Analisis Implementasi Sistem OLAP dan Klasifikasi Ketepatan Waktu Lulus dan Undur Diri Mahasiswa Teknik Informatika Universitas Telkom Menggunakan *Random forest*

Pramudita Oktaviani[1], Ibnu Asror, S.T., M.T.[2], Dr. Moch. Arif Bijaksana[3]

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

Abstrak

Informasi kelulusan dan undur diri mahasiswa merupakan salah satu tolak ukur untuk mengevaluasi keberhasilan sebuah universitas. Begitu pula dengan program studi S1 Teknik Informatika, Universitas Telkom, yang memanfaatkan informasi kelulusan dan undur diri sebagai salah satu pendukung dalam kegiatan perencanaan dan evaluasi dalam mempertahankan kualitas kelulusan dan akreditasi program studi.
Pada kenyataanya, pihak prodi memiliki permasalahan dalam melakukan evaluasi kelulusan, dikarenakan prodi belum bisa mendapatkan informasi yang lengkap, cepat dan akurat, padahal setiap tahunnya permasalahan mengenai kelulusan yaitu jumlah mahasiswa lulus tidak tepat waktu yang lebih besar dibanding dengan jumlah mahasiswa yang lulus tepat waktu dapat mempengaruhi kualitas kelulusan dan akreditasi prodi.
Pada tugas akhir, dilakukan pembangunan sistem OLAP yang meliputi ektraksi data operasional ke dalam sebuah *data warehouse* untuk kemudian dilanjutkan dengan kegiatan analisis data menggunakan teknik klasifikasi *data mining* dengan *random forest* untuk menganalisis pola dari penyebab ketepatan waktu lulus dan undur diri mahasiswa.
Hasil klasifikasi dievaluasi menggunakan *micro average f1-score* untuk mengetahui performansi sistem. Berdasarkan data akademik yang digunakan untuk klasifikasi menggunakan *Random forest,* nilai *micro average f1-score* tertinggi yang diperoleh sebesar 77%.

Kata Kunci:*Data mining, random forest, Online Analytical Processing* (OLAP), *data warehouse.*

Abstract

Information graduation and student retirement is one of the benchmarks for a university. Similarly, S1 Informatics Engineering course, Telkom University, which uses existing information and one of the most effective in planning and evaluation activities in graduation and accreditation courses.
In fact, the parties that determine in the evaluation of graduation, because the program has not been able to get complete information, fast and accurate, whereas people who produce graduation ie the number of students is not in accordance with the time is greater than the number of students who can successfully affect the quality of graduation and accreditation of study program. Many factors that affect the timeliness of passing and not repeating are some of the grades of subjects, GPA, presence and more.
In the final stages, the OLAP system is developed which includes the extraction of operational data into a *data warehouse* and then proceeded by using *data mining* techniques with *random forest* to analyze the patterns of various timeliness of pass and retreat students. From the result of the grouping that has been done using student academic data of the most influential factor yait
The classification results are evaluated using the average micro-f1-score to determine the performance of the system. Based on the academic data used to use *Random forest,* the highest average micro value obtained is 77%.

Keywords :*Data mining, random forest, Online Analytical Processing* (OLAP), *data warehouse.*

1. Pendahuluan

Latar Belakang

Saat ini perkembangan teknologi informasi telah berkembang sangat pesat mengakibatkan ketersediaan dan keberagaman data pun semakin meningkat. Begitu pula yang dirasakan oleh pihak program studi Informatika, Universitas Telkom. Seiring dengan perkembangan teknologi informasi, pihak universitas dituntut untuk memiliki keunggulan bersaing dengan memanfaatkan semua sumber daya yang dimiliki. Pemanfaatan data yang ada tidak cukup hanya mengandalkan data operasional saja, tetapi diperlukan suatu analisis data untuk menggali informasi – informasi yang ada.

Pada kenyataannya pemanfaatan data infomasi saat ini belum dimanfaatkan secara maksimal dan efisien, pihak prodi masih merasa kesulitan dalam menganalisis evaluasi tingkat kelulusan mahasiswa dikarenakan data yang belum terintegrasi dalam sebuah basis data.Padahal pihak prodi perlu untuk melakukan evaluasi tingkat kelulusan untuk meningkatkan dan mempertahankan kualitas kelulusan dan akreditasi program studi, sehingga untuk memaksimalkan data informasi kelulusan diperlukan adanya integrasi data pada sebuah basis data yaitu *data warehouse* yang selanjutnya dari *data warehouse* tersebut dapat dilakukan penggalian informasi atau pola yang penting yang disebut dengan *data mining*. penggunaan teknik *data mining* diharapkan dapat memberikan pengetahuan-pengetahuan yang sebelumnya tersembunyi.

