The Study of Decision Tree Algorithm For Classification of Student Graduation (Case Study : Faculty of Economics, University of Garut)

Fikri Fahru Roji[1*], Rahmawati[2], Dendi Ramdani[1]
[1]Program Studi Bisnis Digital, Fakultas Ekonomi, Universitas Garut,
Jl. Hampor No.52A, Mekarwangi, Garut, Jawa Barat 44151, Indonesia
[2,3]Program Studi Sistem Informasi, Fakultas Ilmu Terapan Dan Sains, Institut Pendidikan Indonesia, Jl. Terusan Pahlawan No.32, Sukagalih, Garut, Jawa Barat 44151, Indonesia
*Penulis koresponden, *e-mail : fikrifahruroji@uniga.ac.id*

***Abstract*:** *The academic performance is one of aspect which has remained the benchmark of the success in learning activities at the university. The indicator of academic performance in the university is the students able to complete their studies on time. Unfortunately, the problem regarding academic performance was associated with the completion time of student studies in Faculty of Economics,University of Garut. In this research explore the model that able to classify the graduation of student through the data mining classification technique by Decision Tree Algorithm. The classification conducted by evaluating the academic performance based on Semester Performance Index (IPS) and Semester Credit Unit (SKS) during two years in the beginning and use the demographics of students as attributes that will be used in the dataset. Based on the examinations that conducted by using k-fold cross validation, there are 8 attributes that influence the graduation of students. The model that represented able to applied for classify the active students in 2016 and 2017 that indicate accuracy value of 79.53% and recall value was 19.23%.*

Keywords: *data mining; classifications; student graduation; decision tree; k-fold cross validation.*

Abstrak: Kinerja akademik menjadi salah satu dari sekian banyak aspek yang menjadi tolak ukur keberhasilan dari sebuah perguruan tinggi dalam menyelenggarakan kegiatan pembelajaran. Salah satu indikator kinerja akademik yang baik bagi suatu perguruan tinggi adalah mahasiswa dapat menyelesaikan masa studi dengan batas minimal waktu yang telah ditentukan, atau bisa dikatakan mahasiswa bisa lulus dengan tepat waktu. Permasalahan mengenai kinerja akademik mahasiswa dalam hal menyelesaikan masa studi secara tepat waktu masih banyak ditemui di Fakultas Ekonomi Universitas Garut. Dalam penelitian ini akan dilakukan pencarian suatu model yang bisa mengklasifikasi pola kelulusan mahasiswa dengan menggunakan teknik Data Mining dengan algoritma Decision Tree. Klasifikasi digunakan dengan cara mengevaluasi hasil kinerja akademik mahasiswa berupa IPS dan jumlah SKS pada tahun pertama dan kedua serta menggunakan demografi mahasiswa sebagai atribut yang akan digunakan dalam dataset. Berdasarkan beberapa skenario uji dengan menggunakan metode k-fold cross validation, terdapat 8 atribut yang mempengaruhi kelulusan mahasiswa. Model yang dihasilkan bisa diimplementasikan untuk melakukan klasifikasi mahasiswa aktif angkatan 2016 dan 2017 dengan nilai accuracy sebesar 79.53% dan nilai recall sebesar 19.23%.

Kata kunci: data mining; klasifikasi; kelulusan mahasiswa; decision tree; k-fold cross validation.

PENDAHULUAN

Mahasiswa dapat menyelesaikan masa studi dengan batas minimal waktu yang telah ditentukan, atau bisa dikatakan mahasiswa bisa lulus dengan tepat waktu merupakan salah satu indikator kinerja akademik yang baik bagi suatu perguruan tinggi. Ketentuan masa studi sudah

diatur dalam ketetapan Kementerian Pendidikan dan Kebudayaan Direktorat Jendral Perguruan Tinggi (KEMENRISTEK DIKTI) tentang sistem pendidikan perguruan tinggi yang menjelaskan bahwa kompetensi lulusan harus menyelesaikan beban wajib dengan minimal capaian 144 SKS (Satuan Kredit Semester) untuk tingkat Strata Satu (S1) dan 108 SKS untuk Diploma Tiga (D3), yang diselesaikan dalam 8 semester untuk S1 dan 4 semester untuk D3, yang dapat ditempuh dalam waktu paling lama sampai 14 semester untuk S1 dan 10 semester untuk D3 (KEMENRISTEK DIKTI NO 44, 2015).

