Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK) DOI: 10.25126/jtiik.202072093
Vol. 7, No. 3, Juni 2020, hlm. 443-452 p-ISSN: 2355-7699
Akreditasi KEMENRISTEKDIKTI, No. 30/E/KPT/2018 e-ISSN: 2528-6579

KOMPARASI *DATA MINING NAIVE BAYES* DAN *NEURAL NETWORK* MEMPREDIKSI MASA STUDI MAHASISWA S1

Azahari*[1], Yulindawati[2], Dewi Rosita[3], dan **Syamsuddin Mallala[4]**

[1,4]Sistem Informasi, STMIK Widya Cipta Dharma, [2]Teknik Informatika, STMIK Widya Cipta Dharma
[3]Program Studi Pendidikan Ilmu Komputer, Jurusan Pendidikan MIPA, FKIP Universitas Mulawarman
Email: [1]azaharilathyf@yahoo.com, [2]yuli.linda08@yahoo.com, [3]dew.rosita@gmail.com, [4]mallala_s@yahoo.co.id
*Penulis Korespondensi

(Naskah masuk: 05 Juli 2019, diterima untuk diterbitkan: 22 April 2020)

Abstrak

Prediksi kelulusan dibutuhkan oleh manajemen perguruan tinggi dalam menentukan kebijakan preventif terkait pencegahan dini kasus drop out. Lama masa studi setiap mahasiswa bisa disebabkan dengan berbagai faktor. Dengan menggunakan *data mining* algoritma *naive bayes* dan *neural network* dapat dilakukan prediksi kelulusan mahasiswa di STMIK Widya Cipta Dharma (WiCiDa) Samarinda . Atribut yang digunakan yaitu, umur saat masuk kuliah, klasifikasi kota asal Sekolah Menengah Atas, pekerjaan ayah, program studi, kelas, jumlah saudara, dan Indeks Prestasi Kumulatif (IPK). Sampel mahasiswa yang lulus dan *drop-out* pada tahun 2011 sampai 2019 dijadikan sebagai data *training* dan data *testing*. Sedangkan angkatan 2015–2018 digunakan sebagai data target yang akan diprediksi masa studinya. Sebanyak 3229 mahasiswa, 1769 sebagai data *training*, 321 sebagai data *testing*, dan 1139 sebagai data target. Semua data diambil dari data mahasiswa program strata 1, dan tidak mengikut sertakan data mahasiswa D3 dan alih jenjang/transfer. Dari data *testing* diperoleh tingkat akurasi hanya 57,63%. Hasil penelitian menunjukkan banyaknya kelemahan dari hasil prediksi *naive bayes* dikarenakan tingkat akurasi kevalidannya tergolong tidak terlalu tinggi. Sedangkan akurasi prediksi *neural network* adalah 72,58%, sehingga metode alternatif inilah yang lebih baik. Proses evaluasi dan analisis dilakukan untuk melihat dimana letak kesalahan dan kebenaran dalam hasil prediksi masa studi.

Kata kunci: *Data mining, Naive bayes, Prediksi, Kelulusan, Mahasiswa*

COMPARATION OF NAIVE BAYES AND NEURAL NETWORK DATA MINING IN PREDICTING STUDY PERIOD OF UNDERGRADUATE STUDENT

Abstract

Graduation predictions are required by the higher education institution preventive policies related to the early prevention of drop-out cases. The duration of study, for each student can be caused by various factors. By using the data mining algorithm Naive bayes and neural network, the student graduation in STMIK Widya Cipta Dharma (WiCiDa) can be predicted. The attributes used are as follows: age at admission, classification of cities from high school, father's occupation, study program, class, number of siblings, and grade point average (GPA). Samples of students who graduated and dropped out between year 2011 and 2019 were used as training data and testing data. While the year class of 2015to 2018 is used as the target data, which will be predicted during the study period. According to the data mining algorithm Naive bayes, there are 3229 students; 1769 as training data, 321 as testing data, and 1139 as target data. All data is taken from students enrolled in undergraduate program and does not include data on diploma students and transfer student. From the testing data, an accuracy rate only 57.63%. The other side, prediction accuracy of the neural network is 72.58%, so this alternative method is the best chosen. The research results show the many weaknesses of the results of prediction of Naive bayes because the level of accuracy of its validity is not high. The evaluation and analysis process are conducted to see where the errors and truths are in the results of the study period predictions.

Keywords: *Data mining, Naive bayes, Predictions, Graduation, College Student*

1. PENDAHULUAN

Kualitas perguruan tinggi, khususnya program studi di Indonesia diukur berdasarkan akreditasi yang dilaksanakan oleh Badan Akreditasi Nasional Perguruan Tinggi atau BAN PT (Instrument BAN PT, 2011) kualitas tersebut diukur berdasarkan 7 standar utama, salah satunya adalah Mahasiswa dan Lulusan. Terkait dengan kualitas perguruan tinggi menjadi salah satu butir dari akreditasi yaitu mahasiswa yang lulus tepat waktu. Berdasarkan Buku Pedoman Akademik STMIK Widya Cipta Dharma, Program Sarjana (S-1) adalah program pendidikan akademik yang memiliki beban studi berkisar 144 – 160 SKS yang dijadwalkan antara 7-8 semester yang dapat ditempuh maksimal 14 semester (Buku Pedoman Akademik, 2018). Adanya informasi mengenai kelulusan mahasiswa tepat waktu dapat menjadikan suatu pengambilan keputusan bagi manajemen Perguruan Tinggi tersebut dalam pengambilan langkah selanjutnya.

Pemanfaatan data yang ada dalam sistem informasi akademik untuk menunjang kegiatan pengambilan keputusan tidak cukup hanya mengandalkan data operasional saja, diperlukan suatu analisis data untuk menggali potensi-potensi informasi yang ada. Jadi, proses pengubahan data menjadi informasi dan dari informasi yang ada akan diambil polanya agar menjadi pengetahuan. Dalam mengambil keputusan, kebanyakan mereka hanya amenggali informasi yang berguna. Hal ini mendorong munculnya cabang ilmu baru untuk mengatasi masalah penggalian informasi atau pola yang penting atau menarik dari data dalam jumlah besar, yang disebut dengan *data mining*.

STMIK Widya Cipta Dharma (WiCiDa) memiliki 3 program studi dimana ada 2 program studi jenjang S1 (Teknik Informatika dan Sistem Informai) dan 1 program studi jenjang D3 (Manajemen Informatika). Dengan memanfaatkan *data mining* pada data bidang pendidikan, sebuah institusi perguruan tinggi bisa memperoleh suatu informasi yang berguna, dimana selanjutnya informasi tersebut dapat menjadi suatu landasan untuk melakukan perbaikan untuk meningkatkan kualitas perguruan tinggi.