Pada penelitian tugas akhir ini penulis melakukakan penelitian terhadap data kelulusan salah satu program studi di universitas Telkom, yaitu Program studi S1 Teknik Informatika. Hal yang akan dilakukan dalam penelitian diantaranya adalah dilakukan pembangunan sistem OLAP yang meliputi ektraksi data operasional ke dalam sebuah *data warehouse* untuk kemudian dilanjutkan dengan kegiatan teknik *data mining* yaitu berupa klasifikasi ketepatan waktu lulus dan undur diri mahasiswa untuk mengetahui pola ketepatan waktu lulus mahasiswa.

Pemilihan OLAP dan *data warehouse* dikarenakan OLAP memiliki kemampuan memanipulasi data secara efisien dari berbagai pandangan sehingga analisis data dapat dilakukan dengan cepat dan mudah sedangkan pemilihan algoritma *random forest* sebagai algoritma klasifikasi pada bagian *data mining* dipilih dikarenakan *random forest* memiliki kemampuan dalam mengatasi *noise* dan *missing value*, dengan kemampuan tersebut algoritma *random forest* mampu mengatasi permasalahan data yang dijadikan dataset. Selain itu, *random forest* dapat mengatasi jumlah data dan fitur yang besar.

Dengan adanya penelitian tugas akhir ini diharapkan akan memberikan kemudahan bagi pihak prodi untuk mengetahui pola, penyebab ketepatan waktu lulus dan undur diri mahasiswa S1 Teknik Informatika.

Topik dan Batasannya

Pada penelitian tugas akhir ini untuk membangun sistem yang mampu menghasilkan informasi berupa hasil analisis dan kriteria yang mempengaruhi ketepatan waktu lulus dan undur diri mahasiswa diperlukan beberapa tahap diantaranya adalah *Online analytical processing* (OLAP) yaitu sebuah perangkat yang menggambarkan teknologi menggunakan visualisasi multidimensi sejumlah data untuk menyediakan akses yang lebih cepat bagi strategi informasi dengan tujuan mempercepat analisis [1]. Untuk membangun suatu sistem OLAP dibutuhkan pula membangun sebuah *data warehouse* yaitu data yang menampung data – data transaksional suatu institusi / perusahaan, pada penelitian ini data yang digunakan merupakan data akademik mahasiswa yang berasal dari *Igracias*. Data hasil OLAP kemudian dianalisis kembali mengunakan metode klasifikasi *data mining* dengan algoritma *random forest* untuk mendapatkan model klasifikasi berdasarkan data analisis yang telah didapat pada OLAP sebelumnya.

Adapun batasan – batasan dari pengerjaan penelitian tugas akhir ini diantaranya :

1. Data yang digunakan merupakan data akademik alumni mahasiswa S1 Teknik Informatika reguler berstatus alumni angkatan tahun 2010 – 2017.
2. Pembangunan *data warehouse* hanya meliputi data yang berkaitan dengan data akademik seperti mata kuliah yang diambil mahasiswa, rata-rata presensi dan nilai akademik .
3. Keluaran yang dihasilkan dari penelitian tugas akhir ini untuk proses analisis *data mining* merupakan model klasifikasi mahasiswa berdasarkan ketepatan waktu lulus dan undur diri.

Tujuan

Tujuan yang ingin dicapai dari penelitian tugas akhir ini adalah memberikan informasi dari hasil analisis data akademik mahasiswa yaitu berupa model klasifikasi mahasiswa berdasarkan ketepatan waktu lulus dan undur diri mahasiswa S1 Teknik Informatika, Universitas Telkom.

Organisasi Tulisan

2. Studi Terkait

2.1 *Online analytical processing* (OLAP)

Online analytical processing (OLAP) adalah sebuah perangkat yang menggambarkan teknologi menggunakan visualisasi multidimensi sejumlah data untuk menyediakan akses yang lebih cepat bagi strategi informasi dengan tujuan mempercepat analisis. Dalam model data OLAP, informasi di gambarkan secara konseptual seperti kubus *(cube),* yang terdiri atas kategori deskriptif (dimensions) dan nilai kuantitif *(measures).*[1]

2.2 *Data warehouse*

Data warehouse merupakan penyimpanan data yang berorientasi objek, terintegrasi, mempunyai *variant* waktu, dan menyimpan data dalam bentuk *nonvolatile* sebagai pendukung manejemen dalam proses pengambilan keputusan.

