificial Neural Network* (ANN) dengan arsitektur *Multilayer Perceptron* (MP) diperoleh bahwa lama studi mahasiswa dipengaruhi oleh Indeks Prestasi Kumulatif (IPK), jumlah mata kuliah yang diambil, jumlah mata kuliah yang mengulang dan pengambilan jumlah matakuliah tertentu [3].

S. Anupama Kumar, dkk.. melakukan penelitian untuk memprediksi tingkat keberhasilan siswa yang cenderung gagal atau lulus berdasarkan nilai studi siswa menggunakan algoritma *Tree* C4.5 kemudian membandingkan dengan algoritma ID3 dengan hasil prediksi dapat menemukan pola yang dapat membantu siswa yang cenderung lemah dapat ditingkatkan nilai yang lebih baik [4].

Lillyan Hadjaratie melakukan penelitian prediksi tingkat kelulusan mahasiswa dengan *Artificial Neural Network* (ANN) metode propagasi balik (*Back Propagation*). Prediksi tingkat kelulusan adalah lama studi dan Indeks Prestasi Kumulatif (IPK). Variabel inputnya berupa nilai angka mutu dari 16 (enam belas) mata kuliah dari 2 (dua) semester pada tahun pertama program perkuliahan. Variabel outputnya berupa lama studi Indeks Prestasi Akademik (IPK) [5].

Sujana melakukan penelitian dengan aplikasi *mining data* mahasiswa dengan metode klasifikasi *Decision Tree*. Pada penerapan algoritma C4.5 apakah IPK seorang mahasiswa dapat diperkirakan berdasarkan nilai beberapa mata kuliah yang dianggap paling signifikan dalam menentukan IPK seorang mahasiswa? Peneliti mengambil matakuliah hanya sampai semester IV dan rata-rata 2 matakuliah, sehingga dosen dapat melihat IPK dan menentukan matakuliah yang harus diulang atau diambil agar IPK seorang mahasiswa dapat meningkat [6].

Dari hasil uji, prosentase dari beberapa matakuliah terhadap nilai IPK seorang mahasiswa. Prosentase *error rate* yang dihasilkan pada hasil *testing* rata-rata adalah di bawah 50%, bahkan ada yang 26 %. Itu menandakan bahwa *rule* yang dihasilkan sudah cukup baik. Hasil ini diperoleh dari data *training* pada matakuliah Algoritma I, Algoritma II, Basis Data, Statistika, dan Struktur Data. Data *training* pada matakuliah tersebut menghasilkan *rule* yang digunakan untuk data *testing* dengan prosentase *error rate* yang sangat kecil. Semakin besar prosentase nilai *error rate* yang dihasilkan pada data *testing*, maka *rule* yang dihasilkan pun tidak baik. Sebaliknya, semakin kecil prosentase *error rate* yang dihasilkan pada data *testing*, maka akan menghasilkan *rule* yang baik pula.

Berdasarkan beberapa pandangan penelitian terkait tersebut di atas penulis melakukan penelitian dengan mengambil data dari indeks prestasi mahasiswa yang diperoleh dari pengambilan matakuliah yang ditawarkan pada setiap semesternya sesuai dengan kurikulum yang berlaku saat ini. Persentase kurikulum pada STMIK Adhiguna mengacu pada Keputusan Menteri Pendidikan Nasional Republik Indonesia Nomor 045/u/2002 tentang kurikulum inti pendidikan tinggi.

Penelitian dilakukan dengan mengambil data dari indeks prestasi mahasiswa yang diperoleh dari pengambilan matakuliah yang ditawarkan pada setiap semesternya. Kelulusan tepat waktu merupakan indikator keberhasilan mahasiswa dalam proses pendidikan dilihat dari nilai yang diperoleh, indeks prestasi serta faktor lain yang dapat menyebabkan kegagalan masa studi mahasiswa sehingga perlu adanya *tool* yang mampu memprediksi. Penekanan pada penelitian ini adalah mahasiswa angkatan 2009 sebagai *training set* yang akan diprediksi kelulusannya ditinjau dari usia, jenis kelamin, indeks prestasi selama 4 semester pertama yaitu semester 1, semester 2, semester 3 dan semester 4. Selanjutnya dilakukan analisis komparasi algoritma klasifikasi yaitu *Algoritma Decision Tree C4.5*, *ID3* dan *Chaid* sehingga dapat diketahui algoritma yang paling akurat untuk memprediksi kelulusan mahasiswa lulus tepat waktu pada STMIK Adhiguna.

Diharapkan penelitian ini dapat bermanfaat dan dapat dijadikan sebagai bahan pertimbangan bagi program studi untuk melakukan langkah-langkah dalam meningkatkan persentase tingkat kelulusan mahasiswa yang ditetapkan dan untuk mencegah secara dini kegagalan akademik mahasiswa.

2. TINJAUAN PUSTAKA

2.1. Penelitian Terkait

Marselina Silvia Suhartinah [2], prediksi kelulusan mahasiswa dengan waktu studi menggunakan algoritma *naïve bayes* dan *C4.5* teknik klastering, analisis diskriminan, *teorema bayes*, *decision tree*, *artificial neural networks*, *support vector machine*, *regresi linear*, *support vector regresi*. Tujuan penelitian ini mencari dan menemukan pola yang terdapat pada data mahasiswa berdasarkan data NEM, IP DNS semester 1, IP DNS semester 2, IPK DNU semester 1-2, gaji orang tua dan pekerjaan orang tua, untuk memprediksi mahasiswa yang lulus atau tidak lulus sesuai dengan waktu studi dengan menggunakan algoritma *naïve bayes* dan *C4.5*, kemudian membandingkan hasil dan akurasi kedua algoritma tersebut diperoleh akurasi ketepatan hasil prediksi *naive bayes* adalah : $((17/21) \times 100\%) = 80,85\%$. Sementara Presentase kesalahan adalah : $((4/21) \times 100\%) = 19,05\%$. Akurasi ketepatan hasil prediksi *C4.5* adalah : $((18/21) \times 100\%) = 85,7\%$. Sedangkan nilai kesalahan pada penelitian dengan algoritma *C4.5* adalah : $((3/21) \times 100\%) = 14,3\%$. Disimpulkan, algoritma *decision tree* memiliki kompleksitas yang lebih besar, karena *entropi* masing-masing nilai digunakan untuk mencari ukuran *purity* dari atribut yang dinyatakan dengan *information gain* sehingga membentuk pola pohon keputusan.