Data-data bidang pendidikan pada umumnya bisa berupa data profil mahasiswa, mata kuliah, KRS (kartu rencana studi), KHS (Kartu Hasil Studi) dan sebagainya, yang biasanya tersimpan dalam database Sistem Informasi Akademik (SIAK) dalam jumlah yang besar, dimana sebenarnya dari data bidang pendidikan tersebut dapat digunakan untuk menggali sebuah informasi. Data yang diperoleh dari melalui proses mining digunakan untuk mendesain model *naive bayes* dan *neural network*. Dari hasil model ini digunakan untuk memprediksi kelulusan mahasiswa pada program S1.

Naive bayes adalah salah satu algoritma pembelajaran induktif yang paling efektif dan efisien untuk *machine learning* dan *data mining* (Mohammad dkk, 2018). *Naïve Bayes* merupakan pengklasifikasian dengan metode probabilitas dan statistik yang dikemukakan oleh ilmuwan Inggris Thomas Bayes, yaitu memprediksi peluang di masa depan berdasarkan pengalaman di masa sebelumnya sehingga dikenal sebagai teorema *Bayes*. Teorema tersebut dikombinasikan dengan ”naive” dimana diasumsikan kondisi antar atribut saling bebas. Algoritma *Naive bayes* adalah pengklasifikasian statistik yang dapat digunakan untuk memprediksi probabilitas keanggotaan suatu *class*. *Bayesian classification* didasarkan pada teorema *Bayes* yang memiliki kemampuan klasifikasi serupa dengan *decision tree* dan *neural network* (Jananto, 2013; Gao dkk, 2018).

Penelitian memprediksi waktu kelulusan mahasiswa juga telah banyak dilakukan, misalnya yang dilakukan Nurhuda & Rosita (2017) tetapi menggunakan jaringan saraf tiruan, adapula penelitian Wijayanti, dkk (2017) dan Sun, dkk (2017) bukan memprediksi tetapi melakukan pengelompokan lulusan (*clustering*). Selain untuk memprediksi waktu kelulusan, penelitian *naive bayes* juga dilakukan untuk memprediksi mahasiswa *drop-out* (Tasnim, dkk, 2017) dan kinerja mahasiswa baru (Adekitan & Noma, 2017). Untuk meningkatkan keakuratan algoritma beberapa penelitian mengkombinasikannya dengan metode *data mining* lainnya, seperti dengan C.45 (Kurniawan, 20187) atau *K-Nearest Neighbor* (Devita,dkk, 2017; Tempola,dkk, 2018).

Sedangkan pada Penelitian ini dilakukan penerapan algoritma *Naive bayes* dan *neural network* untuk memprediksi kelulusan mahasiswa S1 pada program studi Teknik Informatika dan Sistem Informasi. Penelelitian tidak dilakukan pada mahasiswa jenjang D3 Manajemen Informatika karena pendidikan vokasi telah memprogram mahasiswa D3 harus lulus di semester ke-6. Penelitian juga tidak dilakukan pada mahasiswa jenjang S1 yang merupakan mahasiswa transfer atau pindahan. Kontribusi secara umum yang diberikan dari penelitian ini adalah institusi dapat mengambil langkah dan kebijakan yang tepat untuk memaksimalkan jumlah mahasiswa jenjang S1 yang lulus tepat waktu, dan meminimkan jumlah mahasiswa yang *drop-out* atau mengundurkan diri.

Penelitian ini meliputi sebuah penerapan metode algoritma *naive bayes* untuk memprediksi kelulusan mahasiswa. Hasil dari penerapan metode *naive bayes* diimplementasikan degan menggunakan aplikasi *data mining*. Melihat keunggulan dan kekurangan dari metode prediksi yang dipakai. Apabila hasil persentase prediksi kurang memuaskan, maka metode *neural network* sebagai alternatiflah yang akan digunakan. Kedua metode ini dipiliah karena dari sangat tepat dalam menemukan

pola dari data mahasiswa yang beragam (Nurhuda & Rosita, 2017; Sun, dkk ,2017).

Data yang akan diprediksi angkatan 2015 – 2018 dan data yang akan *di-training* adalah mahasiswa yang lulus dan *drop-out* pada tahun 2011 – 2019 yang terdiri dari 2 program studi dengan beberapa atribut yaitu, umur (saat masuk kuliah), klasifikasi kota asal Sekolah Menengah Atas, pekerjaan ayah, program Studi, kelas (Pagi/Malam), jumlah tanggungan dalam keluarga (jumlah saudara), dan IPK. Penelitin ini menggunakan variabel yang tidak hanya berpatok pada prestasi akademik mahasiswa saja, tetapi juga mempertimbangkan variabel dan fator kondisi latar belakang keluarganya. Hal ini dilakukan untuk melihat pengaruh faktor kondisi keluarga terhadap lamanya masa studi, disinilah yang menjadi kebaruan penelitian ini dibanding penelitian-penelitian sebelumnya.

2. METODE PENELITIAN

Data mining adalah Proses menganalisis data yang banyak dan membuat suatu pola untuk menjadi informasi yang berguna. Lain hal nya pula menurut Witten, Frank and Mark (2016) *Data mining* didefinisikan sebagai proses menemukan pola-pola dalam data. Proses ini otomatis atau seringnya semiotomatis. Pola yang penuh arti dan pola tersebut memberikan keuntungan, biasanya keuntungan secara ekonomi. Data yang dibutuhkan dalam jumlah besar.

Data yang digunakan dalam penelitian ini merupakan data primer yang diambil dari data Sistem Informasi Akademik (SIAK) STMIK WiCiDa. Atribut yang diambil dalam penelitian ini antara lain, umur (saat masuk kuliah) diambil dari tanggal lahir dibandingkan dengan tahun masuk, klasifikasi kota asal Sekolah Menengah Atas (diklasifikasikan berasal dari Ibu Kota, Kota, atau Kabupaten), program studi (Teknik Informatika atau Sistem Informasi berdasarkan pada kode prodi.), kelas pagi atau malam, pekerjaan ayah, dan jumlah saudara yang menjadi tanggungan dalam keluarga.

Sedangkan pada mahasiswa yang telah lulus atau *drop-out* juga diambil atribut masa studi berdasarkan tahun kelulusan dan tahun mengundurkan diri. Dari data diperoleh dari Sistem Informasi Akademik STMIK WiCiDa ditampilkan pada tabel 1.