Data warehouse menyatukan dan 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 *On-line Analitycal Processing* (OLAP), sebuah kakas yang digunakan untuk menganalisis secara interaktif dari bentuk multidimensi yang mempunyai data yang rinci. Sehingga dapat memfasilitasi secara efektif *data generalization* dan *data mining*. [2]

2.3 *Data mining*

Data mining, sering juga disebut *knowledge discovery in database* (KDD), adalah kegiatan meliputi pengumpulan, pemakaian data historis untuk menemukan keteraturan, pola atau hubungan dalam set data berukuran besar. [8]

Secara garis besar *data mining* dapat dikelompokkan menjadi 2 kategori utama, yaitu :

a) *Descriptive mining*, yaitu proses untuk menemukan karakteristik penting dari data dalam suatu basis data. Teknik *data mining* yang termasuk dalam *descriptive mining* adalah *clustering, association*,dan *sequential mining*.

b) *Predictive*, yaitu proses untuk menemukan pola dari data dengan menggunakan beberapa variabel lain di masa depan. Salah satu teknik yang terdapat dalam *predictive mining* adalah klasifikasi.[10]

2.4 *Random forest*

Random forest merupakan pengembangan dari metode CART (*Classification and Regression* Tree) dengan menerapkan metode *boostrap aggregating (bagging)* dan *random feature selection.[6]*.

Pada Metode ini *classifier* yang digunakan hanya satu yaitu *decision tree*, yang mana pada setiap pembentukan tree akan dihasilkan banyak tree yang akan menghasilkan *output* sebanyak model yang dibangun. Dari sejumlah *output* yang dihasilkan akan dilakukan *voting* untuk mendapatkan satu *output*.

Pada *Random forest* dalam pembentukan tree digunakan teknik *impurty* berupa *gini index* adapun persamaan perhitungan *gini index* adalah sebagai berikut :

$$Gini\ (D) = 1 - \sum_{i=1}^{m} p_i^2 \qquad (1)$$

2.5 Evaluasi

Evaluasi merupakan tahapan yang dikerjakan dalam penelitian dengan tujuan untuk mengukur seberapa besar nilai performansi dari sistem yang dibuat. Pada evaluasi penelitian tugas akhir ini dilakukan evaluasi yang terdapat pada hasil klasifikasi terhadap algoritma yang digunakan menggunakan confusion matrix. Table 2.1 berikut merupakan ilustrasi dari confusion matrix.

Tabel 2-1 Confusion Matrix

		Predict Class	
		Yes	No
Actual Class	Yes	TP	FN
	No	FP	TN

Dengan :
TP = *True Positive* adalah jumlah yang diprediksi positif dan nilai yang sebenarnya juga positif.
TN =*True Negative* adalah jumlah yang diprediksi negatif dan nilai yang sebenarnya negatif.
FP = *False Positive* adalah jumlah yang diprediksi positif namun nilai yang sebenarnya negatif.
FN =*False Negative* adalah jumlah yang diprediksi negatif namun nilai yang sebenarnya positif.

Berdasarkan confusion matrix, pengukuran performansi menggunakan rumus *Precision, Recall* dan *Micro Average F1-score.*

1. *Precision*, mengevaluasi seberapa baik ketepatan jumlah prediksi sistem terhadap suatu kelas yang berhasil diprediksi dengan benar. Rumus perhitungan *precision* ditulis dengan persamaan

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

2. *Recall* adalah rumus yang digunakan untuk mengevaluasi seberapa baik ketepatan jumlah prediksi sistem terhadap suatu kelas yang berhasil diprediksi dengan benar dari total jumlah data yang memiliki label kelas tersebut. Rumus perhitungan *recall* ditulis dengan persamaan.

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

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

$$F1 - Score = 2 \times \frac{Prec \times Rec}{Prec + Rec} \quad (4)$$

3. Sistem yang Dibangun

Gambar 3.1 merupakan gambaran umum perancangan sistem yang akan dibangun.

