Analisis Implementasi Sistem OLAP dan Klasifikasi Ketepatan Waktu Lulus dan Undur Diri Mahasiswa S1 Teknik Informatika Universitas Telkom Menggunakan Decision Tree C5.0

Jihan Ratnasari[1], Ibnu Asror, S.T., M.T.[2], Dr. Moch Arif Bijaksana, Ir. M.Tech[3]

[1,2,3]Fakultas Informatika, Universitas Telkom, Bandung
[1]jihanratnasari@students.telkomuniversity.ac.id, [2]iasror@telkomuniversity.ac.id,
[3]arifbijaksana@telkomuniversity.ac.id

Abstrak

Bagi suatu Universitas ketepatan lulus dan undur diri mahasiswa merupakan hal yang sangat penting karena sebagai patokan akreditasi. Universitas Telkom merupakan salah satu Universitas Swasta yang berada di Bandung yang merekomondasikan berbagai program studi, salah satunya S1 Teknik Informatika. Program studi S1 Teknik Informatika menjadikan ketepatan lulus dan undur diri mahasiswa sebagai patokan kesuksesan prodi. Namun faktanya pihak prodi masih kesulitan dalam menentukan pola tentang ketepatan waktu lulus dan undur diri mahasiswa dikarenakan tidak diimbangi informasi yang memadai.

Pada tugas akhir ini, membangun sebuah sistem menggunakan *OLAP (Online Analytical Processing)* berupa perancangan *data warehouse* dan *data mining* dengan metode klasifikasi dengan *algoritma C5.0* untuk menganalisis pola pada ketepatan waktu lulus dan undur diri mahasiswa. Selanjutnya hasil dari klasifikasi algoritma C5.0 dievaluasi dengan mempertimbangkan nilai *Precision, Recall, dan Micro Average F1-Score* untuk mendapatkan performansi sistem.

Hasil klasifikasi yang dievaluasi menggunakan dengan mempertimbangkan nilai *Precision, Recall, dan Micro Average F1-Score* untuk mengetahui nilai performansi. Berdasarkan performansi dari 2 pengujian yang pertama dengan menggunakan *k-fold cross validation* di dapat oleh 10-fold dengan nilai performansi 85%, dan pengujian kedua pada perubahan atribut untuk klasifikasi adalah penggunaan atribut data keseluuhan yang mendapatkan nilai akurasi 85%.

Kata kunci : ***data warehouse , Online Analytical Processing (OLAP), data mining, algoritma C5.0***

Abstract

For a university, graduating accuracy and student retirement is very important because as a benchmark of accreditation. Telkom University is one of the Private University located in Bandung that recommends various study programs, one of them is S1 Informatics Engineering. S1 program of Informatics Engineering makes graduation accuracy and student turn as a benchmark of success of study program. But in fact the prodi is still difficult in determining the pattern about the timeliness of graduation and student retreat due to not balanced with adequate information.

In this final project, build a system using OLAP (Online Analytical Processing) in the form of data warehouse and data mining with classification method with C5.0 algorithm to analyze the pattern on the timeliness of pass and the student retreat. The result of C5.0 algorithm classification is evaluated by considering Precision, Recall, and Micro Average F1-Score to get the system performance.

The classification results are evaluated using precision, Recall, and Micro Average F1-Score to determine the value of performance. Based on the performance of the first two tests using k-fold cross validation can be done 10 times with the performance value of 85%, and the tests performed for attributes are attribute data that has an accuracy value of 85%.

Keywords: data warehouse, Online Analytical Processing (OLAP), data mining, C5.0 algorithm

1. Pendahuluan

Latar Belakang

Seiring dengan perkembangan kemajuan teknologi informasi, pada suatu Perguruan Tinggi merupakan bagian terpenting terlebih kebutuhan akan informasi studi mahasiswa yang akurat sangat dibutuhkan oleh sebuah Universitas Telkom, khususnya pada Fakultas Informatika. Peran Prodi merupakan salah satu unsur bagian terpenting pada suatu Fakultas, seperti informasi perkuliahan, data mahasiswa, dan informasi data kelulusan mahasiswa. Sehingga informasi tersebut akan menjadi suatu elemen penting dalam keberhasilan fakultas. Namun penyajian kebutuhan informasi yang digunakan sebagai bahan evaluasi untuk analisis pola ketepatan waktu lulus sebagai penentuan strategi kedepannya, tidak diimbangi dengan penyajian infromasi yang memadai, sering kali informasi tersebut masih harus di gali ulang dengan wkatu yang cukup lama karena beberapa data kelulusan mahasiswa masih dikelola secara manual dari data yang jumlahnya sangat besar.

Pemanfaatan data yang ada di dalam sistem informasi tidak cukup jika hanya mengandalkan data operasional saja, diperlukan suatu analisis data untuk menggali suatu informasi pada studi mahasiswa. Untuk mengatasi permasalahan tersebut diperlukan suatu mekanisme pengolahan data, salah satunnya adalah dengan menggunakan teknologi OLAP *(Online Analitycal Processing).* Analisis data menggunakan OLAP dapat memberikan tingkatan analisis dengan kapabilitas *query* yang kompleks, *data mining* serta *reporting*. Pada pemilihan OLAP untuk mengetahui pola dari *data warehouse*.

Namun pada Universitas Telkom, khususnya Fakultas Teknik Informatika belum tersedianya pengelolaan dan analisis pada data studi akademik mahasiswa. Suatu perancangan yang mengimplementasikan sistem OLAP menggunakan *Decision Tree C5.0* untuk membantu penganalisisan data sehingga dapat menghasilkan informasi pola yang tepat , menghemat waktu, dan efisien. Karena dilihat dari kelebihan algoritma C5.0 itu sendiri dapat menganalisis *basis data* subtansial yang berisi banyak record dan field numerik dan nominal. Data yang digunakan untuk tugas akhir ini adalah data studi akademik mahasiswa yang terdapat pada Igracias.

Dengan demikian, diharapkan system tersebut dapat mempermudah pihak prodi dalam mengetahui pola studi mahasiswa dengan yang mana mahasiswa yang masa belajarnya tepat waktu, lebih dari 4 tahun, dan undur diri sehingga tujuan dari keberhasilan sebuah fakultas dapat tercapai.