Tabel 1. Atribut data yang dibutuhkan

No	NIM	Umur Kuliah	Asal Sekolah	Pekerjaan Ayah	Kode prodi	kelas	Jumlah saudara	IPK	Masa studi	Tgl. lulus
1	0611001	20	ibu kota	Sudah Meninggal	55201	PB	0	1,32	*drop-out*	-
2	0611002	19	kabupaten	Sudah Meninggal	55201	MA	0	2,75	*drop-out*	-
3	0641004	21	ibu kota	Karyawan Swasta	57201	MA	0	2,65	Sangat lambat	26/07/2013
..	..	..	..	..	..	..	..	..	..	..
3907	1843002	19	kabupaten	PNS	55201	PA	2	3,05	-	-

Dapat dipahami dari tabel 1, asal sekolah dari "ibu kota" adalah mahasiswa dari SMA/MA atau sederajat yang berasal dari Kota Samarinda, sedangkan "kota" berasal dari Kota Balikpapan dan Bontang, dan "kabupaten" berasal dari Kabupaten Kutai Kartanegara, Kutai Barat, Berau, dan lainnya. Sedangkan mahasiswa yang berasal dari sekolah luar daerah Kalimantan Timur juga akan diklasifikasikan berdasarkan kota asal sekolahnya.

Atribut pekerjaan ayah dibagi menjadi 9 jenis pekerjaan yaitu, PNS, TNI/POLRI, Karyawan Swasta, Wiraswasta, Karyawan BUMN, Pensiunan, Petani/Buruh, Sudah Meninggal, dan lain-lain bagi pekerjaan yang tidak dapat diklasifikasikan. Sedangkan kelas PA,PB,PC adalah kelas pagi dan MA,MB adalah kelas malam. Masa studi juga diklasifikasikan dalam 5 jenis yaitu, "cepat" bagi mahasiswa S1 yang lulus 3,5 tahun, "normal" bagi yang lulus 4 – 4,5 tahun, "lambat" bagi mahasiswa yang lulus 5 – 5,5 tahun, "sangat lambat" bagi mahasiswa yang lulus 6 tahun atau lebih, dan "*drop-out"* bagi mahasiswa yang mengundurkan diri atau kehabisan waktu masa studi.

Dari tabel 1 juga diperoleh 1220 mahasiswa yang telah lulus jenjang strata 1 dari tahun 2011-2019, 871 mahasiswa yang *drop-out* dari angkatan 2006-2010, dan 1816 mahasiswa yang masih aktif kuliah dari angkatan 2011 sampai 2018. Data yang akan digunakan sebagai data *training* adalah 1769 mahasiswa S1 yang lulus dan *drop-out* pada tahun 2011 – 2017, ditampilkan pada tabel 2.

Sedangkan yang akan menjadi data *testing* adalah 321 mahasiswa, dimana 172 mahasiswa lulus pada tahun 2018 – 2019, dan 149 *drop-out* pada tahun 2018. Belum ada mahasiswa yang *drop-out* di tahun 2019 pada saat penelitian ini berlangsung. Data mahasiswa angkatan 2015 – 2018 yang akan diprediksi masa studinya berjumlah 1139 mahasiswa. Sedangkan sisanya 678 mahasiswa aktif angkatan 2011 – 2014 tidak diolah datanya karena masa studi yang sudah lebih dari 4,5 tahun, dikategori lambat dan sangat lambat. Bahkan ada 142 mahasiswa angkatan 2011 yang akan di-*drop-out* atau harus mengundurkan diri di akhir tahun 2019 ini.

Tabel 2. Data mahasiswa yang lulus dan *drop-out* dari tahun 2011-2017

No	NIM	Umur Kuliah	Asal Sekolah	Pekerjaan Ayah	Kode Prodi.	kelas	Jumlah saudara	IPK	Masa studi
1	0611001	20	ibu kota	Sudah Meninggal	55201	PB	0	1,32	*drop-out*
2	0611002	19	kabupaten	Sudah Meninggal	55201	MA	0	2,75	*drop-out*
3	0741023	37	ibu kota	Karyawan Swasta	57021	MA	6	3,05	lambat
..	..	..	..	..	..	..	..	..	..
1769	1343118	18	Ibu kota	Karyawan Swasta	55201	PA	2	3,19	cepat

2.1 Tahapan Penelitian

Penelitian ini bertujuan untuk memprediksi kelulusan mahasiswa S1 STMIK WiCiDa, jenis penelitian adalah penelitian eksperimen dengan tahapan sebagai berikut:

1. Pengumpulan data
   Pengumpulan data lebih dari sekedar mengambil data yang ada tetapi harus mampu mendeskripsikan data yang ada, serta memiliki kontribusi terhadap pengetahuan. Data tersebut harus jelas, memiliki relasi, dapat diukur, dapat diprediksi, memiliki generalisasi serta teori (Masters, 2018).
2. Pengolahan awal data
   Data yang sudah dikumpulkan diolah dengan algoritma *soft-computing* untuk mengurangi data yang tidak relevan, atau data dengan atribut yang hilang. Pengolahan juga dapat berupa konversi nilai nilai redundan atau nilai yang terlalu beragam kedalam kelompok yang lebih kecil untuk mempermudah pembentukan model.
3. Model/metode yang diusulkan
   Menggambarkan alur metode yang diusulkan serta menjelaskan cara kerja model yang diusulkan dari *naive bayes* dan *neural network*. Kedua metode ini akan digambarkan secara skematik dan disertai dengan formula perhitungan. Model akan dibentuk dari data yang sudah diolah, dan hasil pengolahan model akan diukur dengan model yang ada saat ini.
4. Eksperimen dan pengujian model
   Menjabarkan bagaimana eksperimen yang dilakukan hingga terbentuknya model, serta menjelaskan cara menguji model yang terbentuk.
5. Evaluasi dan validasi hasil
   Evaluasi dilakukan dengan mengamati hasil prediksi menggunakan algoritma *soft computing*. Validasi dilakukan dengan mengukur hasil prediksi dibandingkan dengan data asal. Pengukuran kinerja dilakukan dengan membandingkan nilai *error* hasil prediksi antara algoritma yang diusulkan dan algoritma alternatif, sehingga dapat diketahui algoritma yang lebih akurat. Dalam hal ini metode alternatif yang kami gunakan adalah *neural network.*

3. HASIL DAN PEMBAHASAN

Penelitian ini mengolah sebanyak 1769 mahasiswa sebagai data *training,* 321 mahasiswa sebagai data *testing,* dan 1139 mahasiswa yang akan diprediksi masa studinya. Terdapat 7 atribut dan 1 *class* dalam setiap data yaitu:

1. XI : Umur Kuliah
2. X2 : Asal sekolah
3. X3 : Pekerjaan ayah
4. X4 : Program Studi
5. X5 : Kelas
6. X6 : Jumlah saudara
7. X7 : IPK
8. Class : masa studi

Pengonversian semua teks dan kategori menjadi angka agar data dapat diolah dengan aplikasi *data mining,* pengonversian dilakukan pada atribut berikut:

1. Pengonversian asal sekolah:
   1) Ibu kota = 1
   2) Kota = 2
   3) Kabupaten = 3
2. Pengonversian Pekerjaan Ayah
   1) PNS = 1
   2) TNI/POLRI = 2
   3) Pegawai BUMN = 3
   4) Karyawan Swasta = 4
   5) Wiraswasta = 5
   6) Lain-lain = 6
   7) Pensiunan = 7
   8) Petani/Buruh = 8
   9) Sudah Meninggal = 0
3. Pengonversian Kelas
   1) PA/PB/PC = 1
   2) MA/MB = 2

Sedangkan umur, program studi, jumlah saudara, dam IPK tidak dikonversi karena telah berbentuk angka. Hasil konversi data ditampilkan pada tabel 3, yaitu 1769 data mahasiswa yang lulus dan *drop-out* dari tahun 2011-2017 yang menjadi data *training*.