Gambar 3.1 Flowchart Gambaran Umum Sistem

Berdasarkan Gambar 3.1 , berikut ini merupakan tahapan kerja yang akan dilakukan :

1. Menganalisis data dan mengumpulkan data transaksional. data yang digunakan merupakan data akademik alumni mahasiswa S1 Teknik Informatika Universitas Telkom angakatan 2010-2017 yang diambil dari *igracias*.
2. Dari data transaksional tersebut kemudian dilakukan proses *Extract Transform Load (ETL).* Proses ETL *(Extract, Transform, Load*) adalah sistem dasar dari *data warehouse*.

Pada perancangan ETL yang baik perlu dilakukan ekstraksi sumber data yang mengedepankan kualitas data dan standar yang konsisten, sehingga dapat diintegrasikan dan dapat memberikan format data untuk di representasikan.[11].

Pada penelitian ini data yang telah dianalisis dan dikumpulkan. Pada proses *Extract Transform Load (ETL)* ini dilakukan pengubahan data ke dalam suatu format yang berguna untuk proses transformasi dengan memilih atribut mana yang akan digunakan..

3. pada data yang telah dianalisis dan dikumpulkan. Pada proses *Extract Transform Load (ETL)* ini dilakukan pengubahan data ke dalam suatu format yang berguna untuk proses transformasi dengan memilih atribut mana yang akan digunakan.
4. Hasil dari proses *Extract Transform Load (ETL)* tersebut kemudian masuk ke *data warehouse* .
5. Selanjutnya dari data yang ada pada *data warehouse*, dilakukan pengambilan query yang dijadikan dataset klasifikasi.
6. Selanjutnya dilakukanlah klasifikasi menggunakan *random forest* untuk menentukan kelulusan tepat waktu dan undur diri mahasiswa menggunakan data yang berasal dari *data warehouse.* Data pada *data warehouse* tersebut dijadikan sebagai dataset untuk klasifikasi.

3.1 Perancangan *Data warehouse*

Pemodelan Data Dimensional

Pada tahap ini dilakukan pemodelan data dimensional, sebagai berikut :

1. Pemilihan proses

   Langkah awal dalam pemodelan data dimensional adalah memilih bisnis proses. Pada tugas akhir ini yang menjadi fokus utama penelitian yaitu berkaitan dengan akreditasi program studi mengenai kelulusan dan undur diri, berikut ini merupakan bisnis proses beserta deskripsinya yang akan menjadi fokus utama pada tugas akhir ini.

Table 3.2 Bisnis Proses

No	Bisnis Proses	Deskripsi
1.	Pendataan Kelulusan tepat waktu Mahasiswa	Merupakan proses pemantauan kelulusan tepat waktu berdasarkan data akademik mahasiswa
2.	Pendataan Mahasiswa Undur diri	Merupakan proses pemantaun undur diri berdasarkan data akademik mahasiswa

2. Pemilihan *Grain*

   Grain merupakan penampung *summary* data yang merupakan komponen untuk menghitung nilai dari setiap indikator mutu, adapun *grain* pada *data warehouse* mahasiswa ini meliputi: jumlah mahasiswa lulus tepat waktu, jumlah mahasiswa lulus tidak tepat waktu, jumlah mahasiswa DO / Undur diri, Total sks, IPS.
3. Pemilihan dimensi

   Pada tahap ini dilakukan penyesuaian dimensi untuk *data warehouse* mahasiswa.

   Dimensi yang digunakan diantaranya : dimensi mahasiswa, dimensi mata kuliah, dimensi tahun_ajaran, dimensi nilai dan dimensi status.
4. Pemilihan Fakta

   Selanjutnya adalah pemilihan tabel fakta yang dapat diperoleh dalam proses grain. tabel fakta yaitu tabel yang merepresentasikan *measure*, sebagai pusat data.
   - Studi Mahasiswa

     Tabel fakta studi Mahasiswa meliputi Id_Mahasiswa, id_Matakuliah, id_tahunajaran,id_nilai, id_status, IPS, Total_SKS,rata-rata presensi
   - Kelulusan

     Meliputi : Id_tahunAjaran , Jumlah Mahasiswa lulus tepat waktu , jumlah mahasiswa tidak tepat waktu, jumlah mahasiswa DO/undur diri.
5. Penyimpanan *Pre-Calculation* di tabel fakta

   Setelah dilakukan pemilihan tabel fakta , selanjutnya adalah dilakukan nya pengkajian ulang untuk menentukan apakah ada fakta – fakta yang dapat di terapkan pada kalkulasi awal.
6. Memastikan tabel dimensi

   Pada tabel dimensi ditambahkan gambaran teks terhadap dimensi yang memungkinkan.

