

Klasifikasi Mahasiswa Berpotensi *Dropout* Menggunakan Metode Regresi Logistik

Classification Of Students With Potential To Drop Out Using The Logistic Regression Method

Dikko Rizky Bintang Lubis^{*1}, Armansyah²

Program Studi Ilmu Komputer, Fakultas Sains dan Teknologi, Universitas Islam Negeri Sumatera Utara, Medan, Indonesia

E-mail : dikkorizky20@gmail.com^{*1}, armansyah@uinsu.ac.id²

Received 7 July 2025; Revised 16 July 2025; Accepted 23 July 2025

Abstrak - *Dropout* merupakan masalah serius yang sering terjadi di dunia perkuliahan. Beragam penyebab yang membuat seorang mahasiswa harus di *dropout* dari kampus mulai dari biaya kuliah yang mahal, jarak tempat kuliah yang jauh, rasa malas dan lain sebagainya. Setiap kampus pasti menginginkan kelulusan untuk setiap mahasiswanya dikarenakan masalah *dropout* ini dapat memberikan penilaian buruk bagi pihak kampus. Maka dari itu, langkah yang dapat dilakukan untuk mengatasi permasalahan ini adalah dengan melakukan klasifikasi. Salah satu metode klasifikasi yang dapat digunakan adalah Regresi Logistik. Metode ini mampu melakukan klasifikasi dengan baik dan memberikan hasil akurasi yang baik juga. Pada penelitian ini, objek yang menjadi bahan penelitian adalah Program Studi Sistem Informasi dan *dataset* yang digunakan adalah data mahasiswa dari Angkatan 2015 sampai 2024. Hasil dari penelitian ini memberikan nilai akurasi yang sangat tinggi yaitu 99%. Sedangkan presisinya 99%, *recall* 95% dan *f1-score* nya adalah 97%.

Kata Kunci – Klasifikasi; Regersi Logistik; *Dropout*; Universitas

Abstract - *Dropout* is a serious problem that often occurs in the world of college. There are various causes that make a student have to drop out of campus. Various causes ranging from expensive tuition fees, far distance to the place of study, laziness and so on. Every campus definitely wants graduation for every student because this dropout problem can give a bad assessment to the campus. Therefore, the steps that can be taken to overcome this problem are by classifying. One of the classification methods that can be used is Logistic Regression. This method is able to classify well and provide good accuracy results too. In this study, the object of the research is the Information Systems Study Program and the dataset used is student data from the 2015 to 2024 Classes. The results of this study provide a very high accuracy value of 99%. While the precision is 99%, the recall is 95% and the *f1-score* is 97%.

Keywords – Classification; Logistic Regression; *Dropout*; University

1. PENDAHULUAN

Dropout selalu menjadi masalah yang serius pada setiap kampus. Masalah ini juga dapat menimbulkan masalah juga bagi pihak kampus itu sendiri, salah satunya akan berdampak bagi penilaian saat proses akreditasi kampus. Setiap kampus juga terus berusaha untuk menekan angka *dropout* agar dapat menurun dan juga mengurangi mahasiswa yang terindikasi berpotensi *dropout*. Namun tidak selalu mudah untuk dilakukan, karena banyaknya faktor-faktor yang menjadi penyebab mahasiswa terindikasi berpotensi *dropout*. Faktor-faktor tersebut diantaranya adalah rasa malas pada mahasiwa, biaya kuliah yang mahal, jarak tempat kuliah dari rumah yang jauh dan lain sebagainya. Dari hasil indikasi tersebut membuat para mahasiswa berpotensi untuk di *dropout* dan jika terus berlanjut maka pihak kampus terpaksa untuk mendropout mahasiswa tersebut. Selain itu juga, setiap kampus pasti memiliki batas akhir dalam melaksanakan

perkuliahan. Seperti yang tertuang pada PERMENRIS/ETDIKTI No 44 Tahun 2015, yang menyatakan bahwa batas waktu kelulusan maksimal adalah 14 Semester atau 7 tahun [1].