Tabel 3. Data *training* yang telah dikonversi

No	X1	X2	X3	X4	X5	X6	X7	Class
1	20	1	8	55201	1	0	1,32	*drop-out*
2	19	3	8	55201	2	0	2,75	*drop-out*
3	37	1	4	57021	2	6	3,05	lambat
..	..	..	..	..	..	..	..	..
1769	18	1	4	55201	1	2	3,19	cepat

Setelah dilakukan pengkonversian dalam beberapa atribut. Pada 1769 data mahasiswa yang dijadikan data *training* dilakukan penambangan data menggunakan algoritma *Naive bayes* sehingga terbentuk sebuah model. Pada model yang telah dibangun digunakan untuk diterapkan pada data *testing* dengan prediksi masa studi 321 mahasiswa, dimana 172 mahasiswa lulus pada tahun 2018 – 2019, dan 149 *drop-out* pada tahun 2018. Hasil prediksi pada data *testing* ditampilkan pada tabel 4. Tabel 4 menampilkan perbedaan dan persamaan hasil prediksi dengan masa studi yang sebenarnya telah terjadi.

Tabel 4. Validasi hasil prediksi masa studi pada data *testing*

No	X1	X2	X3	X4	X5	X6	X7	Class	Prediksi	Valid
1	29	1	6	57201	2	5	1.63	*drop-out*	*drop-out*	*True*
2	22	1	6	57201	2	0	2.07	*drop-out*	*drop-out*	*True*
3	19	1	1	57201	2	3	2.77	*drop-out*	sangat lambat	*False*
4	23	1	4	57201	2	0	2.96	sangat lambat	sangat lambat	*True*
..	..	..	..	..	..	..	..	..	..	..
321	21	1	5	57201	2	3	3.24	normal	lambat	*False*

Dari 321 mahasiswa sebagai data *testing* yang diuji terhadap model yang dibangun, didapat 185 data valid dan 136 data tidak valid. Data valid berarti hasil prediksi masa studi dan *class* masa studi menunjukkan hasil yang sama. Maka model yang diujikan memiliki nilai kevalidan 185 ÷ 321 × 100% = 57,63% dengan nilai *error* 42,37%. Tingkat kevalidan model yang diajukan tidak terlalu tinggi, maka model diuji ulang terhadap data *training* yang membentuknya, ditampilkan pada tabel 5.

Tabel 5. Validasi hasil prediksi masa studi pada data *training*

No	X1	X2	X3	X4	X5	X6	X7	Class	Prediksi	Valid
1	20	1	8	55201	1	0	1,32	*drop-out*	*drop-out*	*True*
2	19	3	8	55201	2	0	2,75	*qdrop-out*	*drop-out*	*True*
3	37	1	4	57021	2	6	3,05	lambat	cepat	*False*
4	18	1	2	57201	1	0	2.56	sangat lambat	*drop-out*	*False*
..	..	..	..	..	..	..	..	..	..	..
1769	18	1	4	55201	1	2	3,19	normal	normal	*True*

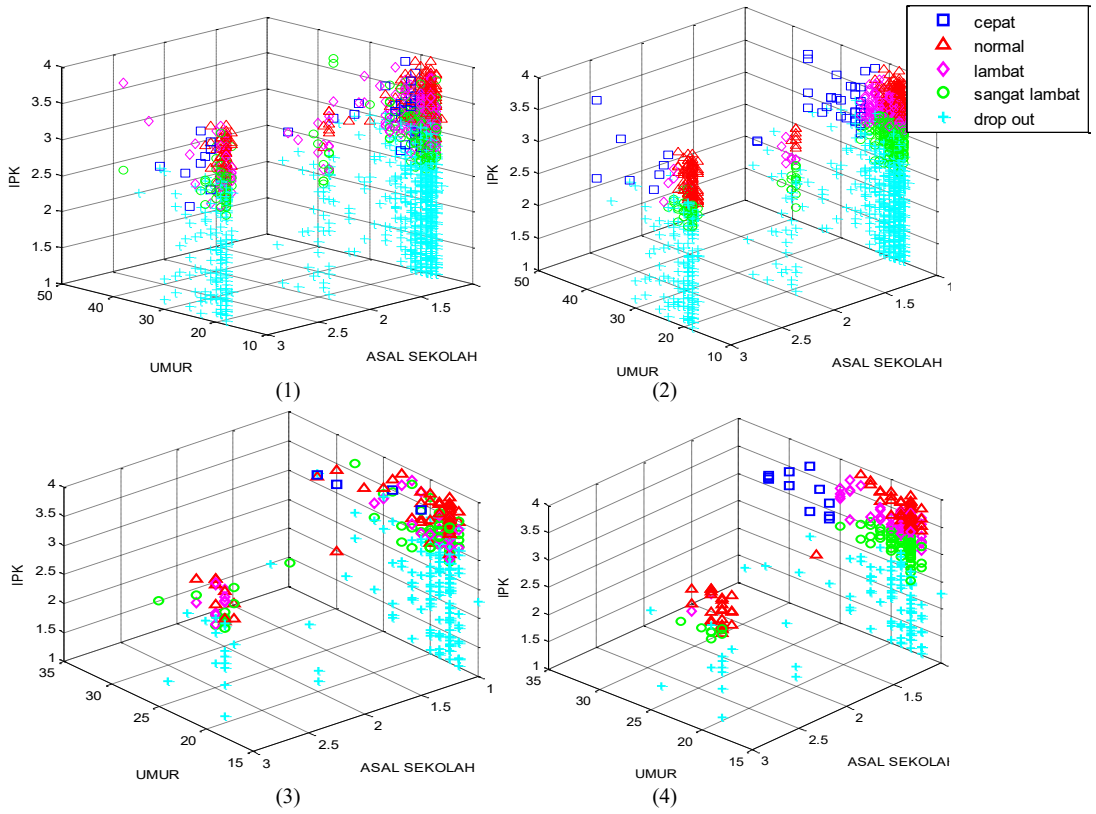
Hasil pengujian ulang dari model yang dibangun terhadap 1769 mahasiswa yang sebelumnya digunakan sebagai data *training*. Didapat 1187 data valid dan 582 data tidak valid. Maka model yang diujikan memiliki nilai kevalidan 1187 ÷ 1769 × 100% = 67,1 % dengan nilai *error* 49,03 %. Nilai kevalidan model lebih tinggi dari pada sebelumnya di data *testing*. Hal ini menunjukkan bahwa model yang dibangun masih dalam kategori dapat diterima untuk diusulkan.