Table 3.3 Rounding out dimention

Dimensi	Field
Mata Kuliah	Kode_mk, mata_kuliah, sks_mk, id_mk
Mahasiswa	Id_mahasiswa, NIM, Tahun_masuk
Nilai	Id_nilai, indeks_nilai, konversi_nilai
Status	Id_status, kode_status, status
TahunAjaran	Id_tahun, tahunajaran, semester

7. Pemilihan durasi database
   Pemilihan durasi data histori yang akan dibuat.

Table 3.4 Durasi data

Nama *data warehouse*	Data yang masuk ke *data warehouse*	Data dalam *data warehouse*
DWHMAHASISWA13	2010-2017	7 tahun

8. Melacak perubahan dari dimensi secara perlahan
   Memantau perubahan yang terjadi pada dimensi. Pada tahap ini dilakukan update data untuk menjaga keakuratan data karena adanya atribut dari tabel yang memiliki nilai yang dapat berubah.
9. Penentuan prioritas *Query* dan *Mode query*
   Mempertimbangkan pengaruh dari rancangan fisik, seperti penyortiran urutan tabel fakta atau penjumlahan.

3.2 Penerapan *data mining*

a. Seleksi data

Sumber data yang digunakan untuk klasifikasi ini berasal dari query yang diambil pada *data warehouse*, jumlah data yang diperoleh adalah sebanyak 1118 record data, alumni mahasiswa tahun 2010 sampai 2017 . Dataset alumnimahasiswa terdiri dari 12 atribut yang meliputi data akademik mahasiswa seperti NIM, mk_nilai_a, mk_nilai_ab, mk_nilai_b, mk_nilai_bc, mk_nilai_c, mk_nilai_d,mk_nilai_e, IPK,IPK_TPB,PRESENSI_MHS,STATUS.

b. Pre processing

- **Pembersihan data**

  Tahap kedua adalah proses pembersihan data yaitu melakukan pembersihan terhadap data yang *redundant dan missing value.* pada penelitian tugas akhir ini tahap pembersihan data tidak dilakukan dikarenakan pada tahap sebelumnya yaitu proses ETL telah dilakukan pembersihan data sehingga tidak ada *redundant* atau pun *missing value.*

c. Klasifikasi Menggunakan *Random forest*

Pada tahap ini dilakukan klasifikasi menggunakan metode *Random forest* dengan mengimplementasikan CART (*Classification and Regression Tree*) sebagai algoritma *decision tree*-nya..

Berikut ini adalah algoritma dari *random forest* [7] :

- Lakukan penarikan contoh acak berukuran *n* dengan pemulihan pada gugus data. Tahapan ini merupakan tahapan bootstrap.
- Dengan menggunakan contoh *bootstrap*, pohon dibangun sampai mencapai ukuran maksimum (tanpa pemangkasan). Pada setiap simpul, pemilihan pemilah dilakukan dengan memilih m peubah penjelas secara acak, dimana m << p. Pemilah terbaik dipilih dari *m* peubah penjelas tersebut. Tahapan ini adalah tahapan *random feature selection*.
- Ulangi langkah 1 dan 2 sebanyak k kali, sehingga terbentuk sebuah hutan yang terdiri atas k pohon.

4. Evaluasi

4.1 Hasil Pengujian

Pada Tugas akhir ini dilakukan 4 Skenario pengujian diantaranya :

a. Skenario pengujian proses ETL pada *data warehouse*

Berikut ini merupakan analisis data pada *data warehouse*, adapun analisis yang dilakukan adalah dengan membandingkan kesamaan hasil dari data pada *data warehouse* dengan data real yang berasal dari data source.

Table 4.1 Eavalusi *data warehouse*

	Data dari data source	Data dari *Data warehouse*
Mahasiswa	1118	1118
Mata Kuliah	159	159
Status	3	3
TahunAjaran	7	7