Permasalahan mengenai kinerja akademik mahasiswa dalam hal menyelesaikan masa studi secara tepat waktu masih banyak ditemui di FEKON UNIGA. Tabel 1 menunjukkan persentase lulusan mahasiswa FEKON UNIGA yang bisa menyelesaikan masa studi secara tepat waktu dari tahun 2015 sampai dengan tahun 2018.

Tabel 1 Persentase Lulusan Tepat Waktu Mahasiswa FEKON UNIGA tahun 2015-2018

Tahun	Akuntansi D3	Akuntansi S1	Manajemen S1
2015	0%	7%	10%
2016	0%	14%	11%
2017	36%	19%	10%
2018	0%	23%	16%

Seiring dengan pertumbuhan teknologi informasi yang pesat ini telah menciptakan akumulasi data yang sangat besar namun minim informasi, dalam pencarian informasi tersebut muncullah istilah Data Mining. Data Mining adalah proses menemukan pola dan pengetahuan menarik dari kumpulan data dari database, data warehouses, data marts dan repositori (Han, 2012) yang fokus dalam melakukan ekstraksi pengetahuan yang tersembunyi. Metode klasifikasi merupakan salah satu metode dalam Data Mining, klasifikasi banyak dimanfaatkan untuk memprediksi kinerja siswa (Shahiri, 2015).

Dalam penelitian ini akan dilakukan pencarian suatu model yang bisa mengklasifikasi pola kelulusan mahasiswa dengan menggunakan teknik Data Mining klasifikasi dengan algoritma *Decision Tree*. Klasifikasi digunakan dengan cara mengevaluasi hasil kinerja akademik mahasiswa berupa IPS dan jumlah SKS pada tahun pertama dan kedua serta menggunakan demografi mahasiswa sebagai atribut yang akan digunakan dalam dataset. Model dari hasil klasifikasi diimplementasikan pada suatu grafik dan *list* mahasiswa yang sudah diklasifikasikan kelulusannya sebagai visualisasi data dari hasil klasifikasi kelulusan mahasiswa agar mudah dibaca oleh pihak Prodi dan dosen wali. Dengan adanya klasifikasi kelulusan mahasiswa diharapkan bisa membantu pengambil keputusan untuk dapat melakukan tindakan-tindakan yang diperlukan jika masa studi mahasiswa dapat diketahui lebih dini.

KAJIAN PUSTAKA

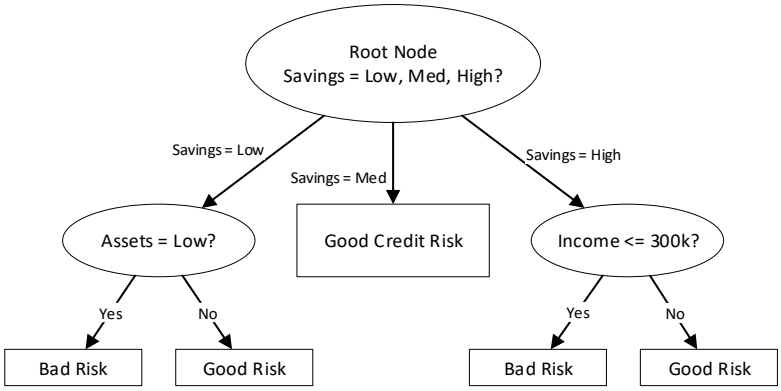
Pada hakikatnya, Data Mining merupakan disiplin ilmu yang tujuan utamanya untuk menemukan, menggali, atau menambang pengetahuan dari data atau informasi yang kita miliki dengan proses menganalisis data dari perspektif yang berbeda dan dapat menyimpulkan menjadi informasi-informasi penting yang bisa digunakan untuk memperkecil biaya pengeluaran, meningkatkan keuntungan atau keduanya (Mabrur, 2012). Dengan kata lain Data Mining bisa dikatakan sebagai kegiatan mengumpulkan data yang diekstraksi menjadi suatu pengetahuan baru berupa pola, kecenderungan atau keteraturan yang dapat digunakan sebagai acuan dalam rangka pengambilan keputusan atau untuk keperluan suatu bisnis. Fungsi umum dalam data mining terbagi menjadi enam (Larose, 2014), yaitu Deskripsi, Estimasi, Prediksi, Klasifikasi, Klustering dan Asosiasi.