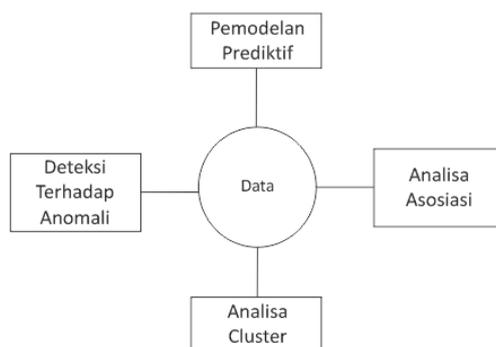
Muhamad Hanief Meninanda, dkk. [3] memprediksi masa studi sarjana dengan *Artificial Neural Network*. Variabel prediktor dari data akademis yang berpengaruh terhadap masa studi dan pembuatan model ANN untuk prediksi masa studi, model prediksi digunakan model *multiple regression* sebagai model pembanding. Data input 1289 mahasiswa dengan variabel ID mahasiswa, masa studi, kode mata kuliah, nama mata kuliah, pengambilan mata kuliah, nilai, bobot, sks, ipk, jumlah mata kuliah dan jumlah mengulang mata kuliah. Eksperimen penelitian menggunakan algoritma ANN *Multilayer Preceptron*, *Linier Regression* dan *Spearman Correlation*. Diperoleh uji beda menghasilkan nilai *p-value* sebesar 0.65 nilai tersebut lebih besar dari nilai kritis ($\alpha=0.05$) dengan tingkat kepercayaan 95% tidak ada perbedaan signifikan antara nilai masa studi aktual dengan nilai masa studi berdasarkan model prediksi.

2.2. Landasan Teori

2.2.1 Data mining

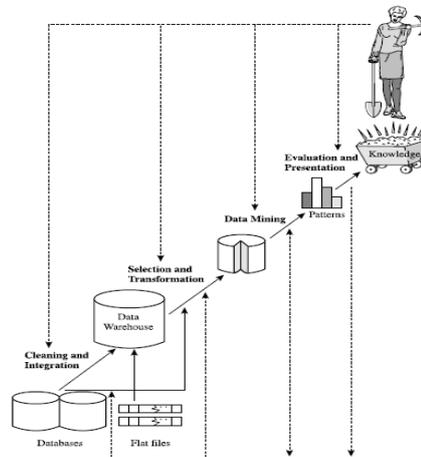
Data mining adalah proses yang menggunakan statistik, matematika, kecerdasan buatan, dan *machine learning* untuk mengekstraksi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai *database* besar. *Data mining* adalah analisis pengamatan *Dataset* untuk menemukan hubungan tak terduga dan untuk meringkas data dengan cara baru yang baik dimengerti dan berguna untuk pengguna data [7].

Inti dari tugas *data mining* adalah pemodelan prediktif, analisa asosiasi, analisa *cluster* dan deteksi terhadap anomali. Prediksi menurut Kamus Besar Bahasa Indonesia adalah “ramalan” atau “prakiraan” [8] Prediksi dalam penelitian ini adalah memperkirakan masa studi mahasiswa yang dinyatakan lulus tepat waktu dan mahasiswa yang dinyatakan lulus tidak tepat waktu dengan menggunakan sumber data pada transaksi aktifitas mahasiswa kuliah dan transaksi nilai.



Gambar 1. Empat Tugas Inti *Data Mining*

Data mining adalah suatu algoritma didalam menggali informasi yang tersembunyi pada suatu koleksi data (*database*). Analisa *data mining* berjalan pada data yang cenderung terus membesar untuk mendapatkan kesimpulan dan keputusan paling layak. *Datamining* memiliki beberapa sebutan atau nama lain yaitu: *Knowledge discovery(mining) indatabases* (KDD), ekstraksi pengetahuan (*knowledge extraction*), analisa data/pola, kecerdasan bisnis (*business intelligence*), dll [9]



Sumber : Data Minig “ Concepts and Techniques”, Jiawei Han and Micheline Kember, 2006.

Gambar 2. Proses dalam KDD

Knowledge Discovery in Database (KDD) yaitu sebuah proses mencari pengetahuan yang bermanfaat dari data. Tahapan proses dalam *data mining* secara garis besar dimulai dari data sumber dan berakhir dengan adanya informasi yang dihasilkan dari beberapa tahapan. Proses KDD secara garis besar sebagai berikut [9] :

- a. Seleksi Data
Pemilihan (seleksi) data baru dari sekumpulan data operasional perlu dilakukan sebelum tahap penggalian informasi dalam KDD dimulai. Data hasil seleksi yang akan digunakan untuk proses *data mining*, disimpan dalam suatu berkas, terpisah dari basis data operasional.
- b. Pembersihan data (*Cleaning*)
Sebelum proses *data mining* dapat dilaksanakan, perlu dilakukan proses pembersihan pada data yang menjadi fokus KDD. Proses pembersihan mencakup antara lain membuang *duplikasi data*, memeriksa data yang *inkonsisten*, dan memperbaiki kesalahan pada data, seperti kesalahan cetak (*tipografi*).
- c. Transformasi
Pada tahap transformasi data diubah kedalam bentuk yang sesuai untuk *dimining*. Beberapa teknik *data mining* membutuhkan format data yang khusus sebelum bisa diaplikasikan. Sebagai contoh, beberapa teknik standar seperti analisis asosiasi dan klastering hanya bisa menerima input data kategorikal. Disini juga dilakukan pemilihan data yang diperlukan oleh teknik *data mining* yang dipakai.
- d. *Data mining*
Data mining adalah proses mencari pola atau informasi menarik dalam data terpilih dengan menggunakan teknik atau metode tertentu. Teknik, metode, atau *algoritma* dalam *data mining* sangat bervariasi. Pemilihan metode atau algoritma yang tepat sangat bergantung pada tujuan dan proses KDD secara keseluruhan.
- e. Interpretasi/Evaluasi
Pola informasi yang dihasilkan dari proses *data mining* perlu ditampilkan dalam bentuk yang mudah dimengerti oleh pihak yang berkepentingan. Tahap ini merupakan bagian dari proses KDD

yang disebut dengan *interpretation*. Tahap ini mencakup pemeriksaan apakah pola atau informasi yang ditemukan bertentangan dengan fakta atau hipotesa yang ada sebelumnya?

2.2.2 Decision Tree C4.5

Algoritma C4.5 pengembangan dari algoritma ID3. Algoritma C4.5 dan ID3 diciptakan oleh *J. Rose quinlan* pada akhir tahun 1970-an. Algoritma C4.5 membuat pohon keputusan dari atas ke bawah, atribut paling atas merupakan akar, dan yang paling bawah dinamakan daun [10].