Perbandingan hasil prediksi dan masa studi yang sebenarnya telah terjadi, disajikan dalam grafik surface pada gambar 1. Dimana ada 3 atribut yang dipakai yaitu, umur(x), IPK(y), dan asal sekolah(z).

Dapat dilihat pada gambar 1, klasifikasi masa studi cepat ditampilkan dengan simbol persegi biru, masa studi normal dengan simbol segitiga merah, masa studi lambat dengan simbol *diamond* ungu, masa studi sangat lambat dengan simbol lingkaran hijau, dan *drop-out* dengan simbol tambah biru muda. Jumlah mahasiswa yang kuliah di STMIK WiCiDa lebih banyak berasal dari ibu kota dibandingkan dari kabupaten dan kota. Mahasiswa dengan IPK tinggi lebih banyak yang lulus dibandingkan mahasiswa dengan IPK rendah yang sebagian besar *drop-out*.

Gambar 1 (1) dan (2) menampilkan, sebagian besar mahasiswa yang mendaftar sebagai mahasiswa baru berumur sekitar 20 tahun-an, hasil prediksi yang salah menampilkan mahasiswa yang berumur lebih tua lebih cepat lulus, sedangkan pada kenyataannya mahasiswa yang berumur tua lebih banyak menempuh masa studi lebih lama dapat dilihat dari simbol persegi biru lebih banyak tersebar pada umur di atas 20 tahun di gambar 1 (2) dan (4).

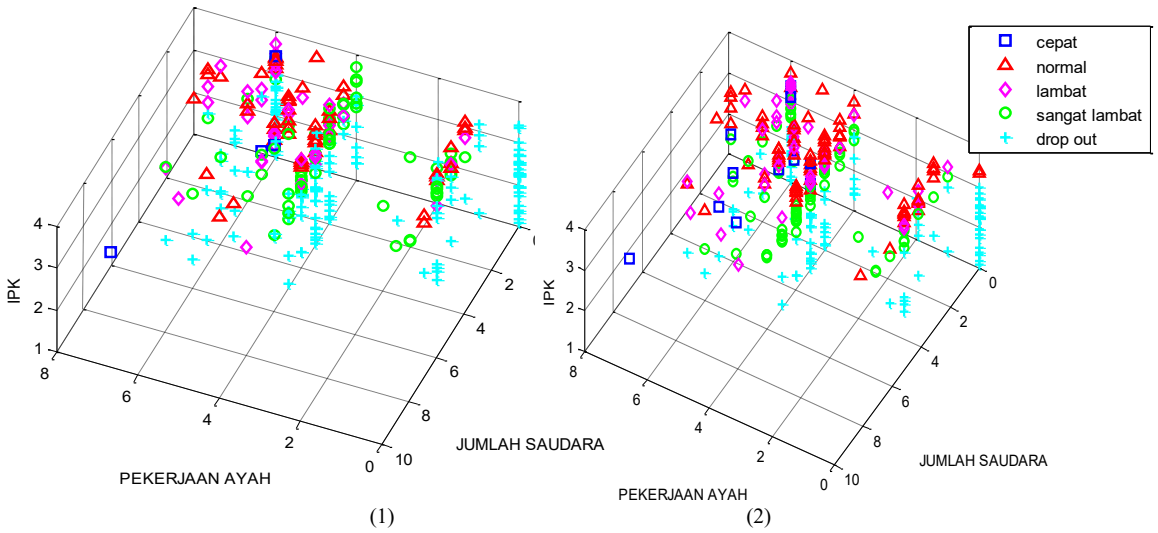
Perbedaan prediksi juga terjadi pada IPK di antara 2,50 sampai 3,00. Lihat persebaran simbol lingkaran hijau pada gambar 1 (2) dan (4), hasil prediksi menampilkan banyak mahasiswa dengan IPK diantara rentang tersebut lulus dengan sangat lambat, tetapi pada kenyataannya pada gambar 1 (1) mahasiswa lulus dengan lambat dan normal, dan pada gambar 1 (3) mahasiswa dengan IPK tersebut banyak yang lulus dengan lambat dan *drop-out.*



Gambar 1. Grafik *surface* perbandingan klasifikasi masa studi sebenarnya (1) dan (3) dan hasil prediksi (2) dan (4), pada data *training* (1) dan (2) serta data *testing* (3) dan (4)

Data *testing* pada gambar 1 (3) dan (4) juga menampilkan perbedaan prediksi pada masa studi mahasiswa lulusan yang berasal dari kabupaten. Prediksi menampilkan mahasiswa yang berasal dari kabupaten lebih banyak yang lulus normal, tetapi pada kenyataannya banyak yang lulus dengan lambat dan sangat lambat.

Perbandingan hasil prediksi juga dilakukan pada atribut yang lain yaitu, pekerjaan ayah (x), IPK (y), dan jumlah saudara (z), dengan tujuan melihat apakah ada pengaruh atribut pekerjaan ayah dan jumlah saudara terhadap nilai IPK dan lama studi. Grafik *surface* pada gambar 2 menampilkan perbedaan hasil prediksi pada data *testing* pada ketiga atribut tersebut.



Gambar 2. Grafik *surface* perbandingan klasifikasi masa studi sebenarnya (1) dan hasil prediksi (2) pada data *testing* dengan 3 atribut IPK, pekerjaan ayah, dan jumlah saudara

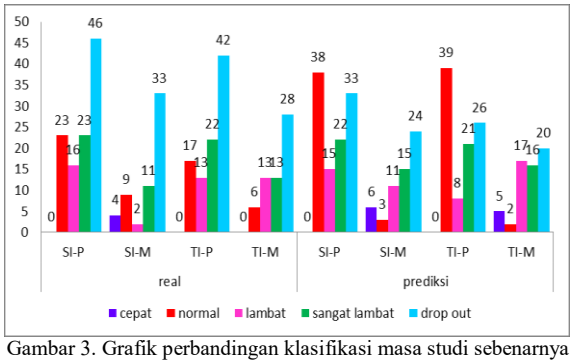
Lihat pada gambar 2 pada pekerjaan ayah 7 (pensiunan), hasil prediksi dengan masa studi sebenarnya menampilkan perbedaan pada pekerjaan ayah "pensiunan". Hasil prediksi menampilkan mahasiswa dengan ayah pensiunan lulus dengan masa studi normal dan lambat, sedangkan pada kenyataannya mahasiswa tersebut lebih banyak yang *drop-out*. Tetapi grafik data *testing* di gambar 2 menampilkan kesamaan prediksi, pada pekerjaan ayah "sudah meninggal". Mahasiswa dengan ayah sudah meninggal sebagian besar telah *drop-out.* Grafik juga menampilkan kesamaan mahasiswa dengan ayah sudah meninggal dan mahasiswa tersebut adalah anak tunggal (jumlah saudara 0) lebih banyak yang *drop-out.*