Educational Data Mining (EDM) merupakan konsep Data Mining yang digunakan untuk mengeksplorasi data pendidikan untuk mengetahui pola dan prediksi deskriptif yang mencirikan perilaku dan pencapaian siswa, pengetahuan konten domain, penilaian, fungsi pendidikan dan aplikasi pendidikan (Peña-Ayala, 2014). EDM menganalisis data yang dihasilkan oleh berbagai jenis sistem informasi yang mendukung pembelajaran atau pendidikan. Data yang digunakan tidak terbatas pada interaksi siswa individu dengan sistem pendidikan seperti nilai ujian, nilai kehadiran dan nilai latihan, tetapi mungkin juga mencakup data administratif, data demografis, aktivitas siswa dan sebagainya. Data dalam pendidikan memiliki karakteristik khas seperti beberapa tingkat hierarki (subjek, penugasan, tingkat pertanyaan), konteks (siswa tertentu di kelas tertentu yang menemui pertanyaan tertentu pada waktu tertentu dan pada tanggal tertentu) dan longitudinal (banyak data yang direkam selama banyak sesi untuk jangka waktu yang panjang) (Romero, 2013).

Decision Tree adalah model prediksi yang dapat digunakan untuk mewakili kedua pengklasifikasi dan model regresi, dalam pohon keputusan penelitian operasi mengacu pada model hierarki keputusan dan konsekuensinya (Rokach, 2015). Decision Tree juga merupakan struktur flowchart yang menyerupai pohon (tree), dimana setiap simpul internal menandakan suatu tes pada atribut, setiap cabang hasil tes, dan simpul daun mempresentasikan kelas atau distribusi kelas (Han, 2012). Decision Tree terdiri dari node yang membentuk pohon berakar. Tree diarahkan dengan node yang disebut "root" yang tidak memiliki masukan. Semua node lain memiliki tepat satu tepi masukan. Sebuah node dengan tepi keluar disebut sebagai internal node atau test node dan semua node lain disebut leaves (Rokach, 2015).

Dalam decision tree, masing-masing node internal membagi ruang instance menjadi dua atau lebih sub-ruang sesuai dengan fungsi diskrit tertentu dari nilai atribut input. Dimulai pada simpul akar, yang dengan Konvensi ditempatkan di bagian atas diagram Decision Tree. Atribut

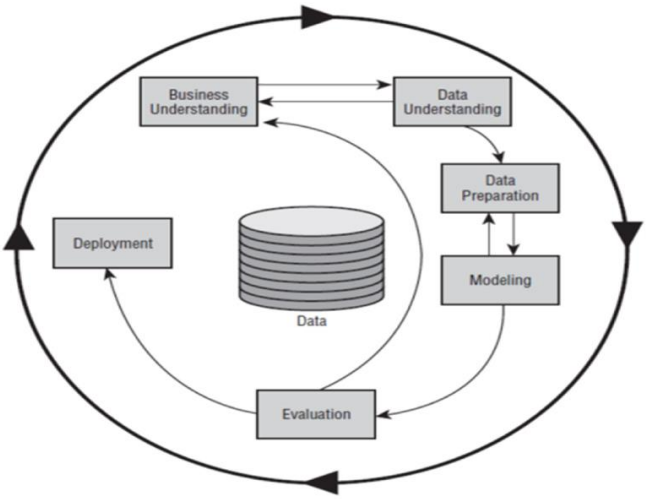
diuji pada node keputusan dengan setiap hasil yang mungkin mengakibatkan cabang. Setiap cabang kemudian mengarah ke simpul keputusan lain atau ke simpul leaves (Larose, 2014). Gambar 1 menunjukkan contoh sederhana dari Decision Tree.