Decision tree adalah algoritma yang paling banyak digunakan untuk masalah klasifikasi. Sebuah *decision tree* terdiri dari beberapa simpul yaitu *tree's roo*, *internal nod* dan *leafs*. Konsep *entropi* digunakan untuk penentuan pada atribut mana sebuah pohon akan terbagi (*split*). Semakin tinggi *entropy* sebuah sampel, semakin tidak murni sampel tersebut [11].

Hasil kualitas keputusan yang didapatkan dari metode pohon keputusan sangat bergantung pada bagaimana pohon tersebut didesain. Sehingga jika pohon keputusan yang dibuat kurang optimal, maka akan berpengaruh pada kualitas dari keputusan yang didapat [12].

Penerapan *feature selection* pada penelitian ini adalah dengan menghitung *information gain* pada setiap atributnya. *Information gain* dari suatu atribut, diperoleh dari nilai *entropy* sebelum pemisahan dikurangi *entropy* setelah pemisahan [13].

$$Entropy(S) = \sum_{i=1}^k (P_i) \log_2(P_i) \dots \dots \dots (2)$$

P_i adalah proporsi data S dengan kelas i , dan k adalah jumlah kelas pada *output S*. Nilai *entropy* setelah pemisahan adalah sebagai berikut:

$$Entropy(S, A) = \sum_{i=1}^v \left(\frac{|S_v|}{S}\right) * Entropy(S_v) \dots \dots \dots (3)$$

Membuat pohon keputusan adalah memilih atribut yang harus diuji pada setiap simpul. Proses ini disebut *information gain*, yang berguna untuk menentukan atribut mana yang akan digunakan pada setiap simpul. *Information Gain* itu sendiri didapatkan dari perhitungan yang menggunakan satuan yang disebut *entropy* [14].

$$Gain(S, A) = Entropy(S) - Entropy(S, A) \dots \dots \dots (4)$$

2.2.3 Algoritma ID3

Algoritma ID3 atau *Iterative Dichotomiser 3* (ID3) merupakan sebuah metode yang digunakan untuk membangkitkan pohon keputusan Secara ringkas, langkah kerja Algoritma ID3 dapat digambarkan sebagai berikut [15].

$$[[Entropy(S) = -P_+ \log_2 P_+ - P_- \log_2 P_-]] \dots \dots \dots (5)$$

S = ruang (data) *sample* yang digunakan untuk *training*. P_+ = jumlah yang bersolusi positif (mendukung) pada data *sample* untuk kriteria tertentu.

$$Gain(S, A) = Entropy(S) - \sum \frac{S_v}{S} Entropy(S_v) \dots \dots \dots (6)$$

S = ruang (data) *sample* yang digunakan untuk *training*.

A = atribut.

V = suatu nilai yang mungkin untuk atribut A .

$Nilai(A)$ = himpunan yang mungkin untuk atribut A .

$|S_v|$ = jumlah sampel untuk nilai V

$|S|$ = jumlah seluruh sampel data.

$Entropy(S_v)$ = *entropy* untuk sampel yang memiliki nilai V

2.2.4 Algoritma Chaid

Metode *Chaid* digunakan untuk menduga variabel dependen berdasarkan variabel-variabel independen. Tahapan yang harus dilakukan dalam analisis *Chaid* untuk menghasilkan sebuah diagram pohon pada dasarnya melalui tiga tahap yaitu penggabungan, pemisahan, dan penghentian. Pada tahap penggabungan dibentuk tabel kontingensi dari variabel dependen dan variabel independen yang telah dikategorikan [16].

$$X^2_{hitung} = \sum_{r=1}^{m_r} \sum_{k=1}^{n_k} \frac{(o_{rk} - e_{rk})^2}{e_{rk}} \dots \dots \dots (7)$$

Keterangan: o_{rk} dan e_{rk} masing-masing adalah jumlah pengamatan dan nilai harapan pada baris ke- r dan kolom ke- k . Kriteria pengujian hipotesisnya adalah tolak H_0 jika $\chi^2_{hitung} > \chi^2_{tabel} = 0,05$ dengan derajat bebas adalah $(k-1)(r-1)$.

$$B = \binom{c-1}{g-1} = \frac{(c-1)!}{(g-1)!((c-1)-(g-1))!} \dots \dots \dots (8)$$

c adalah banyak kategori variabel independen awal ke- i ; g adalah banyak kategori variabel independen ke- i setelah penggabungan; dan i adalah $1, 2, \dots, n$. Apabila skala pengukuran variabelnya adalah skala nominal maka nilai pengali *Bonferroni* dihitung dengan rumus sebagai berikut.

$$B = \sum_{i=0}^{g-1} (-1)^i \frac{(g-1)^c}{i!(g-1)!} \dots \dots \dots (9)$$

Persamaan *Bonferroni* kemudian dihitung dengan rumus sebagai berikut.

$$\alpha = \frac{\pi}{B} \dots \dots \dots (10)$$

Keterangan: α adalah *comparison-wise error rate* (CWER), adalah *family-wide error rate* (FWER) dan B adalah pengali *Bonferroni*.

2.2.5 Pengukuran Kinerja

Permasalahan dalam klasifikasi, pengukuran yang biasa digunakan adalah *precision*, *recall* dan *accuracy* dapat dihitung dengan cara seperti berikut ini [17].

Tabel 1. Penilaian Pengukuran Kinerja

| | | |
|-------------------|-------------------------|-------------------|
| | Lulus Tidak Tepat Waktu | Lulus Tepat Waktu |
| Tidak Tepat Waktu | a | b |
| Tepat Waktu | c | d |

- a. *Precision*
Precision dapat dinyatakan sebagai kepersisan atau persamaan, seberapa persis dokumen tersebut untuk keperluan pencarian informasi. *Precision* adalah bagian data yang di ambil sesuai dengan informasi yang dibutuhkan.

$$Precision = \left(\frac{d}{b + d} \right) \times 100\% \dots \dots \dots (11)$$

- b. *Recall*
Recall adalah pengambilan data yang berhasil dilakukan terhadap bagian data yang relevan dengan *query*.

$$Recall = \left(\frac{d}{c + d} \right) \times 100\% \dots \dots \dots (12)$$

- c. *Accuracy*
Accuracy adalah persentase dari total data yang benar diidentifikasi.

$$Accuracy = \left(\frac{a + d}{totalmhs} \right) \times 100\% \dots \dots \dots (13)$$

2.2.6 n-Fold Cross Validation

n-fold cross validation atau *k-fold cross validation* merupakan salah satu metode yang digunakan untuk mengetahui rata-rata keberhasilan dari suatu sistem dengan cara melakukan perulangan dengan mengacak atribut masukan sehingga sistem tersebut teruji untuk beberapa atribut input yang acak. *n-foldcross validation* diawali dengan membagi data sejumlah *n-fold* yang diinginkan. Dalam proses *crossvalidation* data akan dibagi dalam *n* buah partisi dengan ukuran yang sama $D_1, D_2, D_3 \dots D_n$ selanjutnya proses *testing* dan *training* dilakukan sebanyak *n* kali. Dalam iterasi ke-*i* partisi D_i akan menjadi data *testing* dan sisanya akan menjadi *datatraining*. Untuk penggunaan jumlah *fold* terbaik untuk uji validitas, dianjurkan menggunakan *10-fold cross validation* dalam model [18]. *n-fold cross validation*, data dibagi sejumlah *n* dan data ke-*n* digunakan sebagai data *testing* sedangkan data selain data ke-*n* sebagai data pembentuk pola atau data *training*. Pada proses *fold* ke *n*, Parameter nilai *k* tertentu digunakan untuk menguji validitas data terhadap data *testing* menggunakan algoritma *C4.5*, *ID3* dan *Chaid*.