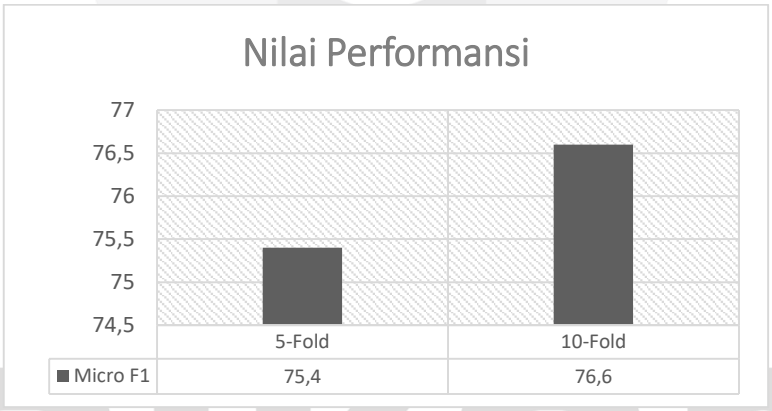
b. Skenario Pengujian *K-fold Cross Validation*

Pengujian pertama menggunakan K-*fold cross validation* dengan tujuan untuk melakukan persebaran data serta membagi data (*training & testing*) sebanyak k *segment*/*fold*. Untuk menunjukkan hasil performansi dari k-*fold cross validation* akan digunakan rata-rata nilai *micro average f1-score* dari setiap *fold*.

Adapaun Nilai K-fold yang dilakukan pengujian diantaranya :

- Uji coba 5-*fold cross validation*
- Uji coba 10-*fold cross validation*

Berikut merupakan hasil pengujian *k-fold cross validation* dengan algoritma *random forest* sebagai berikut :



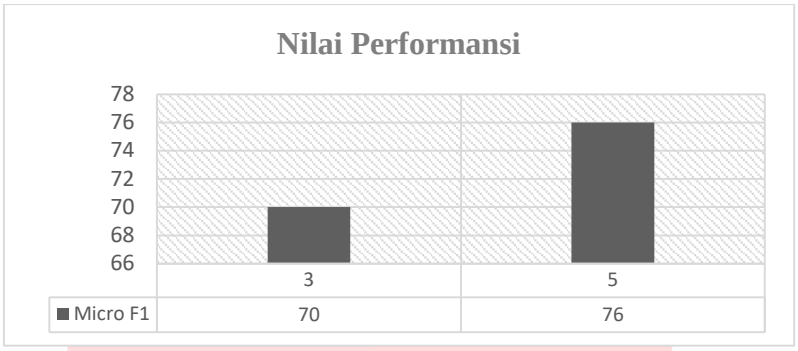
Gambar 4.1 Hasil Performansi k-Fold

c. Skenario Pengujian Perubahan Parameter jumlah K Tree Pada *Random forest*

Pada skenario dua ini akan dilakukan pengujian untuk mengetahui nilai performansi klasifikasi yang terbentuk dari *random forest* menggunakan Perubahan jumlah K tree, adapun jumlah K Tree yang dipakai sebagai pengujian adalan

- K = 3
- K = 5

Berikut hasil performansi dari skenario pengujian ketiga :

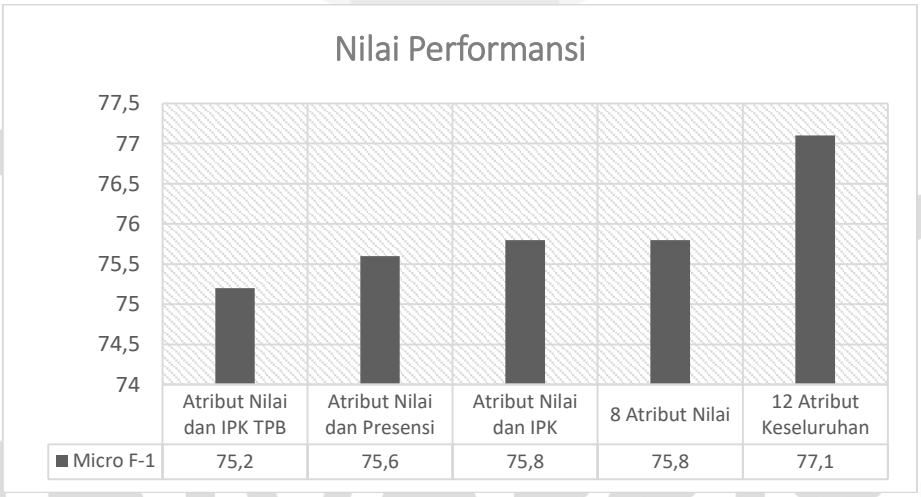


Gambar 4.2 Hasil Performansi Perubahan K tree *Random forest*

d. Skenario Pengujian berdasarkan pengaruh atribut untuk klasifikasi
Pada skenario pengujian ketiga ini, dilakukan pengujian untuk mengetahui nilai performansi yang diperoleh dari klasifikasi menggunakan *random forest* dengan menggunakan atribut untuk klasifikasi. Adapun langkah yang dilakukan untuk melakukan pengujian ini adalah :