Gambar 1 Simple Decision Tree (Larose, 2014)

Cross Validation mirip dengan metode sub sampling acak yang berulang, namun pengambilan sampel dilakukan sedemikian rupa sehingga tidak ada dua set pengujian yang tumpang tindih (Berrar, 2018). Dalam k-fold cross-validation, data awal dipartisi secara acak ke dalam subset k yang saling bertumpuk (fold), D1, D2, ..., Dk, masing-masing berukuran kurang lebih sama (Han, 2012). Dengan kata lain K-fold cross validation adalah metode yang digunakan untuk mengetahui rata-rata keberhasilan dari suatu model dengan cara melakukan perulangan dan mengacak atribut input sehingga sistem tersebut teruji untuk beberapa atribut dengan input yang acak. Untuk klasifikasi perkiraan akurasi adalah jumlah keseluruhan klasifikasi yang benar dari iterasi k, dibagi dengan jumlah total tupel pada data awal. Secara umum, k bernilai 10 dan dianjurkan untuk memperkirakan nilai akurasi (jika daya komputasi memungkinkan bisa lebih) karena bias yang relatif rendah dan varians (Han, 2012)

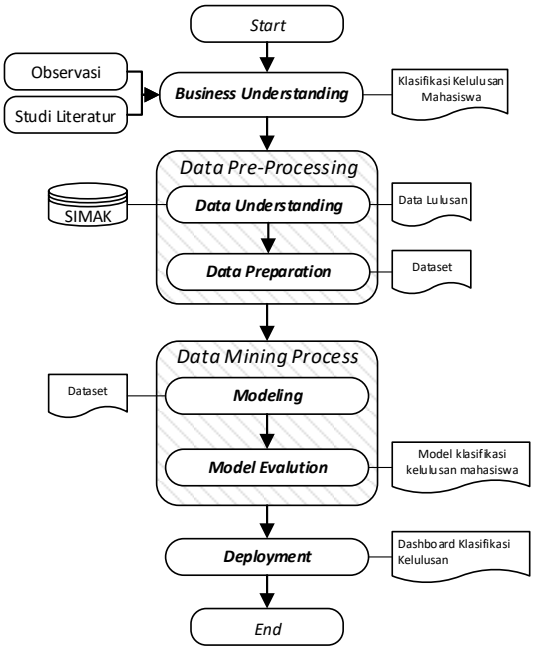
CRoss-Industry Standard Process for Data Mining (CRISP-DM) dikembangkan oleh Daimler Chrysler (kemudian Daimler-Benz), SPSS dan NCR pada tahun 1999, versi CRISP-DM 1.0 diterbitkan dan lengkap serta didokumentasikan. Saat ini model proses untuk Data Mining memberikan gambaran umum dari siklus hidup proyek Data Mining, yang berisi fase proyek, tugas masing-masing dan hubungan antara tugas-tugas dalam proyek (Chapman, 2000). Pada dasarnya, hubungan bisa ada antara tugas-tugas Data Mining tergantung pada tujuan, latar belakang dan minat pengguna dan yang paling penting adalah pada ketersediaan data. Siklus hidup proyek penambangan data terdiri dari enam fase. Gambar 2 menunjukkan fase proses CRISP-DM. Urutan fase tidak kaku yang berarti bergerak bolak-balik di antara beberapa fase yang tergantung pada hasil dari setiap fase.



Gambar 2 CRISP-DM Process Model (Chapman, 2000)

METODE PENELITIAN

Metodologi utama yang digunakan dalam melakukan penelitian ini adalah dengan menggunakan metode CRISP-DM. Gambar 3 menunjukkan langkah-langkah yang akan dilakukan pada penelitian ini.