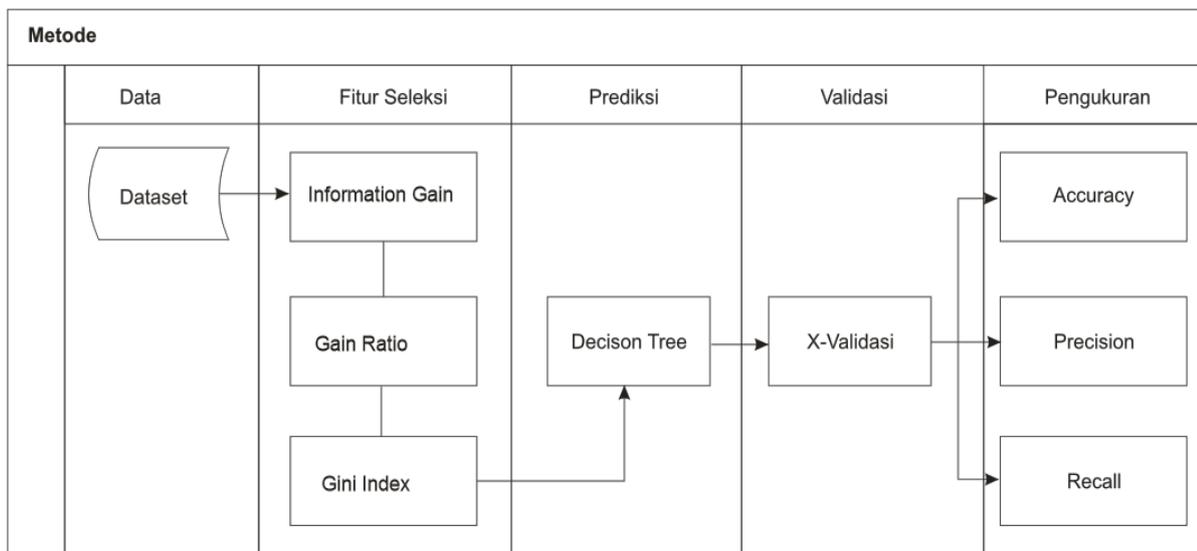
2.3. Prediksi Kelulusan Mahasiswa Tepat Waktu Berdasarkan Usia, Jenis Kelamin, dan Indeks Prestasi Menggunakan Algoritma Decision Tree

Kerangka pemikiran dari penelitian ini bersumber dari fenomena akademik pada STMIK Adhiguna yaitu tidak seimbangny jumlah mahasiswa yang masuk dan yang lulus pada perguruan tinggi sehingga perlunya sebuah solusi yang tepat agar terdapat kesinambungan proses akademik yang relevan dengan mengevaluasi transaksi aktifitas kuliah mahasiswa 4 (empat) semester pertama sebagai acuan pengolahan data diambil berdasarkan usia, jenis kelamin, IP Semester pada 4 (empat) semester pertama.

Untuk mendapatkan nilai akurasi dari masa studi mahasiswa berdasarkan atribut tersebut menggunakan Algoritma *Decision Tree C4.5* kemudian membandingkannya dengan algoritma lain *ID3* dan *Chaid* yang bertujuan menemukan pola yang terdapat pada data mahasiswa berdasarkan data usia, jenis kelamin, dan Indeks Pretasi Kumulatif (IPK) selama 4 (empat) semester pertama untuk memprediksi kelulusan mahasiswa tepat waktu.

3. METODE PENELITIAN

Jenis penelitian ini adalah eksperimen untuk prediksi kelulusan mahasiswa tepat waktu berdasarkan atribut usia, jenis kelamin, indeks prestasi pada empat (4) semester pertama. Tahapan penelitian menggunakan pendekatan teknik *data mining* dengan algoritma *Decision Tree C4.5* untuk memprediksi kelulusan mahasiswa sesuai dengan waktu yang ditentukan dan membandingkan nilai akurasinya dengan algoritma *ID3* dan *Chaid*. Diagram alir metode penelitian ini tercantum pada gambar berikut ini.



Gambar 3. Diagram Alir Metode Penelitian

3.1. Pengumpulan Data

Dataset penelitian ini diambil dari *database* EPSBED STMIK Adhiguna untuk menentukan kelulusan tepat waktu didefinisikan antara usia mahasiswa pada awal masuk perguruan tinggi, jenis kelamin mahasiswa serta perolehan indeks prestasi pada semester 1 (satu) sampai dengan semester 4 (empat).

3.2. Pengolahan Data Awal

Data target diimplementasikan dengan menggunakan *Tools Rapidminer 5.3* sebagai praproses pengolahan data dengan menggunakan algoritma fitur *Decision Tree C4.5*, *ID3* dan *Chaid* untuk mendapatkan fitur perbandingan yang paling tepat antara jarak data *testing* dan data *training*.

3.3. Eksperimen dan Pengujian

Eksperimen penelitian ini untuk mencari dan membandingkan algoritma seleksi atribut yang paling tepat yang akan dipergunakan sebagai masukan ke dalam algoritma *data mining*. Algoritma *data mining* yang digunakan adalah *Decision Tree C4.5*, *ID3* dan *Chaid*. Eksperimen pada tahap ini dengan menggunakan *Tools Rapidminer 5.3*, setelah itu eksperimen dilakukan dengan cara memilih nilai *k*. Parameter nilai *k* digunakan untuk menguji validitas data terhadap data *testing* menggunakan algoritma *C4.5* kemudian membandingkan dengan algoritma *ID3* dan *Chaid* guna mencapai nilai akurasi yang baik.

3.4. Evaluasi dan Validasi Penelitian

Hasil pengujian dari eksperimen berupa tabel akurasi prediksi kelulusan mahasiswa menggunakan *Decision Tree C4.5*, *ID3* dan *Chaid*. Dari data yang diolah menggunakan algoritma tersebut masing-masing mendapat tingkatan nilai akurasi yang berbeda untuk lebih jelasnya dapat dilihat pada Hasil dan Pembahasan.