Berdasarkan masalah tersebut, perlu dilakukan evaluasi ulang secara berkala dalam meminimalisir kasus mahasiswa berpotensi *dropout* maupun yang telah di *dropout*. Tujuan nya adalah agar tidak menjadi citra buruk bagi dunia pendidikan tinggi bahwa kampus tersebut memiliki mahasiswa *dropout* yang sangat banyak. Maka dilakukanlah berbagai cara mulai dari mengedukasi mahasiswa agar mau menyelesaikan perkuliahannya. Membuka pengurangan biaya kuliah agar mahasiswa yang kurang mampu dapat membayar UKT dengan jumlah yang tidak banyak, dan lain sebagainya.

Dalam *machine learning*, ada beberapa pendekatan yang dapat dilakukan dalam menganalisis kasus mahasiswa berpotensi *dropout* maupun yang *dropout*. Tujuannya adalah agar dapat mengubah suatu kebijakan di sebuah instansi sehingga dapat mengevaluasi ulang dalam mengatasi masalah yang muncul di instansi tersebut. Salah satu pendekatan yang sering digunakan banyak orang adalah dengan melakukan klasifikasi data. Dengan melakukan klasifikasi, diharapkan mampu memberikan gambaran kebijakan dan juga langkah kedepannya dalam mengatasi masalah yang muncul dengan cara mengelompokkan data ke dalam kategori atau kelas tertentu berdasarkan fitur atau karakteristik yang dimiliki oleh data tersebut [2].

Regresi Logistik merupakan salah satu metode yang dapat digunakan dalam melakukan klasifikasi. Metode ini merupakan suatu bentuk analisis matematika yang penggunaannya dapat mengamati hubungan antara beberapa variabel bebas ke satu variabel terikat yang bersifat dikotomi. Regresi logistik ini juga dapat menggunakan variabel kategorik maupun numerik sebagai variabel independennya [4]. Cabang dari Regresi logistik ini adalah Regresi Logistik Multinomial. Metode ini dipilih dikarenakan fitur yang digunakan lebih dari dua sehingga Regresi Logistik Multinomial adalah metode yang mampu menyelesaikan permasalahan tersesbut. adalah cabang dari regresi logistik dalam menentukan hubungan antar variabel, baik variabel [5].

Berbagai penelitian sebelumnya telah dilakukan yang berhubungan dengan klasifikasi terhadap potensi *dropout* mahasiswa. Penelitian oleh [5] dimana pada penelitian ini menunjukkan hasil nilai akurasi sebesar 85,5% dengan menggunakan metode Regresi Logistik Multinomial. Penelitian [6] menunjukkan hasil nilai error MAPE sebesar 8% untuk penelitian tersebut. Penelitian [7] memberikan akurasi sebesar 98,36% dengan menggunakan fitur seleksi dan 91,23% jika tidak menggunakan fitur seleksi. Metode yang digunakan adalah *naïve bayes*. Penelitian [8] dengan menggunakan metode *naïve bayes* dan *decision tree*, memberikan hasil akurasi sebesar 97,83% untuk *naïve bayes* dan 99,13% untuk *decision tree*. Penelitian [9] menggunakan metode KNN dengan hasil akurasi sebesar 99,23% dianggap mampu memprediksi mahasiswa berpotensi *dropout*. [10] melakukan klasifikasi mahasiswa berpotensi *dropout* menggunakan algoritma C4.5 dan mendapatkan hasil akurasi sebesar 98,5%. [11] melakukan perbandingan metode antara SVM dan NN *Backpropagation* dan memberikan hasil akurasi 80% dan 77%, yang menunjukkan bahwa SVM adalah metode terbaik dalam mengklasifikasi mahasiswa berpotensi *dropout*. Penelitian yang dilakukan [12] yaitu mengklasifikasi mahasiswa berpotensi putus studi menggunakan algoritma *Naïve Bayes* dan memberikan akurasi sebesar 97%, jika menggunakan *10-fold cross-validation* memberikan akurasi 95%. Penelitian [13] menerapkan algoritma *Decision Tree* C4.5 dalam mengklasifikasi mahasiswa berpotensi *dropout* dan menunjukkan hasil akurasi sebesar 90%. [14] melakukan analisis potensial *dropout* mahasiswa menggunakan metode *K-Means++ Clustering* dan memberikan hasil penelitian berupa potensi *dropout* paling tinggi di Fakultas Tarbiyah, kedua Fakultas Ushuludin dan Filsafat, ketiga Fakultas Ekonomi dan Bisnis Islam dan terakhir Fakultas Syari'ah. Penelitian dari [15] yang melakukan prediksi tingkat mahasiswa *dropout* menggunakan metode *Naïve Bayes* pada Prodi Informatika dan menunjukkan hasil akurasi sebesar 82%. Dan terakhir penelitian yang dilakukan [16], melakukan prediksi siswa putus sekolah dan keberhasilan akademik menggunakan machine learning Regresi Logistik, *Decision Tree*, *Random Forest* dan *K-Neighbors* dan dari hasil pengujian kelima metode, metode yang terbaik dalam memprediksi adalah Regresi Logistik dengan akurasi 87%.