- Melakukan uji coba dengan menggunakan 8 atribut dari 12 atribut yang ada pada dataset untuk dijadikan atribut klasifikasi, kedelapan atribut tersebut merupakan atribut berdasarkan nilai.
- Melakukan uji coba per atribut, diantaranya :
  a. Uji coba dengan menggunakan atribut nilai dengan atribut IPK yaitu atribut pendukung akademik
  b. Uji coba dengan menggunakan atribut nilai dengan atribut IPK TPB yaitu atribut pendukung akademik
  c. Uji coba dengan menggunakan atribut nilai dengan atribut presensi yaitu atribut pendukung akademik.
- Melakukan uji coba dengan menggunakan keseluruhan atribut yang ada pada dataset (12 atribut) untuk dijadikan atibut klasifikasi, kedua belas atribut diantaranya merupakan atribut nilai dan atribut pendukung akademik.

Berikut ini merupakan hasil performansi dari skenario yang telah dilakukan :



Gambar 4.3 Hasil Performansi Atribut

4.2 Analisis Hasil Pengujian

Berdasarkan skenario pengujian yang telah dilakukan berikut hasil yang didapatkan :

1. Analisis pada data
   Pada skenario pengujian pertama dilakukan analisis data pada *data warehouse*, pada hasil skenario yang telah dilakukan dihasilkan bahwa data *real* pada *data source* dan data pada *data warehouse* memiliki kesamaan data dan tidak ada ketimpangan data pada kedua data tersebut.
2. Analisis terhadap *K-fold cross validation*
   Pada skenario pengujian yang pertama dilakukan pengujian menggunakan *k-fold cross validation* dimulai dari pengujian dengan 5-*fold cross validation* yang menghasilkan nilai performansi dari rata-rata nilai *micro average f1-score* sebesar 75,4 %. Sedangkan untuk pengujian 10-*fold cross validation* dihasilkan nilai performansi rata-rata *micro average f1-score* sebesar 76,6 % . Pemilihan Nilai 5 dan 10 *fold cross validation* pada skenario pengujian ini merujuk bahwa 5 dan 10 *fold cross validation* merupakan nilai K yang paling sering digunakan dan umum pada sebuah penelitian. [2] Dilihat dari hasil nilai performansi maka Nilai K=10 inilah yang ditetapkan sebagai *k-Optimal* oleh peneliti, yang selanjutnya dapat digunakan untuk klasifikasi ketepatan waktu lulus dan undur diri mahasiswa.
3. Analisis perubahan jumlah K Tree
   Skenario pengujian yang kedua ini dilakukan pengujian mengunakan perubahan parameter *K Tree* yang dimulai dengan pengujian 3-*tree* yang menghasilkan nilai performansi dari rata rata *micro average f1-score* sebesar 70%, kemudian uji coba juga dilakukan pada 5-*tree* yang menghasilkan nilai performansi sebesar 75,5%.Skenario ini dilakukan untuk menguji seberapa berpengaruhnya jumlah K terhadap nilai performansi.Dari ketiga uji coba tersebut dilihat dari hasil nilai performansi nya maka 5-*tree* lah yang menjadi *K-tree* optimal.
4. Analisis terhadap perubahan atribut dataset klasifikasi
   Untuk pengujian ketiga ini, skenario yang dilakukan adalah melakukan pengujian dengan melakukan pengujian terhadap atribut nilai dengan atribut pendukung akademik yang dijadikan sebagai data klasifikasi. Dataset pertama untuk uji coba adalah atribut nilai dengan atribut IPK TPB yaitu atribut pendukung akademik dihasilkan performansi 75.2%, sedangkan untuk data ke-2 digunakan data yang berisi atribut nilai dengan atribut presensi yaitu atribut pendukung akademik dan dihasilkan performansi 75.6 %, untuk dataset ketiga uji coba dilakukan dengan menggunakan atribut nilai dan IPK dihasilkan performansi 75,8%, selanjutnya uji coba dengan menggunakan atribut yang berhubungan dengan nilai akademik mahasiswa yaitu berupa data yang berisi nilai mahasiswa yang berisi 8 atribut , dari dataset tersebut dihasilkan nilai performansi sebesar 75,8 %, dan untuk uji coba terakhir dilakukan ujicoba dengan seluruh atribut yang ada yaitu 12 atribut diantaranya atribut nilai dan atribut pendukung akademik, hasil uji coba tersebut dihasilkan nilai performansi sebesar 77.2%. Sehingga atribut optimal merupakan data yang memiliki 12 atribut, yaitu atribut yang berisi nilai dan juga atribut pendukung akademik mahasiswa.