Topik dan Batasannya

Dalam membangun sistem yang dapat menghasilkan informasi tentang ketepatan waktu lulus dan undur diri, dilakukan analisis dan beberapa tahap perancangan sebuah sistem menggunakan *OLAP (Online Analytical Processing)*. Metode OLAP itu sendiri dipergunakan pada data berjumlah besar dan menghasilkan bentuk output berupa analisis sehingga dapat digunakan untuk mendukung keputusan. Dengan demikian diharapkan sistem ini nantinya akan dapat menunjang kelancaran proses penganalisaan data, sehingga dapat menghemat waktu dan menghasilkan perhitungan yang lebih tepat. Setelah itu dianalisis kembali menggunakan *data mining* dengan metode klasifikasi dengan *algoritma C5.0*.

Batasan – batasan dalam penelitian tugas akhir ini diantaranya adalah :

1. Data yang digunakan adalah data studi akademik mahasiswa SI Teknik Informatika reguler yang berstatus alumni angkatan 2010 – 2017.
2. Hasil output pada data warehouse dan laporan yang mengenai hasil mahasiswa. Dan hasil proses data mining adalah model klasifikasi mengunakan algoritma C5.0.

Tujuan

Tujuan yang dicapai dari penelitian tugas akhir ini untuk membantu program studi dalam menyajikan informasi data mahasiswa dengan menerapkan sistem OLAP serta model klasifikasi dengan menggunakan algoritma C5.0 dengan menggunakan data studi akademik mahasiswa dari igracias yang berdasarkan masa kelulusan studinya apakah tepat waktu atau tidak.

Dan dalam evaluasi pengujian terhadap model yang terbentuk nantinya dilakukan dengan menghitung performansi dengan *Confusion Matrix* dan *K- Fold Cross Validation.* Pada *Confusion matrix* mempertimbangkan nilai *accuracy*, *precision*, *recall*, dan *f-measure*. Pengujian dilakukan untuk mengetahui apakah model tersebut cukup baik untuk menyelesaikan kasus ini.

Organisasi Tulisan

2. Studi Terkait

2.1 Data Warehouse

Dalam *data warehouse* dapat menggabungkan data dalam bentuk multidimensi. Pembangunan *data warehouse* meliputi pembersihan data, penyatuan data dan transformasi data dan dapat dilihat sebagai praproses yang penting untuk digunakan dalam *data mining*. Selain itu *data warehouse* mendukung *Online Analitycal Processing* (OLAP), sebuah kasus yang digunakan untuk menganalisis secara interaktif dari bentuk multidimensi yang mempunyai data yang rinci. *Data warehouse* juga dapat digubakan untuk mendukung pengambilan keputusan. [12[

Data warehouse adalah sebuah sistem yang mengambil dan menyatukan data secara periodik dari sistem sumber menuju ke penyimpanan data dimensional atau penyimpanan data normalisasi [Reinardi,2018]. Biasanya data yang tersimpan di dalamnya merupakan data sejarah (*history data*) yang digunakan untuk melakukan analisa untuk mendukung proses pengambilan keputusan. Selain itu data diperbarui secara berkelompok bukan setiap saat ketika proses transaksi berjalan pada sistem sumber. [2]

2.2 Data Mining

Data mining merupakan sebuah proses menganalisis sekumpulan data hasil penelitian, dengan tujuan untuk menemukan hubungan antar data, dan untuk meringkas data sehingga data menjadi mudah dimengerti dan berguna bagi pemilik data. [3] *Data mining* juga didefinisikan sebagai suatu komponen area yang menarik dari *machine learning* dan komputasi yang mampu beradaptasi. *Data mining* juga dapat didefinisikan sebagai analisis (besar) dari sekumpulan data pengamatan untuk menemukan hubungan yang tidak terduga dan menyimpulkan data dengan cara baru yang mana keduanya mudah dimengerti dan juga berguna bagi si pemilik data. [8]

Kumpulan data dalam jumlah besar membuat organisasi memiliki data dalam jumlah banyak, namun tanpa adanya pengolahan data, data tersebut belum dapat menghasilkan informasi ataupun *knowlegde* yang berguna. Oleh karena itu, data mining diperlukan untuk menemukan polapola dari data yang ada, meningkatkan nilai intrinsik data, dan mengubah data menjadi *knowledge.*

2.3 Online Analytical Processing (OLAP)

Menurut Han et al, *Online Analytical Processing* (OLAP) terdiri atas seperangkat *tool* untuk membantu proses analisis dan perbandingan data dalam *database*. Tool dan metode OLAP membantu pengguna menganalisis data pada sebuah *data warehouse* dengan menyediakan berbagai tampilan data, dan didukung dengan representasi data grafik yang dinamis. [3]

Yang menjadi inti dari beberapa sistem OLAP adalah konsep kubik OLAP atau biasa dinamakan kubik multidimensional atau *hypercube.*Kubik tersebup terdiri dari beberapa fakta yang disebut *measures* dimana dikategorikan oleh dimensi. Kubik metadata ini biasanya terbentuk dari sebuah skema star atau skema snowflake dari tabel relasi *database*. *Measures* berasal dari *record* dalam tabel fakta dan dimensi berasal dari tabel dimensi.

2.4 Algoritma C5

Algoritma C5.0 adalah salah satu algoritma yang terdapat dalam klasifikasi *data mining*, yang khususnya diterapkan pada teknik *decision tree* C5.0 merupakan penyempurnaan algoritma terdahulu ID3.