Membandingkan hasil prediksi juga dilakukan pada presensi grafik dari 2 atribut program studi dan kelas, ditampilkan pada gambar 3. SI-P adalah kelas Sistem Informasi Pagi, SI-M adalah kelas Sistem Informasi Malam, TI-P adalah kelas Teknik Informatika Pagi, dan TI-M adalah kelas Teknik Informatika Malam. Ditampilkan perbandingan masa studi sebenarnya dan hasil prediksi pada data *testing.*

Dapat dilihat pada gambar 3, terlihat perbedaan jumlah mahasiswa *drop-out* (grafik batang berwarna biru muda) antara keadaan nyata dan prediksi. Hasil prediksi menampilkan lebih banyak mahasiswa yang lulus dengan masa studi normal (grafik batang berwarna merah) pada kelas SI-P dan TI-P, padahal pada kenyataannya lebih banyak mahasiswa yang *drop-out* pada kelas tersebut.

Tetapi hasil prediksi menampilkan banyak kesamaan pada mahasiswa yang lulus dengan masa studi lambat dan sangat lambat pada setiap kelas. Sebenarnya dapat kita analisis pada data tersebut bahwa mahasiswa kelas malam memiliki persentase lebih banyak yang di *drop-out* dibandingkan mahasiswa kelas pagi, walaupun sisi jumlah mahasiswa pagi lebih banyak yang *drop-out*, karena jumlah mahasiswa kelas pagi lebih banyak dari pada kelas malam. Hasil prediksi pada kelas SI-M dan TI-M juga menunjukkan bahwa mahasiswa kelas malam lebih banyak yang *drop-out* dari pada yang lulus.

Dari sekian proses evaluasi kevalidan model *naive bayes* yang dibangun. Dilakukan penerapan model tersebut terhadap data yang akan di prediksi. Data yang akan diprediksi adalah mahasiswa angkatan 2015 – 2018 berjumlah 1139 mahasiswa. Setelah memalui proses pengkonversian nilai atribut dan penerapan model *naive bayes* didapat hasil prediksi pada tabel 6.



Gambar 3. Grafik perbandingan klasifikasi masa studi sebenarnya (*real*) dan hasil prediksi pada data *testing* dengan 2 atribut program studi dan kelas

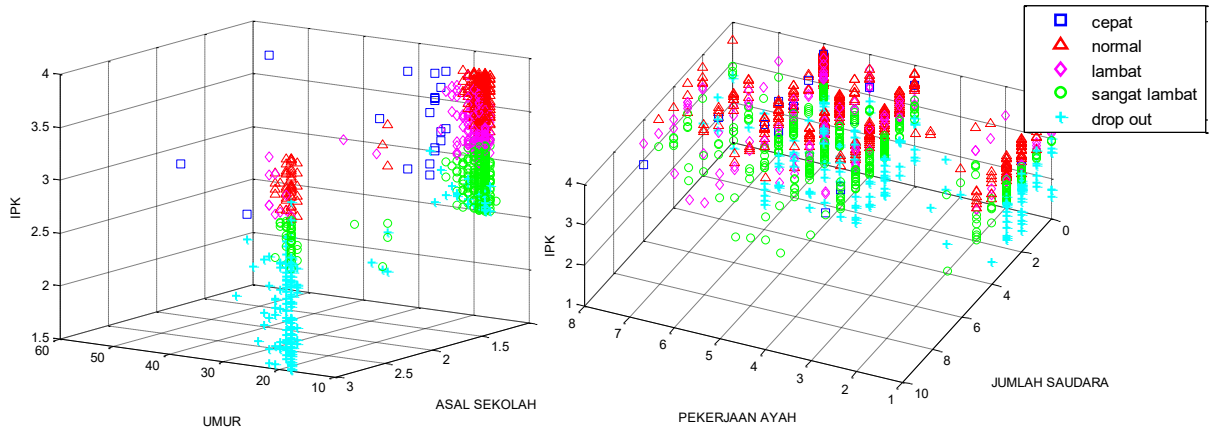
Tabel 6. Hasil prediksi masa studi mahasiswa angkatan 2015 sampai dengan 2018

No	NIM	X1	X2	X3	X4	X5	X6	X7	Prediksi
1	1541001	18	1	5	57201	2	3	3.37	lambat
2	1541002	18	3	7	57201	2	1	2.39	*drop-out*
3	1541003	19	3	1	57201	1	4	3.12	normal
4	1541004	18	3	5	57201	1	1	1.80	*drop-out*
..		..	..	..	..	..	..	..	..
1139	1843124	20	1	4	55201	1	4	3.53	normal

Persebaran data mahasiswa ditampilkan pada grafik surface pada gambar 4. Dimana ada 3 atribut yang digunakan pada gambar 4 (1) yaitu, umur, asal sekolah dan IPK. Sedangkan pada gambar 4 (2) atribut yang digunakan yaitu, pekerjaan ayah, jumlah saudara dan IPK.

Dapat dianalisis hasil prediksi pada gambar 4 (1), bahwa mahasiswa dengan IPK yang tinggi lebih banyak yang lulus cepat, dan IPK yang rendah lebih banyak yang lulus sangat lambat dan *drop-out*. Hasil prediksi juga menampilkan lebih banyak mahasiswa yang *drop-out* berasal dari asal sekolah kabupaten dibandingkan dari kota. Mahasiswa dengan umur di atas 20 tahun saat menjadi mahasiswa baru, lebih banyak yang diprediksi lulus cepat dibandingkan umur di bawah 20 tahun, hasil prediksi inilah yang menurut perbandingan dengan data testing dan training menunjukkan kesalahan.

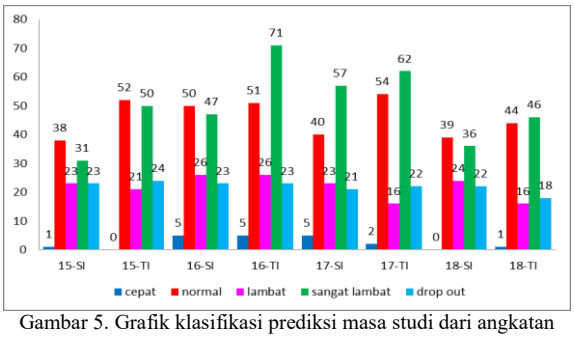
Sedangkan hasil prediksi pada gambar 4 (2), tidak menampilkan pola data class yang mencolok dari masa studi terhadap pekerjaan ayah dan jumlah saudara. Hanya terlihat dari persebaran simbol persegi biru, mahasiswa yang diprediksi lulus cepat mempunyai ayah dengan pekerjaan karyawan swasta dan "lain-lain".