Gambar 3. Metodologi Penelitian

HASIL DAN PEMBAHASAN

Business Understanding

Ketentuan masa studi di perguruan tinggi negeri maupun swasta sudah diatur dalam ketetapan KEMENRISTEK DIKTI. KEMENRISTEK DIKTI menetapkan bahwa

kompetensi lulusan pendidikan perguruan tinggi harus memenuhi beban SKS (Satuan Kredit Semester) sesuai dengan tingkatannya. Permasalahan mengenai kinerja akademik mahasiswa dalam hal menyelesaikan masa studi secara tepat waktu masih banyak ditemui di FEKON UNIGA. FEKON UNIGA membutuhkan sistem atau suatu model bisnis yang bisa menekan mahasiswa yang menyelesaikan masa studinya secara tidak tepat waktu dengan menemukan pola kelulusan. FEKON UNIGA mengembangkan suatu model Data Mining dengan tujuan:

A. Melakukan klasifikasi kelulusan mahasiswa pada tahun kedua (semester keempat) berdasarkan data karakteristik mahasiswa dan beberapa pencapaian selama dua tahun.
B. Menemukan faktor yang paling mempengaruhi selama dua tahun baik dari sisi demografi mahasiswa maupun hasil akademik mahasiswa.
C. Memberikan kemudahan kepada dosen wali dan Prodi dalam memberikan perlakuan khusus kepada mahasiswa yang teridentifikasi tidak akan lulus tepat waktu.
D. Meningkatkan kualitas mahasiswa serta kualitas sistem pendidikan dan pembelajaran di FEKON UNIGA.

Data Understanding

Penelitian ini menggunakan dua jenis data yaitu data demografi dan data hasil akademik mahasiswa FEKON UNIGA dari lulusan tahun 2015 sampai tahun2018 yang terdiri atas Prodi Akuntansi D3, Prodi Akuntansi S1 dan Prodi Manajemen S1.

A. Data Akademik

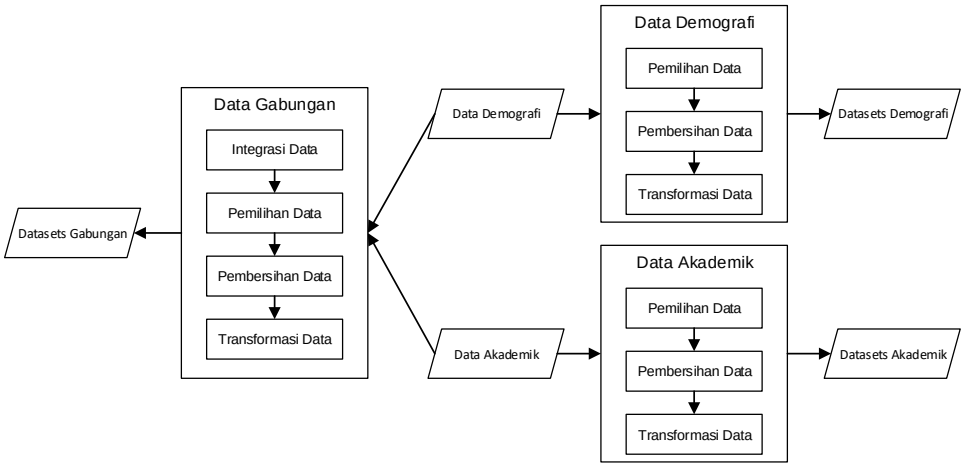
   Data akademik merupakan rekapan nilai akademis mahasiswa yang sudah lulus pada tahun 2015 sampai 2018 dengan jumlah 1373 record data dan 17 atribut.

B. Data Demografi

   Data demografi merupakan data pribadi mahasiswa yang sudah lulus pada tahun 2015 sampai 2018 dengan jumlah 1373 record data dan 19 atribut.

Data Preparation

Pada penelitian ini, proses persiapan data dilakukan terpisah pada data demografi, data akademik serta data gabungan. Karena setiap data memiliki atribut dan format yang berbeda, langkah-langkah yang dilakukan dalam proses persiapan data akan berbeda pula. Langkah-langkah persiapan untuk masing-masing data ditampilkan pada Gambar 4. Langkah ini pada akhirnya menghasilkan tiga dataset berupa data demografi, data akademik dan data gabungan.