Menurut Ernawati, menjelaskan bahwa dalam algoritma C5.0, pemilihan atribut yang akan diproses menggunakan *information gain*. Secara heuristik akan dipilih atribut yang menghasilkan simpul yang paling bersih (*purest*). Jika dalam cabang suatu *decision tree* anggotanya berasal dari satu kelas maka cabang ini disebut *pure.*Kriteria yang digunakan adalah *information gain*. Jadi dalam memilih atribut

untuk memecah obyek dalam beberapa kelas harus kita pilih atribut yang menghasilkan *information gain* paling besar. [8]

Ukuran *information gain* digunakan untuk memilih atribut uji pada setiap *node* di dalam *tree*. Ukuran ini digunakan untuk memilih atribut atau *node* pada pohon. Atribut dengan nilai *informationgain* tertinggi akan terpilih sebagai *parent* bagi *node* selanjutnya. Algoritma ini membentuk pohon keputusan dengan cara pembagian dan menguasai sampel secara rekursif dari atas ke bawah. Algoritma ini dimulai dengan semua data yang dijadikan akar dari pohon keputusan sedangkan atribut yang dipilih akan menjadi pembagi bagi sampel tersebut. Pada pembentukan model klasifikasi, untuk mengklasifikasikan sampel yang digunakan maka diperlukan informasi dengan menggunakan formula 1. Setelah itu lakukan perhitungan untuk mendapatkan informasi nilai subset dari suatu atribut A dengan menggunakan formula 2. Kemudian untuk mengetahui nilai *information gain* dari atribut A, maka digunakan formula 3. Formula untuk *information g*ain [8].

$$I(S_1, S_2, \dots, S_m) \quad = \sum_{i=1}^{m} p_i log_2(p_i) \quad \dots\dots 1$$

$$E(\mathrm{A}) = \sum_{j=1}^{y} \frac{S_{1j}+\cdots+S_{mj}}{S} I\left(S_{1j\,,\dots.}\, S_{mj}\right) \dots\dots 2$$

$$\text{Gain (A)} = \text{I (S1, S2, … Sm)} - \text{E (A)} \dots\dots 3$$

dengan :
S : jumlah data sampel
m : jumlah sample pada kelas I
Pi : proporsi kelas dan diestimasikan dengan *si /s*
E(A) : nilai subset dari atribut A
S_j : sampel pada S yang bernilai A_j
A_{ij}: jumlah sampel pada kelas *Ci* dalam sebuah subset S_j
$\frac{S_{1j}+S_{mj}}{S}$: jumlah subset j yang dibagi dengan jumlah sampel S
Gain(A) : nilai information gain dari atribut A

Setelah itu lakukan terus menerus langkah diatas yaitu menghitung nilai tiap atribut berdasarkan nilai *information gain* yang tertinggi hingga semua *record* terpartisi. Proses dari *decision tree* ini akan berhenti jika semua *record* dalam simpul N mendapat kelas yang sama, tidak ada atribut di dalam *record* yang dipartisi lagi, dan tidak ada record di dalam cabang yang kosong.

2.5 Evaluasi dan Validasi

Pada pelitian tugas akhir ini dilakukan performansi menggunakan Confusion matrix dengan mempertimbangkan nilai *Precision, Recall, dan Micro Average F1-Score.*

Confusion matrix itu sendiri merupakan salah satu metode yang dapat digunakan untuk mengukur kinerja suatu metode klasifikasi. Pada dasarnya confusion matrix mengandung informasi yang membandingkan hasil klasifikasi yang dilakukan oleh sistem dengan hasil klasifikasi yang seharusnya. [13]

Pengukuran performansi menggunakan *Precision, Recall* dan *Micro Average F1-score.*

1. *Precision,* merupakan persentase dari item yang diprediksi benar dan terbukti benar. *Precision* digunakan untuk mengukur seberapa tepat kah prediksi suatu sistem dalam menyatakan prediksinya.

Persamaan :
$$Precision = \frac{TP}{TP + FP} \quad (1)$$

2. *Recall* adalah persentase dari item yang memang benar dan berhasil di prediksi benar. *Recall* digunakan untuk mengukur seberapa banyak item memang benar yang berhasil di prediksi benar.

Persamaan :
$$Recall = \frac{TP}{TP + FN} \quad (2)$$

3. *F1-Score* merupakan kombinasi pengukuran terhadap *Precision* dan *Recall* (*Harmonic Mean*). *F1-score* digunakan untuk mengukur kombinasi nilai yang telah dihasilkan dari *Precision* dan *Recall* sehingga menjadi satu nilai pengukuran.

Persamaan : $$F1 - Score = 2 \times \frac{Prec \times Rec}{Prec + Rec} \quad (3)$$

Tabel 2 -3 *Confusion Matrix*

		Predicted	
		Negative	Positive
Actual	Negative	TN	FP
	Positive	FN	TP

Pada pengukuran kinerja menggunakan *confusion matrix*, terdapat 4 (empat) istilah sebagai representasi hasil proses klasifikasi. [13] Keempat istilah tersebut adalah :

1. *True Positive* (TP) merupakan jumlah nilai yang diprediksi benar faktanya benar.
2. *True Negative* (TN) merupakan jumlah nilai yang diprediksi tidak benar faktanya tidak benar.
3. *False Positive* (FP) merupakan jumlah nilai yang diprediksi benar namun faktanya tidak benar.
4. *False Negative* (FN) merupakan jumlah nilai yang diprediksi tidak benar faktanya benar.

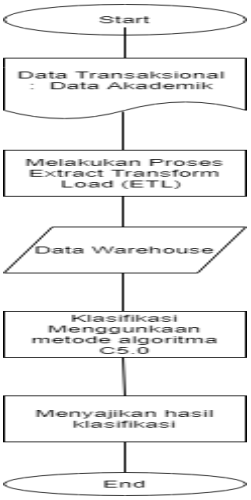
Selanjutnya pada validasi pengujian terhadap model sistem menggunakan *k-fold cross validation. K-Fold Cross Validation* itu sendiri merupakan salah satu metode yang digunakan untuk mengetahui rata-rata keberhasilan dari suatu sistem dengan cara melakukan perulangan dengan mengacak atribut masukan sehingga sistem tersebut teruji untuk beberapa atribut input yang acak. Sebuah dataset D secara acak dibagi menjadi subset D1, D2, …, Dk yang disebut *fold*. untuk k = jumlah *fold* digunakan. Setiap *fold*, kemudian digunakan sebagai data testing, sedangkan *fold* lainnya sebagai data training. Lakukan langkah tersebut sampai seluruh *fold* telah digunakan sebagai data uji. Pada penentu jumlah k nya dapat di atur sesuai yang diiinginkan. [14]

3. Sistem yang Dibangun

3.1 Gambaran Umum Perancangan Sistem

Tahap perancangan *system* ini merupakan tahap pengidentifisian dari kebutuhan – kebutuhan fungsional untuk perisiapan rancang membangun implementasi yang bertujuan untuk mendesain *system* dalam memenuhi kebutuhan pemakai *system*.