(1) (2)

Gambar 4. Grafik surface hasil prediksi masa studi mahasiswa angkatan 2015 sampai dengan 2018

Hasil prediksi juga ditampilkan pada grafik dari 2 atribut angkatan dan program studi, ditampilkan pada gambar 5. Grafik menampilkan prediksi masa studi dari mulai angkatan 2015 sampai 2018 untuk program studi Sistem Informasi dan Teknik Informatika.



Gambar 5. Grafik klasifikasi prediksi masa studi dari angkatan 2015-2018 untuk program studi Sistem Informasi dan Teknik Informatika

Hasil prediksi pada gambar 5 menunjukkan, kebanyakan mahasiswa lulus dengan masa studi normal dan sangat lambat, kecil sekali jumlah mahasiswa yang lulus cepat dari kedua program studi. Banyaknya mahasiswa yang diprediksi lulus dengan sangat lambat, besar kemungkinan akan meningkatkan jumlah mahasiswa yang *drop-out* di masa yang akan datang.

Rendahnya tingkat kevalidan hasil prediksi *naive bayes*, yang hanya menghasilkan nilai kevalidan *data tesing* 57,63%, membuat kami harus menggunakan algoritma *neural network* sebagai metode alternatif. *Data training* dan *data testing* yang digunakan adalah data yang sama digunakan pada pengujian sebelumnya. Penilaian performa prediksi atau pengukuran keakuratan suatu arsitektur jaringan yang digunakan dalam penelitian adalah faktor kesalahan RMSE *(Root Mean Square Error)*, dimana dari hasil terkecil akan dipilih menjadi model yang paling akurat.

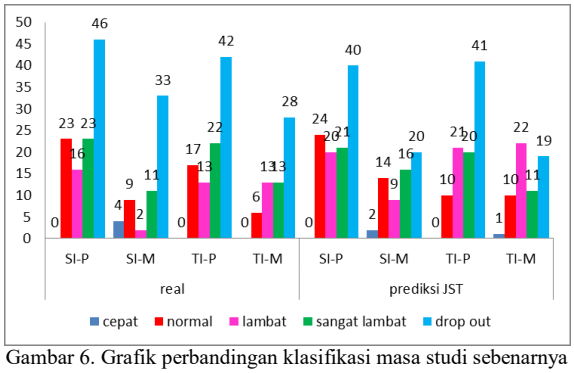
Berdasarkan beberapa eksperimen yang telah dilakukan pada tabel 7, arsitektur jaringan syaraf tiruan teroptimal untuk prediksi kelulusan adalah arsitektur 1 lapisan *hidden layer input* dengan 7 *neuron* dan fungsi pelatihan *traingdx*, dengan MSE *(Mean Square Error)* yang dihasilkan adalah $4{,}36\times10^{-6}$ dan *learning rate* 0,001. MSE tersebut adalah nilai minimum yang dihasilkan dari bebrapa eksperimen dengan fungsi pelatihan *taringd, traingdx, traingda, dan trainlm*. Hasil perbandingan iterasi dan eksperimen dapat dilihat pada tabel 7

Tabel 7. Hasil eksperimen arsitektur *neural network* untuk 7 *neuron* dan 1 *hidden layer*

#	Epochs	Learning rate	Goal	Max Fail	MSE
1	2000	0.1	0.001	6	0.00988
2	3000	0.01	0.0001	6	0.000380
3	2000	0.3	0	2000	0.00643
4	322	0.001	0	6	4.38e-06
5	349	0.01	0	1000	3.30e-06
6	2000	0.1	0	1000	0.00318
7	2000	0.001	0.003	6	*0.00774*
8	2000	0.001	0.002	6	0.00731
9	2000	0.01	0.003	6	0.000857

Pada Tabel 7, kita dapat melihat bahwa pengulangan percobaan pada nomor 4 dihentikan di iterasi ke-322 dengan nilai MSE terkecil, maka jaringan ini yang terbaik untuk digunakan untuk memprediksi kelulusan. Hasil model *training* ini divalidasi pula dengan menguji kembali pada *data testing,* seperti langkah yang dijabarkan sebelumnya di tabel 5. Tetapi kali ini, hasil validasi dari 321 mahasiswa sebagai data *testing*, didapat 232 data valid, yang berarti model diujikan memiliki nilai kevalidan 233 ÷ 321 × 100% = 72,58%. Akurasi lebih tinggi, dibandingkan dengan hasil prediksi dengan *naive bayes*.

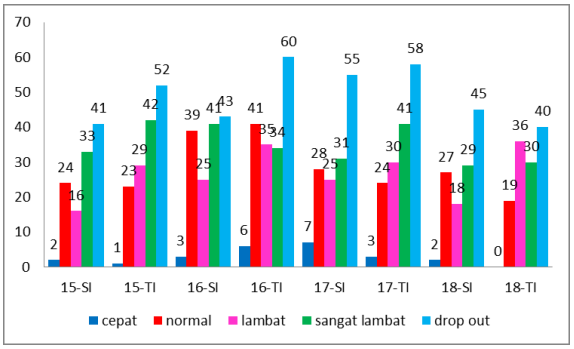
Membandingkan hasil prediksi data *testing* juga dilakukan kembali pada presensi grafik dari 2 atribut program studi dan kelas, ditampilkan pada gambar 6. Dibandingkan hasil prediksi masa studi sebenarnya, dengan hasil prediksi *neural network.*



Gambar 6. Grafik perbandingan klasifikasi masa studi sebenarnya (*real*) dan hasil prediksi *neural network* pada data *testing*

Jika kita bandingkan hasil prediksi *neural network* pada gambar 6, terlihat tidak jauh perbedaan antara jumlah mahasiswa yang *drop-out* (grafik batang berwarna biru muda) dan lulus normal (grafik batang berwarna merah) dalam kondisi nyata dan hasil prediksi. Jika dibandingkan dengan hasil prediksi sebelumnya pada grafik di gambar 3, hasil prediksi *naive bayes* menunjukkan kesalahan dengan menampilkan lebih banyak mahasiswa yang lulus normal dibandingkan dengan yang *drop-out.*

Hasil *prediksi neural* network juga ditampilkan pada grafik 2 atribut (angkatan dan program studi). Gambar 7 menampilkan prediksi masa studi mahasiswa angkatan 2015 sampai 2018 untuk program studi SI dan TI dengan *neural network.*



Gambar 7. Grafik klasifikasi prediksi *neural network,* untuk masa studi dari angkatan 2015-2018 program studi SI dan TI.