4. HASIL PENELITIAN DAN PEMBAHASAN

Data mahasiswa diambil dari tahun angkatan 2009 untuk menentukan prediksi kelulusan tepat waktu. Terdapat 10 atribut salah satunya adalah label prediksi lulus tepat waktu dan lulus tidak tepat waktu. Data mahasiswa terdiri dari 390 data, terdiri dari 136 data lulus tetap waktu dan 389 *record* tidak tepat waktu. Adapun atribut data adalah sebagai berikut.

Tabel 2. Keterangan Atribut Data Set

| Nama Atribut | Keterangan |
|--------------|--|
| Nim | Nomor Induk Mahasiswa |
| Tgl. Lahir | Tanggal Lahir Mahasiswa |
| Usia | Usia Masuk Mahasiswa |
| JK | Jenis Kelamin Mahasiswa |
| IPS-1 | Indeks Prestasi pada Semester Satu (1) |
| IPS-2 | Indeks Prestasi pada Semester Dua (2) |
| IPS-3 | Indeks Prestasi pada Semester Tiga (3) |
| IPS-4 | Indeks Prestasi pada Semester Empat (4) |
| Label | Indikasi Lulus Tepat Waktu dan Lulus Tidak Tepat Waktu |

4.1. Hasil Pengolahan Data Awal

Data diperoleh dari *file database* mahasiswa yang akan dipakai dalam pengujian dalam penelitian ini, adapun datanya dapat dilihat pada tabel 3 adalah sebagai berikut :

Tabel 3. *Dataset* Awal Mahasiswa

| NO | NIM | TGL. LAHIR | USIA | JK | IPS1 | IPS2 | IPS3 | IPS4 |
|-----|------------|------------|------|----|------|------|------|------|
| 1 | 5520109002 | 07/04/1988 | 21 | L | 3,30 | 2,87 | 3,00 | 3,00 |
| 2 | 5520109003 | 02/08/1988 | 20 | L | 2,26 | 3,30 | 3,75 | 3,75 |
| 3 | 5520109005 | 28/02/1992 | 17 | L | 2,71 | 3,30 | 3,60 | 3,60 |
| 4 | 5520109006 | 08/09/1981 | 27 | L | 3,00 | 3,52 | 3,40 | 3,40 |
| 5 | 5520109007 | 28/08/1990 | 18 | L | 2,87 | 3,30 | 3,75 | 3,75 |
| 6 | 5520109008 | 30/01/1990 | 19 | L | 2,74 | 1,57 | 2,75 | 2,99 |
| 7 | 5520109009 | 31/05/1990 | 19 | L | 3,00 | 2,00 | 2,25 | 2,89 |
| 8 | 5520109010 | 09/06/1989 | 20 | L | 2,87 | 2,70 | 2,90 | 2,90 |
| 9 | 5520109011 | 15/06/1992 | 17 | L | 2,74 | 2,39 | 2,60 | 3,63 |
| 10 | 5520109012 | 02/02/1991 | 18 | L | 2,67 | 2,91 | 2,90 | 2,90 |
| 11 | 5520109013 | 07/12/1990 | 18 | P | 3,04 | 3,42 | 3,70 | 3,70 |
| 12 | 5520109014 | 24/02/1986 | 23 | P | 2,94 | 2,57 | 2,09 | 2,10 |
| 13 | 5520109015 | 11/11/1987 | 21 | L | 2,44 | 3,30 | 2,63 | 2,60 |
| 14 | 5520109016 | 19/04/1984 | 25 | L | 3,17 | 3,19 | 3,20 | 3,31 |
| 15 | 5520109018 | 06/06/1988 | 21 | L | 2,26 | 3,61 | 3,45 | 3,45 |
| 16 | 5520109019 | 26/11/1987 | 21 | P | 2,57 | 3,04 | 3,55 | 3,51 |
| 17 | 5520109020 | 01/09/1990 | 18 | L | 2,91 | 2,61 | 2,75 | 2,75 |
| 18 | 5520109022 | 16/11/1991 | 17 | L | 2,96 | 1,48 | 2,00 | 2,25 |
| 19 | 5520109023 | 26/07/1988 | 20 | L | 2,74 | 2,74 | 1,85 | 1,85 |
| 20 | 5520109024 | 15/08/1991 | 17 | L | 3,13 | 3,43 | 2,34 | 2,99 |
| . | . | . | . | . | . | . | . | . |
| 389 | 5520109031 | 10/06/1991 | 18 | L | 2,87 | 1,96 | 2,90 | 2,90 |

Pemilihan model disesuaikan dengan kebutuhan sesuai dengan tujuan penelitian. Data yang akan diprediksi lulus tepat waktu dan lulus tidak tepat waktu untuk angkatan 2009 yang berjumlah 389 *record*. Data *cleaning* digunakan untuk membersihkan nilai yang kosong atau *tuple* yang kosong (*missing values* dan *noise*). Tidak semua atribut dalam tabel data digunakan, hanya atribut-atribut yang dianggap berpengaruh saja yang akan digunakan untuk penelitian.

Tabel 4. Menentukan Label *Dataset* Mahasiswa

| USIA | jk | IPS-1 | IPS-2 | IPS-3 | IPS4 | LABEL |
|------|----|-------|-------|-------|------|-------------------|
| 21 | 1 | 3,30 | 2,87 | 3,00 | 3,00 | TEPAT WAKTU |
| 20 | 1 | 2,26 | 3,30 | 3,75 | 3,75 | TEPAT WAKTU |
| 17 | 1 | 2,71 | 3,30 | 3,60 | 3,60 | TEPAT WAKTU |
| 27 | 1 | 3,00 | 3,52 | 3,40 | 3,40 | TEPAT WAKTU |
| 18 | 1 | 2,87 | 3,30 | 3,75 | 3,75 | TEPAT WAKTU |
| 19 | 1 | 2,74 | 1,57 | 2,75 | 2,99 | TIDAK TEPAT WAKTU |
| 19 | 1 | 3,00 | 2,00 | 2,25 | 2,89 | TIDAK TEPAT WAKTU |
| 20 | 1 | 2,87 | 2,70 | 2,90 | 2,90 | TIDAK TEPAT WAKTU |
| 17 | 1 | 2,74 | 2,39 | 2,60 | 3,63 | TEPAT WAKTU |
| 18 | 1 | 2,67 | 2,91 | 2,90 | 2,90 | TIDAK TEPAT WAKTU |
| 18 | 0 | 3,04 | 3,42 | 3,70 | 3,70 | TEPAT WAKTU |
| 23 | 0 | 2,94 | 2,57 | 2,09 | 2,10 | TIDAK TEPAT WAKTU |
| 21 | 1 | 2,44 | 3,30 | 2,63 | 2,60 | TIDAK TEPAT WAKTU |
| 25 | 1 | 3,17 | 3,19 | 3,20 | 3,31 | TEPAT WAKTU |
| 21 | 1 | 2,26 | 3,61 | 3,45 | 3,45 | TEPAT WAKTU |
| 21 | 0 | 2,57 | 3,04 | 3,55 | 3,51 | TEPAT WAKTU |
| . | . | . | . | . | . | . |
| . | . | . | . | . | . | . |
| 18 | 1 | 2,52 | 2,56 | 3,05 | 3,05 | TEPAT WAKTU |
| 17 | 1 | 3,00 | 3,74 | 2,75 | 2,75 | TIDAK TEPAT WAKTU |