Sistem yang akan dibangun dapat digunakan untuk ketepatan waktu lulus dan undur diri mahasiswa S1 Teknik Informatika Universitas Telkom. Pertama dimulai dari pengumpulan data dari beberapa sumber data yang berhubungan dengan tepat waktu kelulusan dan undur diri mahasiswa pada Fakultas Informatika. Sumber data yang digunakan dalam penelitian ini adalah mengenai ketepatan waktu kelulusan antara lain data akademik mahasiswa, data nilai mahasiswa, dan rata- rata absensi. Setelah itu dilakukan proses *Extract Transform Load (ETL)* pada data yang telah dikumpulkan. Hasil dari ETL kemudian diintegrasikan dalam sebuah data warehouse dan teknik *Online Analyctical Process (OLAP)* yang dihasilkan analisis selanjutnya dilakukan penerapan algoritma *Decision Tree C5.0* untuk mengklasifikasi ketepatan waktu lulus dan undur diri mahasiswa.



Gambar 3.1 Gambaran Umum sistem

Menurut flowchart diatas, berikut merupakan penjelasan tahapan tersebut :

1. Mengumpulkan data transaksional, data yang digunakan merupakan data akademik mahasiswa S1 teknik Informatika Universitas Telkom yang diambil dari Igracias.
   Data tersebut diantaranya data akademik mahasiswa, data nilai mahasiswa , dan rata – rata presensi.
2. Melakukan ETL (Extract Transform Load) pada *data source*. Proses ETL ini merupakan mengetraksi data dari sistem sumber dan menyajikan data dalam berbagai bentuk untuk proses transformasi. Pada proses ini dilakukan pengubahan data ke dalam bentuk format yang berguna untuk proses transformasi dengan memilih atribut mana yang akan digunakan.
3. Kemudian hasil dari ETL masuk ke dalam *data warehouse*.
4. Setelah itu dari data warehouse kemudian dilakukan klasifikasi menggunakan *decision tree C5.0* untuk menentukan kelulusan tepat waktu dan undur diri mahasiswa. Yang menggunakan data dari data warehouse. Dan data pada *data warehouse* dijadikan data set untuk klasifikasi.

3.2 Perancangan Data Warehouse

Pada tahap perancangan ini meliputi perancangan sumber data, perancangan arsitektur *data warehouse* dan pemodelan data dimensional.

3.2.1 Perancangan Sumber Data

Pada tahap ini adalah menganalisis data transaksional yang akan digunakan sebagai sumber data pada data warehouse. Data yang digunakan diantaranya data studi akademik mahasiswa yang didapat dari Igracias. Data studi akademik mahasiswa, diantaranya data nilai, mata kuliah, rata-rata presensi.

3.2.2 Perancangan Arsitektur Data Warehouse

Selanjutnya dilakukan proses ETL setelah melakun perancangan sumber data *Extraction* merupakan langkah pertama pada proses ETL dimana pada proses ini dilakukan ekrtraski data dari sumber-sumber data. Pada proses ini dilakukan pengubahan data ke dalam suatu format yang berguna untuk proses transformasi. Yaitu dengan mengambil data dari data yang berada pada igracias yang berkaitan dengan akademik kemudian memilih atribut mana yang akan digunakan atau yang tidak digunakan.Setelah itu dilakukan proses *transform*, yaitu menggunakan serangkaian aturan atau fungsi untuk mengekstrak data dari sumber dan selanjutnya akan dimasukkan ke data warehouse. Berikut adalah hal-hal yang dilakukan dalam tahapan transformasi:

- Hanya memilih kolom tertentu saja untuk dimasukkan ke dalam data warehouse.
- Menyamakan karakter berupa kode *datasource* dengan kode pada *data warehouse.*
- Menggabungkan data secara bersama-sama dari berbagai sumber.

Setelah proses transform dilakukan, kemudian lanjut ke proses *load* data hasil *ektraksi* dan *transform* ke *data warehouse*. Fase *load* merupakan tahapan yang berfungsi untuk memasukkan data ke dalam target akhir yaitu ke dalam suatu data warehouse. Proses ETL (*Ekstrak Transform Load*) dalam pembangunan *data warehouse* merupakan proses yang penting karena menentukan pembangunan data warehouse selanjutnya. Proses *load* data dari *data source* ke *data warehouse* melalui proses ETL. Selanjutnya dilakukan pemodelan data dimensional. [12]

3.2.3 Pemodelan Data Dimensional

Pemodelan data dimensional Sembilan tahap tersebut adalah:

1. Pemilihan Proses
   Memperjelas batasan atau bisnis proses mengenai *data warehouse* yang dibuat berdasarkan proses.
   a) Ketepatan waktu lulus mahasiswa
      Merupakan proses ketepatan waktu lulus mahasiswa berdasarkan data akademik mahasiswa.
   b) Undur diri mahasiswa
      Merupakan proses undur diri mahasiswa berdasarkan data akademik mahasiswa.
   c) Tidak Tepat waktu lulus mahasiswa
      Merupakan proses tidak tepat waktu mahasiswa berdasarkan data akademik mahasiswa.
2. Pemilihan Grain
   Menentukan secara tepat apa yang direpresentasikan oleh record pada tabel fakta. Dalam tahap ini, analisis yang dilakukan meliputi jumlah sks, rata-rata presensi, jumlah mahasiswa tepat waktu, jumlah mahasiswa tidak tepat waktu, dan jumlah mahasiswa undur diri.
3. Identifikasi dan Penyesuain Dimensi
   penyesuaian dimensi dimensi dan grain dan ditampilkan dalam bentuk matriks. Dimensi yang digunakan adalah dimensi mahasiswa, dimensi matkul, dimensi status, dimensi tahun ajaran, dimensi nilai.
4. Pemilihan Fakta
   memilih fakta yang bisa digunakan dari grain.
   - Studi Mahasiswa
     Tabel fakta studi Mahasiswa meliputi rata-rata presensi id_mahasiswa, id_matakuliah, id_tahunajaran, id_nilai, id_status, id_IPS, id_total_SKS.
   - Kelulusan
     Tabel kelulusan meliputi id_tahunajaran, jumlah Mahasiiswa tepat waktu, jumlah mahasiswa tidak tepat waktu, jumlah mahasiswa undur diri/DO
5. Penyimpanan Pre-calculation pada tabel fakta
   Setelah fakta-fakta dipilih maka dilakukan pemeriksaan ulang untuk menentukan apakah fakta-fakta memungkinkan diterapkan untuk *pre-calculation* dan melakukan penyimpanan pada tabel fakta.
6. Memastikan tabel dimensi
   menambah gambaran dengan teks yang mudah dimengerti user terhadap dimensi yang memungkinkan.