Berbeda dengan hasil prediksi *naive bayes* pada gambar 5, yang menunjukkan kebanyakan mahasiswa lulus dengan masa studi normal dan sangat lambat untuk angkatan 2015-2018. Prediksi *neural network* pada gambar 7 menunjukkan kebanyakan mahasiswa tersebut *drop-out.* Banyaknya mahasiswa yang diprediksi akan *drop-out* harus menjadi pekerjaan rumah bagi akademik STMIK WiCiDa dalam mengambil kebijakan untuk meningkatkan kinerja mahasiswanya.

4. KESIMPULAN

Metode *naive bayes* menggunakan atribut umur saat masuk kuliah, klasifikasi kota asal Sekolah Menengah Atas, pekerjaan ayah, program studi, kelas, jumlah saudara, dan IPK, tingkat kevalidan data *testing*nya hanya 57,63%, untuk memprediksi masa studi mahasiswa S1 di STMIK Widya Cipta Dharma. Dari data *training* yang digunakan sebagai model untuk memprediksi masa studi angkatan 2015-2018 didapatkan paling banyak mahasiswa lulus dengan masa studi sangat lambat dan normal.

Tingkat kevalidan model dengan *naive bayes* yang tidak tinggi menampilkan beberapa kesalahan. Prediksi menampilkan mahasiswa yang berumur lebih tua lebih cepat lulus, sedangkan pada kenyataannya mahasiswa yang berumur tua lebih banyak menempuh masa studi lebih lama. Selain itu diprediksi mahasiswa yang berasal dari sekolah kabupaten lebih banyak yang lulus normal, tetapi pada kenyataannya banyak yang lulus dengan lambat dan sangat lambat. Tetapi rendahnya nilai validasi ini telah di*back-up* dengan algoritma *neural network* sebagai metdode alternatif. Dimana hasil prediksi dari *neural network* memiliki tingkat kevalidan 72,58%.

Hasil prediksi data *testing* yang valid menunjukkan beberapa fakta menarik. Diketahui mahasiswa dengan ayah sudah meninggal atau pensiunan dan mahasiswa tersebut adalah anak tunggal (jumlah saudara 0) lebih banyak yang *drop-out*. Selain itu prediksi juga menunjukkan, mahasiswa kelas malam memiliki persentase lebih banyak yang di *drop-out* dibandingkan mahasiswa kelas pagi. Hasil penelitian ini dapat digunakan STMIK WiCiDa untuk mengambil langkah dan kebijakan yang tepat untuk memaksimalkan jumlah lulusannya.

Penelitian kedepannya dapat menambahkan atribut yang lebih diskriminan, seperti penghasilan orang tua, jalur masuk kuliah, jurusan pada saat sekolah, ataupun IPK di setiap semester. Penggunaan algoritma lain untuk memprediksi seperti C4.5, SVM, atau *Expectation Maximisation* (EM *Algorithm*) diharapkan lebih meningkatkan nilai kevalidan model.

UCAPAN TERIMA KASIH

Direktorat Riset dan Pengabdian Masyarakat
Direktorat Jenderal Penguatan Riset dan
Pengembangan Kementerian Riset, Teknologi, dan
Pendidikan Tinggi sesuai dengan Kontrak Penelitian
Tahun Anggaran 2019

DAFTAR PUSTAKA

ADEKITAN, A. I., & NOMA-OSAGHAE, E. 2019. *Data mining* approach to predicting the performance of first year student in a university using the admission requirements. Education and Information Technologies, 24(2), 1527-1543.

BAN-PT. 2010. Pedoman Evaluasi-Diri Untuk Akreditasi Program Studi Dan Institusi Perguruan Tinggi. Jakarta: DiktiJanner.

DEVITA, R. N., HERWANTO, H. W., & WIBAWA, A. P. 2018. Perbandingan Kinerja Metode *Naive bayes* dan K-Nearest Neighbor

untuk Klasifikasi Artikel Berbahasa indonesia. Jurnal Teknologi Informasi dan Ilmu Komputer, 5(4), 427-434.

GAO, C. Z., CHENG, Q., HE, P., SUSILO, W., & LI, J. 2018. Privacy-preserving Naive Bayes classifiers secure against the substitution-then-comparison attack. Information Sciences, 444, 72-88

JANANTO, A. 2013. Algoritma *Naive bayes* untuk Mencari Perkiraan Waktu Studi Mahasiswa. Dinamik, 18(1).

KURNIAWAN, Y. I. 2018. Perbandingan Algoritma *Naive bayes* dan C. 45 dalam Klasifikasi *Data mining*. Jurnal Teknologi Informasi dan Ilmu Komputer, 5(4), 455-464.

MASTERS, T. 2018. *Data mining* Algorithms in C++.USA: Springer

MOHAMMAD, A. H.,ALWADA'N, T., & AL-MOMANI, O. 2018. Arabic text categorization using support vector machine, Naïve Bayes and neural network.GSTF Journal on Computing (JoC), 5(1)

MUFLIKHAH, L., RATNAWATI, D.E., & PUTRI, R.R.M. 2018. *Data mining*. Malang: Universitas Brawijaya Press.

NURHUDA, A., & ROSITA, D. 2017. Prediction Student Graduation on Time Using Artificial Neural Network on *Data mining* Students STMIK Widya Cipta Dharma Samarinda.. Proceedings of the 2017 International Conference on E-commerce, E-Business and E-Government (pp. 86-89). ACM.

SUN, J., ZHOU, A., KEATES, S., & LIAO, S. 2017. Simultaneous Bayesian Clustering and Feature Selection Through Student's t Mixtures Model. IEEE transactions on neural networks and learning systems, 29(4), 1187-1199.

TASNIM, N., PAUL, M. K., & SATTAR, A. S. 2019. Identification of *Drop Out* Students Using Educational *Data mining*. In 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE) (pp. 1-5). IEEE.

TEMPOLA, F., MUHAMMAD, M., & KHAIRAN, A. 2018. Perbandingan Klasifikasi Antara KNN dan *Naive bayes* pada Penentuan Status Gunung Berapi dengan K-Fold Cross Validation. Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK), 5(5).

TIM PENYUSUN BUKU PEDOMAN STMIK WIDYA CIPTA DHARMA. 2018. Buku Pedoman STMIK Widya Cipta Dharma. Samarinda: STMIK Widya Cipta Dharma

WIJAYANTI, S., & ANDREA, R. 2017. K-Means Cluster Analysis for Students Graduation: Case Study: STMIK Widya Cipta Dharma. In Proceedings of the 2017 International Conference on E-commerce, E-Business and E-Government (pp. 20-23). ACM.

WITTEN, I. H., FRANK, E., HALL, M. A., & PAL, C. J. 2016. *Data mining*: Practical machine learning tools and techniques. Morgan Kaufmann.