Pada Tabel 4 Atribut NIM, Nama Mahasiswa dan Tanggal Lahir Mahasiswa tidak disertakan lagi karena dianggap tidak memiliki pengaruh dalam eksperimen dalam penelitian ini.

4.2. Hasil Ekperimen dan Pengujian Model

Jumlah data yang digunakan berjumlah 389 *record*, 254 dinyatakan tidak tepat waktu dan 135 dinyatakan tepat waktu, maka :

$$I(S, A) = I(254,135) = -\frac{254}{389} \log_2 \frac{254}{389} - \frac{135}{389} \log_2 \frac{135}{389} = 0,931$$

Hasil dari *information gain* untuk *class* dinyatakan lulus tepat waktu dan lulus tidak tepat waktu sebesar 0,931. Nilai *entropy* untuk atribut usia, dengan data sampel adalah usia <=22 sebanyak 351 *class* lulus tepat waktu sebanyak 123 sampel dan tidak tepat waktu 228. Usia<=29 sebanyak 31 dengan sampel lulus tepat waktu sebanyak 9 dan tidak tepat waktu 22 sampel. Usia<=36 sebanyak 7 dengan sampel 3 dinyatakan lulus tepat waktu dan 4 sampel tidak tepat waktu. Sehingga nilai diperoleh adalah sebagai berikut.

$$Entropy(Usia) = \frac{351}{389} (0,977) + \frac{31}{389} (0,360) + \frac{7}{389} (0,122) = 0,912$$

Sehingga dapat diperoleh nilai gain yaitu :

$$Gain(Usia) = I(S, A) - E(Usia) = 0,931 - 0,912 = 0,019$$

Data sampel jenis kelamin (JK) diperoleh jenis kelamin laki-laki sebanyak 230, perempuan sebanyak 159 data, jenis kelamin laki-laki lulus tepat waktu sebanyak 76 sampel dan tidak tepat waktu 154 sampel. Jenis kelamin perempuan sebanyak 59 sampel tepat waktu dan sebanyak 100 sampel tidak tepat waktu, sehingga diperoleh nilai *information gain* adalah sebagai berikut :

$$Entropy(JK) = \frac{230}{389} (0,989) + \frac{159}{389} (0,916) = 0,960$$

Sehingga dapat diperoleh nilai gain yaitu :

$$Gain(JK) = I(S, A) - E(JK) = 0,931 - 0,960 = -0,028$$

Data sampel Indeks Prestasi Semester1 (IPS-1) diperoleh IP semester 2,75 – 3,00 sebanyak 332 sampel dengan data 106 sampel lulus tepat waktu dan 226 sampel tidak tepat waktu. IP semester 3,01 – 3,50 sebanyak 55 sampel dengan data 29 sampel tepat waktu dan 26 sampel tidak tepat waktu. IP semester 3,51 – 4,00 sebanyak 2 sampel tidak tepat waktu sedangkan untuk tidak tepat waktu tidak ada, sehingga diperoleh nilai *information gain* adalah sebagai berikut.

$$\begin{aligned} Entropy(IPS1) &= \frac{332}{389}(0,966) + \frac{55}{389}(0,414) + \frac{2}{389}(0,039) \\ &= 0,883 \end{aligned}$$

Sehingga dapat diperoleh nilai gain yaitu :

$$\begin{aligned} Gain(IPS1) &= I(S, A) - E(IPS1) \\ &= 0,931 - 0,883 \\ &= 0,048 \end{aligned}$$

Data sampel indeks prestasi (IPS-2) dapat diperoleh untuk IP semester 2,75 – 3,00 sebanyak 238 sampel dengan data 81 sampel lulus tepat waktu dan 157 sampel tidak tepat waktu. IP semester 3,01 – 3,50 sebanyak 119 sampel dengan data 44 sampel tepat waktu dan 75 sampel tidak tepat waktu. IP semester 3,51 – 4,00 sebanyak 32 sampel dengan data 10 sampel tepat waktu dan 22 sampel tidak tepat waktu, sehingga diperoleh nilai *information gain* adalah sebagai berikut :

$$\begin{aligned} Entropy(IPS2) &= \frac{238}{389}(1,000) + \frac{119}{389}(0,814) + \frac{32}{389}(0,370) \\ &= 0,891 \end{aligned}$$

Sehingga dapat diperoleh nilai gain yaitu :

$$\begin{aligned} Gain(IPS2) &= I(S, A) - E(IPS2) \\ &= 0,931 - 0,891 \\ &= 0,040 \end{aligned}$$

Data sampel indeks prestasi (IPS-3) dapat diperoleh untuk IP semester 2,75 – 3,00 sebanyak 271 sampel dengan data 55 sampel tepat waktu dan 216 sampel tidak tepat waktu. IP semester 3,01 – 3,50 sebanyak 99 sampel dengan data 63 sampel tepat waktu dan 36 sampel tidak tepat waktu. IP semester 3,51 – 4,00 sebanyak 19 sampel dengan data 17 sampel tepat waktu dan 2 sampel tidak tepat waktu, sehingga diperoleh nilai *information gain* adalah sebagai berikut :

$$\begin{aligned} Entropy(IPS3) &= \frac{271}{389}(0,870) + \frac{99}{389}(0,743) + \frac{19}{389}(0,236) \\ &= 0,807 \end{aligned}$$

Sehingga dapat diperoleh nilai gain yaitu :

$$\begin{aligned} Gain(IPS3) &= I(S, A) - E(IPS3) \\ &= 0,931 - 0,807 \\ &= 0,105 \end{aligned}$$