Tabel 3.1 Rounding Out Dimension

Dimensi	Field
Mahasiswa	Id_mahasiswa, Nim, tahun masuk
Nilai	Id_nilai, konversi_nilai, indeks_nilai

Mata Kuliah	Id_matkul, kode_matkul, mata_kuliah, sks_matkul,
Status	Id_status, kode_status, status
Tahun Ajaran	Id_tahun, tahunajaran, semester

7. Pemilihan Durasi Database
   memilih durasi data histori yang dimiliki.

Tabel 3.2 Durasi Data

Nama data warehouse	Data yang masuk ke data warehouse	Data dalam data warehouse
Alumni2013	2010 - 2017	7 tahun

8. Melacak perubahan dari dimensi secara perlahan
   mengamati perubahan dari dimensi pada *table* dimensi. Dalam tahap ini dilakukan *update* data untuk menjaga konsistensi dan keakuratan data karena adanya atribut dari *table* yang memilihi nilai yang dapat berubah
9. Penentuan prioritas dan model
   Mempertimbangkan pengaruh dari rancangan fisik, seperti penyortiran urutan tabel fakta pada disk dan keberadaan dari penyimpanan awal ringkasan (*summaries)* atau penjumlahan (*aggregate*).

3.3 Perancangan C5.0 untuk Klasifikasi

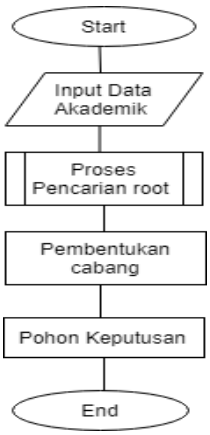
Dalam algoritma ini pemilihan atribut yang akan diproses menggunakan information gain. Dalam memilih atribut untuk pemecah obyek dalam beberapa kelas harus dipilih atribut yang menghasilkan information gain paling besar. Atribut dengan nilai information gain tertinggi akan dipilih sebagai parent bagi node selanjutnya. Perancangan C5.0 ini untuk menentukan pola dan menganalisis ketepatan waktu lulus dan undur diri mahasiswa berdasarkan kategori lulus tepat waktu, lulus tidak tepat waktu, dan undur diri. Kemudian dilakukan proses pencarian root dan pembentukan cabang. Hasil dari proses training berupa pohon keputusan yang disimpan dalam bentuk rule. Tahapan dalam membangun pohon keputusan dalam algoritma C5.0 adalah [16]

1. Mempersiapkan data training. Data training diambil dari data histori yang sudah dikelompokkan dalam kelas-kelas tertentu.
2. Proses klasifikasi data dimulai dengan adanya data *training* dan data *testing.*
3. Lakukan perhitungan terhadap *entropy* dan *information gain* pada setiap atribut
4. Buat Root node berdasarkan atribut yang memiliki information gain tertingi.
5. Hitung infornation gain setiap atribut untuk setiap cabang dari parent nodel, jika ada record di dalam cabang yang kosong dan jika ada atribut di dalam record yang dipartisi lain.
6. Setelah semua record dalam simpul mendapat kelas makan lanjut pembentukan model klasifikasi.

Flowchart dari proses training [16] :

a. Proses pencarian Root
   Proses pencaria *root* dilakukan pertama kali adalah menghitung nilai *entropy*. Selanjutnya dilakukan perhitungan gain untuk mengetahui nilai *root*. Nilai *gain* tertinggi pada masing-masing kategoru akan dijadikan nilai *root.*
b. Pembentukan Cabang

Setelah ditentukan nilai *root*, maka kategori dengan nilai *gain* tertinggi akan dijadikan dasar untuk menentukan pembentukan *node*. Kategori nilai *gain* tertinggi akan dijadikan *node* selanjutnya, kemudian akan dibentuk pohon *node*.



Gambar 3.2 Flowchart Proses Training

Proses awal dari *training* menggunakan algoritma C5.0 adalah pencarian *root*. Kemudian dihitung *information gain* setiap atribut dimana atribut yang memiliki nilai *information gain* tertinggi ditetapkan sebagai *root*. Hasil dari proses pencarian *root* adalah salah satu atribut sebagai *root*.

Di dalam proses pencarian *root*, terdapat proses perhitungan *entropy* dan perhitungan *information gain*. Mendapatkan nilai *entropy* dimulai dengan menghitung keseluruhan total status lulus tepat wkatu, luslus tidak tepat waktu, dan undur diri. Kemudian menghitung nilai *gain* masing-masing atribut. Hasil dari proses tersebut diperoleh nilai *information gain* dari setiap atribut.

Proses selanjutnya setelah pencarian *root* adalah proses pembentukan cabang. Dari hasil penentuan *root*, selanjutnya dilakukan kembali proses perhitungan nilai *entropy* dan gain untuk menentukan simpul selanjutnya. Data masukkan pada proses ini adalah atribut yang telah terpilih menajdi *root*. Untuk menentukan simpul selanjutnya yaitu dengan menghitung nilai *entropy* dan *gain* semua atribut berdasarkan *root* yang didapat sebelumnya. Proses tersebut akan berjalan rekrusif untuk menemukan *node* selanjutnya dan akan berhenti jika atribut sudah berada pada kelas yang sama atau *entropy* kelas bernilai nol dan pada kondisi tersebut leaf akan terbentuk. Hasil dari proses ini adalah pembentukan *decesion tree*. [16]

4. Evaluasi

4.1 Hasil Pengujian

Berikut bebarapa hasil pengujian skenario yang dilakukan terhadap algoritma C5.0.