Gambar 4 Proses Persiapan Data

Modeling

Pada pemodelan dilakukan 3 jenis pemodelan terhadap dataset demografi, dataset akademik, dan dataset gabungan keduanya dengan menggunakan algoritma Decision Tree. Penentuan data train dan data testing pada proses pemodelan ini menggunakan metode K-Fold Cross Calidation dengan nilai 10 fold. Selanjutnya model yang dihasilkan dievaluasi untuk mengetahui dataset mana yang memberikan hasil pemodelan paling baik dan algoritma mana yang memiliki hasil akurasi yang paling baik. Setiap dataset memiliki jumlah record sebanyak 1304 data. Untuk menemukan nilai pembanding, dalam setiap percobaan akan dicari nilai accuracy, recall, precission dan F-measure yang dihasilkan dari confusion matrix.

Atribut yang digunakan pada dataset demografi, dataset akademik, dan dataset gabungan diurutkan berdasarkan information gain. Skenario penelitian ini menggunakan semua atribut pada skenario pertama di setiap dataset, kemudian menghapus atribut dengan nilai information gain terendah pada eksperimen berikutnya. Eksperimen akan dilakukan sampai dengan nilai confusion matrix dari model menunjukkan nilai tertinggi. Tabel 2 menunjukkan nilai information gain pada setiap atribut.

Tabel 2 Information Gain Setiap Atribut

No	Dataset Akademik		Dataset Demografi		Dataset Gabungan	
	Atribut	Information Gain	Atribut	Information Gain	Atribut	Information Gain
1	SKS4	0.104282	PendidikanAyah	0.0307381	SKS4	0.104282
2	SKS3	0.0937728	PendidikanIbu	0.0212307	SKS3	0.0937728

No	Dataset Akademik		Dataset Demografi		Dataset Gabungan	
	Atribut	Information Gain	Atribut	Information Gain	Atribut	Information Gain
3	IPS4	0.0767301	Kelamin	0.0113722	IPS4	0.0767301
4	IPS3	0.0719908	PekerjaanAyah	0.011265	IPS3	0.0719908
5	IPS1	0.0650817	Program	0.0097202	IPS1	0.0650817
6	IPS2	0.0571957	Menikah	0.004005	IPS2	0.0571957
7	Prodi	0.0003442	LuarKota	0.0024583	PendidikanAyah	0.0307381
8			Prodi	0.0003442	PendidikanIbu	0.0212307
9			HidupAyah	0.0003266	Kelamin	0.0113722
10			HidupIbu	0.0002653	PekerjaanAyah	0.011265
11			UsiaMasuk	0	Program	0.0097202
12					Menikah	0.004005
13					LuarKota	0.0024583
14					Prodi	0.0003442
15					HidupAyah	0.0003266
16					HidupIbu	0.0002653
17					UsiaMasuk	0