5. Kesimpulan.

- Berdasarkan hasil pengujian yang telah dilakukan maka dapat disimpulkan bahwa *10-fold cross validation, 5-tree*,dengan 12 atribut merupakan hasil optimal dari pengujian yang telah dilakukan dengan menggunakan *random forest,* didapatkan pula nilai performansi berdasarkan hasil pengujian menggunakan *micro average f1-score* sebesar 77 %.
- Dari hasil klasifikasi yang telah dilakukan menggunakan 12 atribut dari data akademik mahasiswa yang lulus tidak tepat waktu, lulus tepat waktu dan undur diri didapatkan *root node* yaitu atribut Mk nilai A. Adapun hasil sebagian *rule* klasifikasi dari ketepatan waktu lulus dan undur diri mahasiswa adalah sebagai berikut:
  a. Lulus Tepat Waktu
     - Mahasiswa yang memiliki jumlah Mk_Nilai_A ≥ 8 dan jumlah mk_nilai_c ≤ 15
     - Mahasiswa yang memiliki Ipk ≥ 2.73 dan Mk_Nilai_B ≤ 15
     - Mahasiswa yang memiliki presensi ≤ 88,41

  b. Lulus tidak Tepat waktu
     - Mahasiswa yang memiliki ipk ≤ 3.25 and presensi_Mhs ≤ 85
     - Mahasiswa yang memiliki jumlah Mk_Nilai_B ≤ 17 dan Presensi_Mhs ≤ 85,73 dan Mk_Nilai_A ≤ 22

  c. DO / Undur diri
     - Mahasiswa yang memiliki Ipk TPB ≤ 2.7
     - Mahasiswa yang memiliki jumlah Mk Nilai E ≥ 4
     - Mahasiswa yang memiliki Mk_Nilai_B ≤ 15 dan 3 dan Mk_Nilai_D ≤ 8 dan Mk_Nilai_A ≤ 12 And Presensi_Mhs ≤ 67

- Untuk saran pada penelitian selanjutnya diharapkan menambahkan atribut lain, baik atribut akademik yang belum ada pada penelitian ini ataupun atribut non-akademik yang dapat mempengaruhi ketepatanwaktu lulus dan undur diri mahasiswa.

Daftar Pustaka

[1] Reddy, G.S., Srinivasu, R., Rao, M.P.C., Rikkula, S.R., 2010. International Journal on Computer Science and Engineering.Data Warehousing, *Data mining*, OLAP and OLTP Technologies areEssential Elements to SupportDecision-Making Process inIndustries.

[2] Han, J. And Kamber, M. 2006. *Data mining* Concepts and Techniques Second Edition. Morgan Kauffman, San Fransisco.

[3] Hidayanti Nutriana, 2012. PENTAHO SEBAGAI SOLUSI MASALAH PENGOLAHAN DATABASE (Pentaho as a Solution of Database Processing Problems).

[4] Kimball, Ralp., Margy, Ross. The Data warehousing Toolkit. Canada: Willey Computer Publishing.

[5] Suzana Meta, Jemakmun. Suyanto, 2013. Analisis dan Perancangan *Data warehouse* Rumah Sakit Umum Daerah Palembang Bari.

[6] Brieman L, 2001 , *Random forest*s, Machine Learning.

[7] Breiman L. 1996. *Bagging Predictors*. Machine Learning 24, 123-14

[8] B. Santosa, 2007 ,"*Data mining* Teknik Pemanfaata Data untuk Keperluan Bisnis," Yogyakarta: Graha Ilmu.

[9] Hastie, T., Tibshirani, R. and Friedman, J., 2001. The elements of statistical learning. *NY Springer*.

[10] Tan S, Kumar P, Steinbach M. 2006. "*Introduction To Data mining"*. Addison Wesley

[11] Dharayani R , Laksitowening K, Yanuarfiani A, Implementasi ETL (*Extract, Transform, Load*) Pangkalan Data Perguruan Tinggi dengan Menggunakan *State-Space Problem*