Dalam tujuan dari pengujian ini untuk mengetahui performansi terbaik yang diperoleh dari model klasifikasi yang terbentuk dengan algoritma C5.0 dengan mempertimbangkan nilai precesion, recall, dan F1-score.

1. Skenario Pengujian Data pada Data Warehouse

Analisis data pada data warehouse ini, dilakukan analisis dengan membandingkan jumlah hasil data real dengan data warehouse.

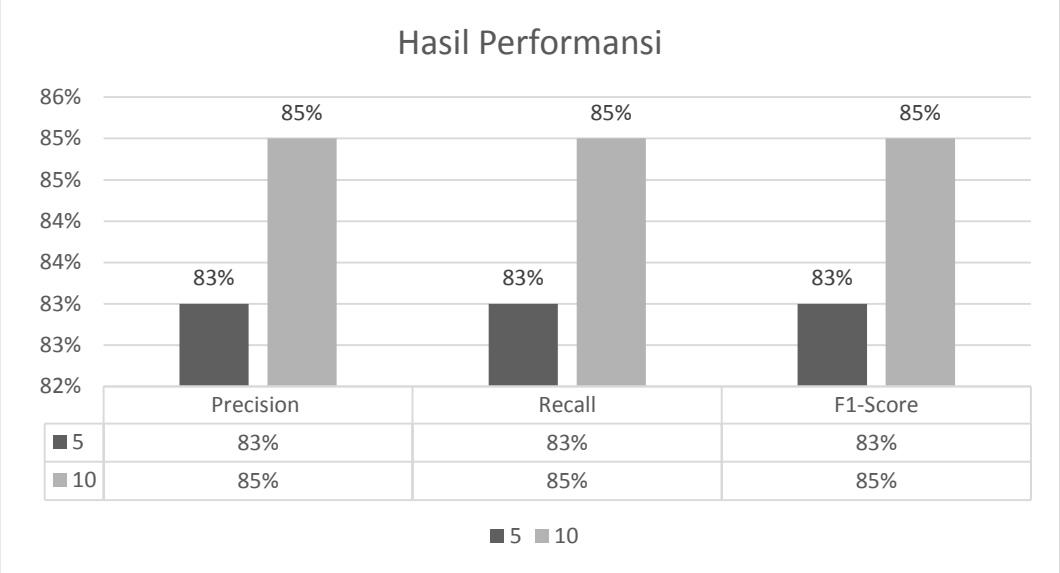
Tabel 4.1 Evaluasi Data Warehouse

	Data dari Data Real	Data dari Data Warehouse
Mahasiswa	1118	1118
Status	3	3

Tahun Ajaran	7	7
Mata Kuliah	159	159

2. Skenario Pengujian dengan Menggunakaan K-fold Cross Validation

Skenario 1, pada skenario 1 ini dilakukan pengujian untuk mengetahui nilai performansi yang diperoleh dengan algoritma C5.0. Disini menggunakan k-fold cross validation untuk melakukan persebaran data dan pembagian data testing dan training. Pada skenario ini dengan mempertimbangkan nilai *Precision, Recall, dan Micro Average F1-Score.* Dengan jumalah k foldnya 5 dan 10.

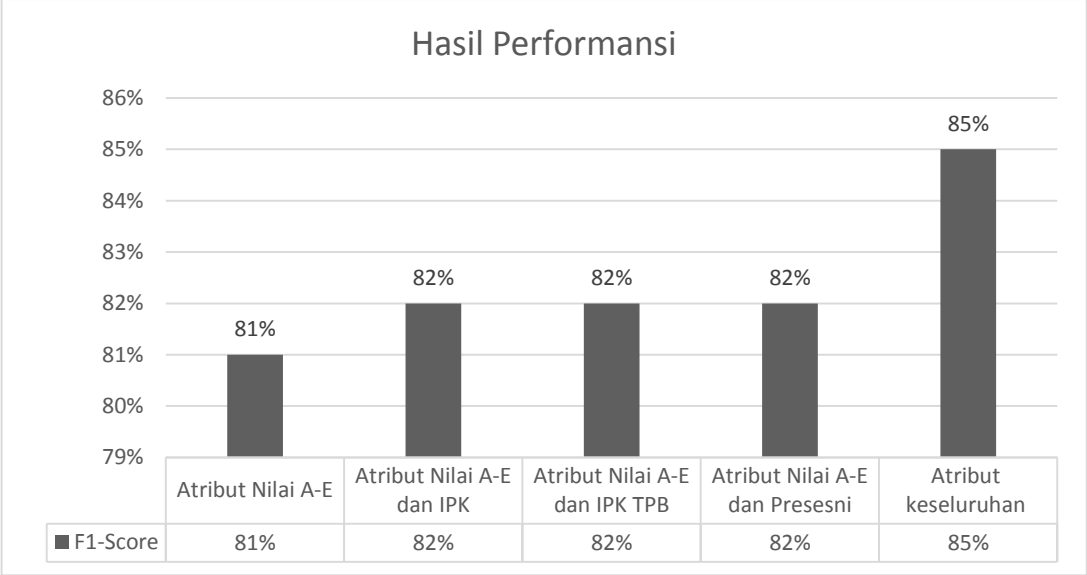


Gambar 4.1 Nilai F1-Score Skenario Pengujian K-Fold Cross Validation

3. Skenario Pengujian pada Perubahan Atribut untuk Klasifikasi

Skenario 3, pada skenario 3 dilakukan pengujian untuk mengetahui nilai performansi yang diperoleh dari algoritma C5.0. Skenario pengujian pada perubahan atribut untuk klasifikasi menggunakan :

a. Uji coba dengan menggunakan nilai mata kuliah A sampai E.
b. Uji coba dengan menggunakan nilai mata kuliah A sampai E dan atribut tambahan pendukung yaitu atribut IPK.
c. Uji coba dengan menggunakan nilai mata kuliah A sampai E dan atribut tambahan pendukung yaitu IPK TPB.
d. Uji coba dengan menggunakan nilai mata kuliah A sampai E dan atribut tambahan pendukung yaitu Presensi Mahasiswa.
e. Uji coba dengan menggunkan semua atribut keseluruhan.