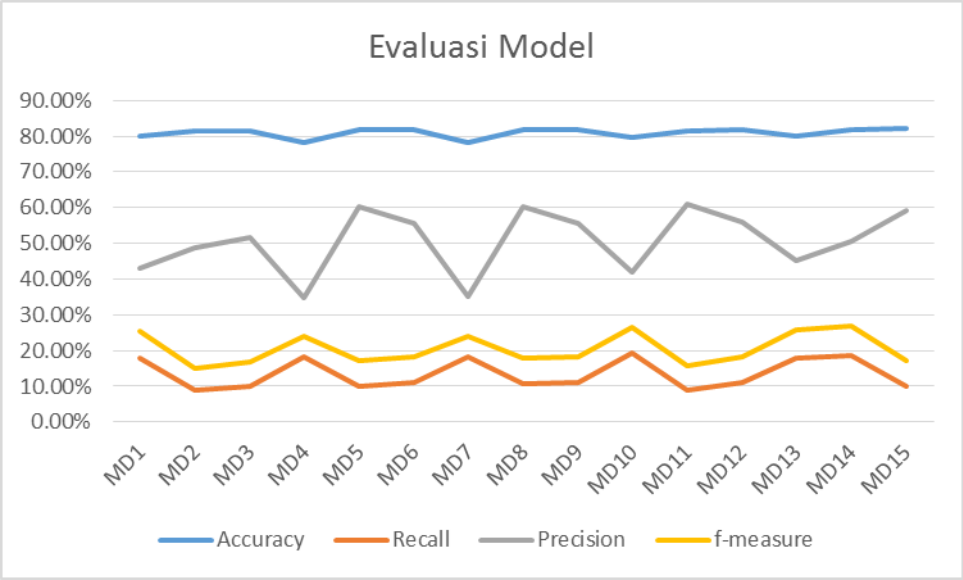
Evaluation

Evaluasi dilakukan terhadap model-model paling baik yang dihasilkan tiap tahap pemodelan. Tujuan evaluasi adalah untuk mengetahui model mana yang paling baik dalam mengklasifikasikan kelulusan mahasiswa. Perbandingannya didapat dataset akademik, dataset demografi dan dataset gabungan, atribut yang digunakan dan algoritma mana yang digunakan. Evaluasi ini dilakukan dengan menganalisis nilai accuracy, recall, precission, dan f-measure yang memiliki kualitas model yang dihasilkan. Nilai accuracy, recall, precission, dan f-measure didapatkan dari hasil perhitungan confusion matrix untuk setiap percobaan dan modelnya. Tabel 3 dan Gambar 5 merupakan hasil evaluasi dari beberapa model yang dihasilkan.

Tabel 3 Hasil evaluasi beberapa model

Nama Model	Accuracy	Recall	Precision	F-measure
MD1	80,21%	17,96%	43,12%	25,36%
MD2	81,59%	8,92%	48,69%	15,08%
MD3	81,52%	10,15%	51,69%	16,97%
MD4	78,38%	18,37%	34,69%	24,02%

MD5	81,90%	10,13%	60,36%	17,35%
MD6	81,82%	10,95%	55,52%	18,29%
MD7	78,37%	18,40%	35,20%	24,17%
MD8	81,90%	10,55%	60,28%	17,96%
MD9	81,82%	10,95%	55,52%	18,29%
MD10	79,53%	**19,23%**	41,83%	26,35%
MD11	81,59%	8,93%	**61,00%**	15,58%
MD12	81,98%	10,95%	56,00%	18,32%
MD13	80,14%	17,97%	45,04%	25,69%
MD14	81,83%	18,45%	50,48%	**27,02%**
MD15	**82,05%**	10,15%	59,33%	17,33%



Gambar 5 Hasil Evaluasi Model

Berdasarkan nilai dari table 3, didapat beberapa model dengan nilai Acuracy, Recall, Precission dan f-measure yang paling tinggi yang disajikan pada table 4.

Tabel 4 Model dengan nilai Acuracy, Recall, Precission dan f-measure paling tinggi

Nilai	**Acuracy**	**Recall**	**Precision**	**f-measure**
Dataset	Gabungan	Demografi	Akademik	Akademik
Nama Model	MD15	MD10	MD11	MD14
Nilai	82.05%	19.23%	61.00%	27.02%

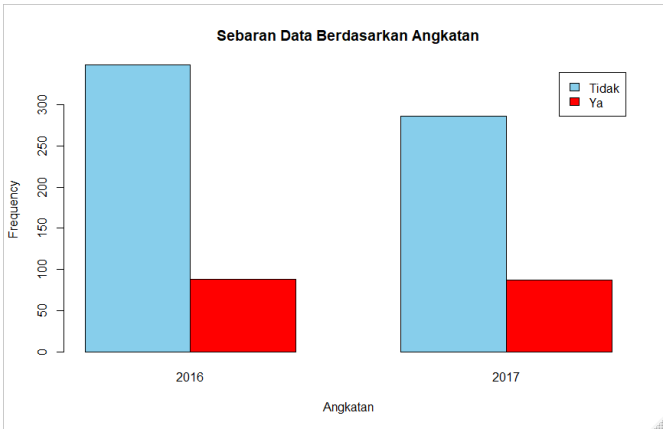
Nilai	Acuracy	Recall	Precision	f-measure
Atribut	SKS4	PendidikanAyah	SKS4	SKS4
	SKS3	PendidikanIbu	SKS3	SKS3
	IPS4	Kelamin	IPS4	IPS4
	IPS3	PekerjaanAyah	IPS3	
	IPS1	Program		
	IPS2	Menikah		
	PendidikanAyah	LuarKota		
	PendidikanIbu	Prodi		
	Kelamin			
	PekerjaanAyah			
	Program			
	Menikah			
	LuarKota			