Data sampel indeks prestasi (IPS-4) diperoleh untuk IP semester 2,75 – 3,00 sebanyak 240 sampel dengan data 8 sampel tepat waktu dan 232 sampel tidak tepat waktu. IP semester 3,01 – 3,50 sebanyak 120 sampel dengan data 100 sampel tepat waktu dan 20 sampel tidak tepat waktu. IP semester 3,51 – 4,00 sebanyak 29 sampel dengan data 27 sampel tepat waktu dan 2 sampel tidak tepat waktu, sehingga diperoleh nilai *information gain* adalah sebagai berikut :

$$\begin{aligned} Entropy(IPS4) &= \frac{240}{389}(0,560) + \frac{120}{389}(0,724) + \frac{29}{389}(0,306) \\ &= 0,592 \end{aligned}$$

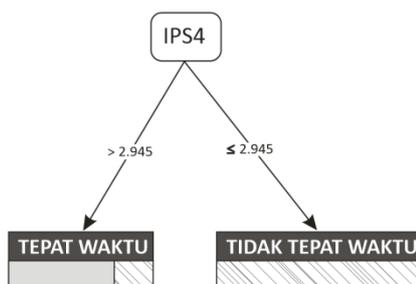
Sehingga dapat diperoleh nilai gain yaitu :

$$\begin{aligned} Gain(IPS4) &= I(S, A) - E(IPS4) \\ &= 0,931 - 0,592 \\ &= 0,340 \end{aligned}$$

Nilai *gain* dari masing – masing atribut diperoleh nilai gain yang tertinggi adalah sebagai berikut.

- Gain(Usia) = 0,019
- Gain(JK) = -0,028
- Gain(IPS1) = 0,048
- Gain(IPS2) = 0,040
- Gain(IPS3) = 0,105
- Gain(IPS4) = 0,340

Dari perhitungan nilai *information gain* pada semua atribut maka atribut yang memiliki nilai *gain* tertinggi adalah atribut yang layak menjadi *root*.



Parameter pilihan atribut yaitu usia, jenis kelamin, indeks prestasi semester 1 (satu) sampai dengan semester 4 (empat) serta atribut classnya mahasiswa lulus dengan tepat waktu dan tidak tepat waktu dengan *k-fold(number of validation)* menggunakan *range* $k=3$, $k=5$ dan $k=10$. Untuk mendapatkan nilai akurasi yang baik maka didapat diuji nilai *accurasi*, *precision* dan *recall*.

Tabel 5. Hasil Pengujian Algoritma C4.5 untuk Nilai *k-fold*

| k= | | Criterion | | |
|----|-----------|--------------|------------------|------------|
| | | gain ratio | information gain | gini index |
| 3 | Accuracy | 91,51 | 90,23 | 87,91 |
| | Precision | 90,06 | 88,72 | 86,91 |
| | Recall | 92,64 | 90,96 | 86,24 |
| 5 | accuracy | 90,75 | 91,00 | 88,94 |
| | precision | 89,30 | 89,54 | 87,62 |
| | recall | 92,22 | 92,25 | 88,24 |
| 10 | accuracy | 90,77 | 90,25 | 86,38 |
| | precision | 89,33 | 88,73 | 85,01 |
| | recall | 92,40 | 91,31 | 84,88 |

Akurasi tertinggi diperoleh pada $k=3$ yaitu sebesar 91,51% (*gain ratio*).

Tabel 6. Hasil Pengujian Algoritma ID3 untuk Nilai *k-fold*

| k= | | Criterion | | |
|----|-----------|--------------|------------------|--------------|
| | | gain ratio | information gain | gini index |
| 3 | accuracy | 40,36 | 35,47 | 35,47 |
| | precision | 68,39 | 57,45 | 57,45 |
| | recal | 54,33 | 50,42 | 50,42 |
| 5 | accuracy | 40,36 | 34,97 | 34,97 |
| | precision | 68,39 | 50,70 | 50,70 |
| | recal | 54,33 | 50,03 | 50,03 |
| 10 | accuracy | 40,37 | 35,22 | 35,22 |
| | precision | 68,39 | 50,70 | 50,70 |
| | recal | 54,33 | 50,05 | 50,05 |

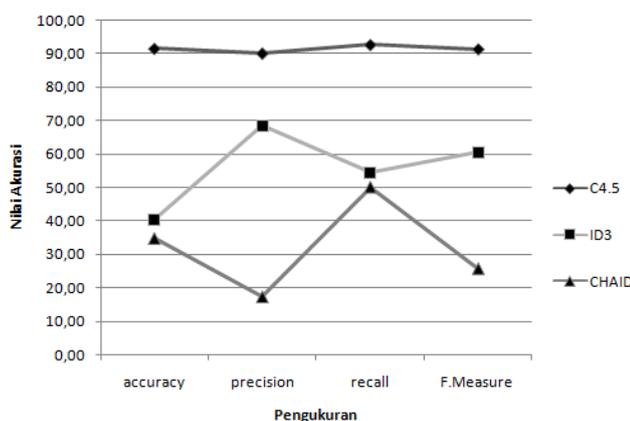
Nilai *Precision* dengan $k=3$ nilai *gain ratio* diperoleh 68,39, nilai pada *information gain* dan *gini index* diperoleh 57,45 dapat diartikan kecocokan data mahasiswa yang digunakan untuk nilai informasi lulus tepat waktu dan tidak tepat waktu.

Tabel 7. Hasil Pengujian Algoritma *Chaid* untuk Nilai *k-fold*

| k= | | Criterion | | |
|----|-----------|--------------|---------------|------------|
| | | minimal gain | maximal depth | confidence |
| 3 | | 0,10 | 20,00 | 0,25 |
| | accuracy | 34,70 | 17,35 | 50,00 |
| | precision | 34,70 | 17,35 | 50,00 |
| | recal | 34,70 | 17,35 | 50,00 |
| 5 | accuracy | 34,71 | 17,35 | 50,00 |
| | precision | 34,71 | 17,35 | 50,00 |
| | recal | 34,71 | 17,35 | 50,00 |
| 10 | accuracy | 34,70 | 17,35 | 50,00 |
| | precision | 34,70 | 17,35 | 50,00 |
| | recal | 34,70 | 17,35 | 50,00 |

Pada Tabel 7 nilai *Recall*, *Precision* dan *Accuracy* mendekati pada angka yang sama dalam artian nilai *recall* adalah proporsi jumlah dokumen yang dapat ditemukan kembali dan kecocokan jumlah dokumen yang ditemukan dan dianggap relevan menentukan pohon keputusan mahasiswa lulus tepat waktu dan tidak tepat waktu.