Gambar 4.2 Nilai F1-Score Skenario Pengujian pada Perubahan Atribut untuk Klasifikasi

4.2 Anlisis pengujian

Berikut ini analsis pengujian yang telah dilakukan pada algoritma C5.0

1. Analisis pada skenario pengujian data pada data warehouse dengan data real. Data yang diujikan yaitu, data mahasiswa, status, tahun ajaran, dan mata kuliah. Dari semua data pada data warehouse, sesuai dengan data real.
2. Analisis pada skenario pengujian yang pertama dengan menggunakan *k-fold cross validation.* Pada gambar grafik diatas dapat dilihat yang menunjukkan dua perbandingan terhadap percobaan masing-masing k-fold. Dimulai dengan 5-fold terlihat bahwa hasil performansi yang mencapai nilai performansi 83% dengan rata-rata pada Precison, Recall, dan F1-score juga menghasilkan nilai performansi yang sama pula. Sedangkan pada 10-fold menghasilkan 85% nilai performansinya. Pada precision, recall, dan F1-score juga menghasilan nilai performansi yang sama. Maka pada saat jumlah k-fold diberikan nilai yang lebih besar, nilai performansinya akan menjadi lebih baik. Merujuk bahwa penentuan jumlah *k* dapat diatur serta sedikit atau banyaknya data juga perlu diperhatikan untuk penentuan jumlah *k* yang digunakan. Jika data yang dimiliki tidak terlalu besar maka jumlah *k* yang digunakan juga sebaiknya tidak terlalu besar dan begitu sebaliknya. Dengan adanya pengujian ini, dapat mengetahui apakah perserbaran datanya yang dipakai seimbang atau tidak. Dan dapat disimpulkan maka k: 10 fold menjadi nilai performansi terbaik dan nilai yang optimal pada skenario pengujian dengan menggunkan *k-fold cross validation* pada algoritma C5.0.

   Dan dalam skenario pengujian pada pemilihan k fold cross validation 5 dan 10, merujuk bahwa k 5 dan 10 merupakan nilai k fold yang paling umum dan sering digunakan untuk berbagai penelitian.[4]
2. Analisis pada skenario pengujian pada perubahan atribut untuk klasifikasi. Pada pengujian ini dilakukan dengan 5 data atribut yang berbeda , yaitu dengan antibut nilai a-e, atribut nilai A-E ditambah atribbut pendukung IPK, atribut nilai A-E ditambah atribbut pendukung IPK TPB, atribut nilai A-E ditambah atribut pendukung prresensi mahasiswa, dan atribut data keseluruhan. Yang pertama dengan menggunakan atribut nilai A-E mendapatkan nilai hasil performansi F1-Score 81%. Kedua atribut nilai A-E ditambah atribbut pendukung IPK mendapatkan nilai hasil performansi F-Score adalah 82%. Ketiga atribut nilai A-E ditambah atribbut pendukung IPK TPB mendapatkan nilai hasil performansi F-Score adalah 82%. Keempat atribut nilai A-E ditambah atribbut pendukung presensi mahasiswa mendapatkan nilai hasil performansi F-Score adalah 82%. Dan terakhir dengan atribut data keseluruhan menunjukkan nilai hasil performansi 85%.

   Dapat disimpulkan bawah hasil performansi pada perubahan atribut untuk klasifikasi yang paling maksimal adalah yang menggunakan atribut keseluruhan. Dengan merujuk pada kelebihan algoritma C5.0 dirancang untuk dapat menganalisis data yang lebih kompleks agar mendapatkan nilai akurasi yang maksimal.

5. Kesimpulan

Pada tugas akhir ini menghasilkan kesimpulan sebagai berikut :

1. Klasifikasi ketepatan waktu dan undur diri mahasiswa menggunakan algoritma C5.0 diawali dengan *data warehouse* kemudian dilakukan klasifikasi menggunakan *decision tree C5.0* untuk menentukan kelulusan tepat waktu dan undur diri mahasiswa. Yang menggunakan data dari data warehouse. Dan data pada data warehouse dijadikan data set untuk klasifikasi. Setelah itu pembentukan model klasifikasi dengan algoritma membagi sampel data berdasarkan atribut yang memiliki nilai information gain tertinggi.
2. Nilai performansi dari 2 skenario yang terbaik pada skenario pengujian yang pertama dengan menggunakan *k-fold cross validation* di dapat oleh 10-fold dengan nilai performansi 85%, dan pengujian kedua pada perubahan atribut untuk klasifikasi adalah penggunaan atribut data keseluuhan yang mendapatkan nilai akurasi 85%.
3. Adapun hasil rule dari klasifikasi ketepatan waktu dan undur diri dari mahasiswa adalah sebagai berikut :

   a. Lulus Tidak Tepat Waktu
   - Mahasiswa yang memiliki matakuliah_nilai_A ≦ 9 and matakuliah_nilai_B > 10 and presensi >66,1
   - Mahasiswa yang memiliki matakuliah_nilai_A > 9 and matakuliah_nilai_E ≦ 2 and IPK ≦ 3,01 and matauliah_nilai_AB and ≦ 11 and IPK_TPB ≦ 3,08 and matakuliah_nilaiD ≦ 7 and IPK ≦ 2,74
   - Mahasiswa yang memiliki matakuliah_nilai_A > 9 and matakuliah_nilai_E ≦ 2 and IPK > 3,01 and matakuliah_nilai_E >0 and matakuliah_nilai_B > 21
   - Mahasiswa yang memiliki matakuliah_nilai_A > 9 and matakuliah_nilai_E ≦ 2 and IPK ≦ 3,01 and matakuliah_nilai_AB ≦ 11 and IPK_TPB ≦ 3,08 and matakuliah_nilai_D ≦ 7 and IPK > 2,74 and matakuliah_nilai_A >15 and IPK ≦ 2,89 then Lulus Tidak Tepat Waktu.
   - Mahasiswa yang memiliki matakuliah_nilai_A > 9 and matakuliah_nilai_E 2 and matakuliah_nilai_E ≦ 9 and matakuliah_nilai_AB ≦ 14 then Lulus Tidak Teapat Waktu