Deployment

Deployment model dilakukan pengambilan data mahasiswa aktif angkatan 2016 dan 2017 dari semua mahasiswa FEKON UNIGA yang mana pada angkatan tersebut sudah memenuhi kriteria klasifikasi. Total ada 999 data yang didapat, di mana mahasiswa angkatan 2017 berjumlah 423 data dan mahasiswa angkatan 2016 berjumlah 576 data. Selanjutnya dilakukan proses data preprocessing dan menyisakan sebanyak 809 data (436 data angkatan 2016 dan 373 angkatan 2017). Tabel 4 merupakan rekapitulasi dari klasifikasi kelulusan mahasiswa angkatan 2016 dan 2018 FEKON UNIGA. Selanjutnya Gambar 5 menunjukkan sebaran data klasifikasi kelulusan berdasarkan angkatan.

Tabel 5 Rekapitulasi Klasifikasi Kelulusan Mahasiswa Angkatan 2016-2018

Prediksi Tepat Waktu	Jumlah	Confidence (Mean)	Persentase
Ya	175	0.222	21.6%
Tidak	634	0.772	78.4%



Gambar 6 Sebaran Data Klasifikasi Kelulusan Berdasarkan Angkatan

KESIMPULAN DAN SARAN

Berdasarkan beberapa skenario uji, faktor-faktor yang bisa mempengaruhi kelulusan mahasiswa adalah Pendidikan Ibu, Jenis Kelamin, Pekerjaan Ayah, Program Kuliah, Status Pernikahan, Kota asal dan Prodi. Jika menilai pada accuracy tertinggi didapatkan pada dataset demografi dengan nilai sebesar 82.05%. Model yang dihasilkan bisa diimplementasikan untuk melakukan klasifikasi mahasiswa aktif angkatan 2016 dan 2017 dengan jumlah data 809 mahasiswa di mana terdapat 634 mahasiswa yang terindikasi tidak akan lulus tepat waktu.

DAFTAR PUSTAKA

Berrar, D. (2018). Cross-Validation. *Encyclopedia of Bioinformatics and Computational Biology*, 1–4.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *CRISP-DM Step-by-step Data Mining Guide*. USA: SPSS Inc.

Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Data mining concepts and techniques* (3rd ed.). Waltham: Morgan Kaufmann Publishers.

Larose, D. T., & Larose, C. D. (2014). *Discovering Knowledge In Data, An Introduction to Data Mining* (2nd ed.). New Jersey: John Wiley & Sons.

Mabrur, A. G., & Lubis, R. (2012). Penerapan Data Mining Untuk Memprediksi Kriteria Nasabah Kredit. *Jurnal Komputer Dan Informatika (KOMPUTA)*, *1*(1), 53–57.

Peña-Ayala, A. (2014). *Educational Data Mining Applications and Trends*. London: Springer. https://doi.org/10.1007/978-3-319-02738-8

Rokach, L., & Maimon, O. (2015). *Data Mining With Decision Trees Theory And Applications* (2nd ed.). Singapore: World Scientific Publishing.

Romero, C., & Ventura, S. (2013). Data Mining in Education. *WIREs Data Mining and Knowledge Discovery*, *3*(1), 12–27.

Shahiri, A. M., Husain, W., & Rashid, N. A. (2015). A Review on Predicting Student's Performance Using Data Mining Techniques. *Procedia Computer Science*, *72*, 414–422. https://doi.org/10.1016/j.procs.2015.12.157