4.3. Perbandingan Model Algoritma



Gambar 4. Perbandingan *Accuracy* Algoritma *Decision Tree* *C4.5*, *ID3* dan *Chaid*

Nilai *recall* (proporsi jumlah dokumen) besar dapat terjadi jika jumlah dokumen yang diberikan juga besar kemungkinan juga nilai *precision*-nya semakin kecil. Nilai perbandingan *accuracy* pada tiap pengujian menggunakan algoritma *DecisionTree* *C4.5*, *ID3* dan *Chaid*, dari dapat disimpulkan algoritma *DecisionTree* *C4.5* memiliki kinerja yang lebih baik dibandingkan algoritma *ID3* dan *Chaid*.

Tabel 8. Perbandingan Akurasi Algoritma *Decision Tree* *C4.5* *ID3* dan *Chaid* dengan Kondisi *K-Fold* Berbeda.

| k = | C4.5 | ID3 | CHAID |
|-----|-------|-------|-------|
| 3 | 91,51 | 40,36 | 34,70 |
| 5 | 90,75 | 40,36 | 34,71 |
| 10 | 90,77 | 40,37 | 34,70 |

Perbandingan kinerja ketiga model algoritma berdasarkan jumlah data secara keseluruhan disimpulkan bahwa algoritma *Decision Tree* *C4.5* memiliki kinerja (*precision*, *recall*, dan *accuracy*) lebih baik dibandingkan dengan algoritma *ID3* dan *Chaid*.

5. KESIMPULAN DAN SARAN

5.1. Kesimpulan

Dari perhitungan nilai *information gain* pada atribut yang digunakan yaitu atribut Indeks Prestasi Semester 4 (IPS-4) mendapat nilai *gain* tertinggi yaitu 0,340 dengan demikian atribut tersebut layak menjadi *root*. Algoritma *Decision Tree C4.5* memiliki kompleksitas yang lebih besar, karena setiap nilai dalam suatu atribut ditelusuri dan diproses untuk mendapatkan *entropi* masing-masing nilai yang akan digunakan untuk mencari ukuran *purity* masing-masing atribut yang dinyatakan dengan *information gain*. Proses penelusuran ini akan membentuk sebuah pola berupa pohon keputusan.

5.2. Saran

Data *training* yang digunakan sebanyak 389 *record* data dan ketidaklengkapan data yang diperoleh penulis. Untuk melihat kinerja yang lebih baik dalam hasil akurasi masing-masing algoritma maka untuk penelitian selanjutnya jumlah *record* data yang digunakan untuk proses *training* sebaiknya ditingkatkan.

UCAPAN TERIMAKASIH

Penelitian ini dapat terselesaikan karena bantuan berbagai pihak, oleh karena itu peneliti berterimakasih kepada pihak-pihak yang mendukung terlaksananya penelitian yaitu para pembimbing penelitian, penguji,, serta pihak-pihak lain yang mendukung terlaksananya penelitian ini.

PERNYATAAN ORIGINALITAS

“Saya menyatakan dan bertanggung jawab dengan sebenarnya bahwa Artikel ini adalah hasil karya saya sendiri kecuali cuplikan dan ringkasan yang masing-masing telah saya jelaskan sumbernya” [Agus Romadhona – P31.2011.00964]

DAFTAR PUSTAKA

- [1] Peraturan Pemerintah Republik Indonesia Nomor 66 tahun 2010 tentang Perubahan atas Peraturan Pemerintah Nomor 17 tahun 2010 tentang Pengelolaan dan Penyelenggaraan Pendidikan.
- [2] Marselina SS, Ernastuti, “Aplikasi Algoritma Naive Bayes dan Algoritma C4.5 dalam Prediksi Kelulusan Mahasiswa Universitas Gunadarma” Universitas Gunadarma, Jakarta, 2010
- [3] Muhamad Hanief Meinanda, dkk.. “Prediksi masa studi sarjana dengan Artificial Neural Network”. Institut Teknologi Bandung, Bandung, 2009.
- [4] S. Anupama Kumar dan Vijayalakshmi M.N, ”Efficiency Of Decision Trees In Predicting Student’s Academic Performance.” R.V.College of Engineering, Bangalore , India, 2011
- [5] Lillyan Hadjaratie, “Jaringan Saraf Tiruan Untuk Prediksi Tingkat Kelulusan Mahasiswa Diploma Program Studi Manajemen Informatika Universitas Negeri Gorontalo.” Institut Pertanian Bogor, Bogor, 2011.
- [6] Sujana, “Aplikasi Mining Data Mahasiswa Dengan Metode Klasifikasi Decision Tree”, Universitas Widyatama, Yogyakarta, 2010.
- [7] Larose, Daniel T. “*Data mining* methods and models”. Department of Mathematical Sciences Central Connecticut State University, 2006.
- [8] Kamus Besar Bahasa Indonesia, <http://kamusbahasaIndonesia.org/prediksi>, diakses Tanggal 10 Bulan Oktober 2013.
- [9] Han Jiwei and Kamber Micheline, “*Data mining : Concepts and Techniques*”, Second Edition, by Elsevier Inc. All Rights Reserved, 2006.
- [10] Budanis Dwi Meilani Achmad dan Fauzi Slamet, “Klasifikasi Data Karyawan untuk Menentukan Jadwal Kerja Menggunakan Metode Decision Tree”, ITATS, 2012.
- [11] Dian Oktfia Dan D. L. Crispina Pardede, “Perbandingan Kinerja Algoritma Decision Tree Dan Naive Bayes Dalam Prediksi Kebangkrutan”, Jakarta.
- [12] Muhammad Hasby, “Penggunaan Pohon Keputusan Dalam Teori Keputusan” Institut Teknologi Bandung, Bandung.
- [13] Warih Maharani, “Klasifikasi Data Menggunakan Jst Backpropagation Momentum Dengan Adaptive Learning Rate Warih Maharani, Institut Teknologi Telkom, Bandung, 2009

- [14] Mohammad Noor. A, “ Penerapan Pohon Untuk Machine Learning”, Sekolah Teknik Elektro dan Informatika Institut Teknologi Bandung, Bandung.
- [15] Holisatul Munawaroh, “Perbandingan Algoritma Id3 Dan C5.0 Dalam Identifikasi Penjurusan Siswa SMA”, Universitas Trunojoyo
- [16] Yustisia Wirania, dkk., “ Pembentukan Pohon Klasifikasi dengan Metode *Chaid*”, UNTAN Pontianak, 2013.
- [17] Sofi Deviyanti, dkk., “Perbandingan Kinerja Algoritma ID3 Dan C4.5 Dalam Klasifikasi Spam-Mail”, Universitas Gunadarma, Jakarta.
- [18] Emerensye S. Y. Pandie, “Sistem Informasi Pengambilan Keputusan Pengajuan Kredit Dengan Algoritma K-Nearest Neighbour (Studi Kasus: Koperasi Simpan Pinjam)”, Universitas Diponegoro, Semarang, 2012.