   b. Lulus tepat waktu
   - Mahasiswa yang memiliki matakuliah_nilai_A > 9 and matakuliah_nilai_E ≦ 2 and IPK > 3,01 and matakuliah_nilai_E ≦ 0
   - Mahasiswa yang memiliki matakuliah_nilai_A > 9 and matakuliah_nilai_E ≦ 2 and IPK > 3,01 and matakuliah_nilai_E >0 and matakuliah_nilai_B ≦ 21
   - Mahasiswa yang memiliki matakuliah_nilai_A > 9 and matakuliah_nilai_E ≦ 2 and IPK ≦ 3,01 and matakuliah_nilai_AB ≦ 11 and IPK_TPB ≦ 3,08 and matakuliah_nilai_D ≦ 7 and IPK > 2,74 and matakuliah_nilai_A ≦ 15
   - Mahasiswa yang memiliki matakuliah_nilai_A > 9 and matakuliah_nilai_E >2 and matakuliah_nilai_E ≦ 9 and matakuliah_nilai_AB > 14

   c.DO/Undur Diri
   - Mahasiswa yang memiliki matakuliah_nilai_A ≦ 9 and matakuliah_nilai_B > 10 and presensi ≦ 66,1
   - Mahasiswa yang memiliki matakuliah_nilai_A ≦ 9 and matakuliah_nilai_B ≦ 10
   - Mahasiswa yang memiliki matakuliah_nilai_A > 9 and matakuliah_nilai_E > 2 and matakuliah_nilai_E > 9

Saran

Untuk pengembangan lebih lanjut, saran yang dapat diberikan :

1. Untuk pengembangan selanjutnya dapat menggunakan sampel data yang lebih banyak atau kompleks agar mendapatkan tingkat akurasi yang lebih tinggi dan lebih baik lagi.
2. Dalam penelitian selanjutnya dapat mengembangkan dengan aplikasi GUI untuk dapat menentukan dan menganalisis ketepatan waktu lulus dan undur diri mahasiswa sesuai dengan aturan atau rule model yang terbentuk.

Daftar Pustaka

[1]. Kimball, Ralph., Margy,Ross. 2002. The Data warehouse Toolkit. Canada : Willey Computer Publishing.
[2]. Supriyatna Adi, 2016. Sistem Analisis Data Mahasiswa Menggunakan Aplikasi Online Analytical Processing (OLAP) Data Warehouse. Karawang.
[3]. Suryanto, Wahyu D.Pengembangan Data warehouse dan aplikasi OLAP Data Tracer Study Alumni IPB Berbasis Web Menggunakan Microsoft Bussiness Intelligence. Bogor.
[4]. Han, J. And Kamber, M. 2006. Data Mining Concepts and Techniques Second Edition. Morgan Kauffman, San Fransisco.
[5]. Budi Santoso. 2013. Analisa Pemrosesan Data Secara Online OLAP Untuk Dunia Pendidikan. Yogyakarta.
[6]. Hidayanti Nutriana, 2012. PENTAHO SEBAGAI SOLUSI MASALAH PENGOLAHAN DATABASE (Pentaho as a Solution of Database Processing Problems).
[7]. Didi, Linda, Nita. Pemanfaatan OLAP untuk Analisis Penjualan Barang Pada CV.Maju Jaya Berkarya. Palembang.
[8]. Kurniarti, Dian, Marji. Penerapan Algoritma decision Tree C5.0 Untuk Peramalan Forex. Malang.
[9]. Holisatul, baim, Yeni. Perbandigan Algoritma ID3 dan C5.0 Dalam Indentifikasi Penjurusan Siswa SMA. Bangkalan
[10]. Rutvija And Jayati, 2015. C5.0 Algorithm to Improved Decision Tree with Feature Selection and Reduced Error Pruning.
[11]. Saeide, Taha, Mohammad, 2014. Prediction the Loyal Student Using Decision Tree Algorithms. Tehran, Iran.
[12]. Radtyo, Johan, Tony, 2008. Aplikasi Data Warehouse untuk Analisis Penjualan Mobil Berbasis Multidimensional Modeling (MDM) dan Star Schema Design. Banjarbaru.
[13]. Indriani Aida. Klasifikasi Data Forum dengan menggunakan Metode Naive Bayes Classifier. Tarakan.
[14]. Mutiara, Irwan, Andi, 2015. Penerapan K-Optimal Pada Algoritma Knn untuk Prediksi Kelulusan Tepat Waktu Mahasiswa Program Studi Ilmu Komputer Fmipa Unlam Berdasarkan IP Sampai Dengan Semester 5. Banjarbaru, Kalimantan selatan.
[15]. Yogi, Rian, Vivianne, 2009. Evaluasi Pemohon Kredit Mobil Di PT "X" Dengan Menggunakan Teknik Data Mining Decision Tree. Bandung.
[16]. Yusuf Yogi, 2007. Perbandingan Performansi Algoritma Decision Tree C5.0, CART, Dan CHAID : Kasus Prediksi Status Resiko Kredit Di Bank X. Bandung
[17]. Hadi Febri, 2017. Penerapn Data Mining Dalam Menganalisa Pemberian Pinjaman Dengan Menggunakan Metode Algoritma C5.0 (Studi Kasus:Koperasi Jasa Keuangan Syariah Kelurahan Lambung Bukik). Padang.
[18]. Rizal dan Ilham, 2017. Penerapan Algoritma C5.0 Pada Sistem Pendukung Keputusan Kelayakan Penerimaan Beras Masyarakat Miskin. Sukabumi.
[19]. Masykur Nuqson, 2010. Aplikasi Data Mining Untuk Menampilkan Informasi Tingkat Kelulusan Mahasiswa (studi Kasus di Fakultas MIPA Universitas Diponegoro). Semarang.
[20]. Maluque, 2015. Cross Validation. Scholar. Harvard
[21]. Bharat Rao, Glenn Fung, Rosales Romer. On the Dangers of Cross Validation An Experimental Evaluation. IKM CKS Siemens Medical Solutions. USA.
[22]. Ron Kohavi, 1995. A Study of Cross Validation and Bootstrap for Accuracy Estimation and Model Selection. Stanford University.