Dari beberapa penelitian terdahulu diatas yang telah dibahas dengan berbagai metode seperti *Naïve Bayes*, *Decision Tree*, KNN, SVM, dan Regresi Logistik Multinomial, sebagian besar penelitian tersebut hanya fokus pada perbandingan akurasi antar algoritma atau terbatas pada penggunaan fitur tertentu tanpa mempertimbangkan kondisi realistik data akademik lokal secara komprehensif. Selain itu, masih sedikit penelitian yang menggunakan dataset mahasiswa dari institusi lokal seperti UINSU dan memanfaatkan pendekatan klasifikasi dengan multi-label (lulus, berpotensi lulus, aktif, berpotensi *dropout* dan *dropout*) dengan menggunakan 7 fitur. *Dataset* yang digunakan juga dari rentang 2015-2024 (10 tahun) yang semakin memperkuat penelitian ini. Dengan demikian, penelitian ini melengkapi kekurangan dari penelitian terdahulu dalam melakukan klasifikasi mahasiswa berpotensi *dropout* menggunakan metode Regresi Logistik Multinomial.

2. METODE PENELITIAN

Pada proses penelitian yang dilakukan, diperlukan suatu kerangka atau tahapan-tahapan dalam melakukan penelitian tersebut. Tujuannya adalah agar penelitian tersebut tersistematis dan dapat dilakukan dengan baik. Tahapan-tahapan penelitian ini diantaranya adalah pengumpulan data, praprosesing data, klasifikasi data, interpretasi data dan hasil penelitian. Metode yang digunakan pada penelitian ini adalah Regresi Logistik Multinomial.



Gambar 1. Alur Penelitian

2.1 Pengumpulan Data

Tahap ini merupakan tahapan awal dalam melakukan penelitian. Pengumpulan data adalah proses mengumpulkan informasi atau fakta yang relevan dan akurat untuk keperluan penelitian, analisis, atau pengambilan keputusan. Data yang dikumpulkan dapat berupa angka (data kuantitatif) atau uraian/kata (data kualitatif), tergantung dari tujuan pengumpulan tersebut [17]. Data yang diperoleh, diambil melalui pihak Prodi Sistem Informasi, Fakultas Sains dan Teknologi. Jenis data yang digunakan adalah data sekunder yaitu merupakan data yang diperoleh dari pihak ketiga. Dengan pendekatan penelitian yaitu kuantitatif. Data yang dikumpulkan juga merupakan data tahun masuk, lama studi, IPS, IPK, SKS semester, Total SKS dan biaya kuliah. Dari hasil pengumpulan data yang telah dilakukan, diperoleh *dataset* yang berjumlah 1943 data. Data ini merupakan data mahasiswa dari Angkatan 2015 sampai 2024. Data yang dikumpulkan ini masih dalam bentuk mentah sehingga perlu dibersihkan agar mendapatkan data yang diperlukan untuk meneliti.

2.2 Praprosesing Data

Praprosesing data merupakan proses penting dalam analisis data mining yang bertujuan untuk membersihkan, mengubah format, dan mempersiapkan data agar lebih mudah dan akurat dalam proses analisis [18]. Pada tahap ini, dilakukan pemrosesan ulang data yang sudah di kumpulkan. Data yang dikumpulkan akan di bersihkan dan membuang kolom-kolom yang tidak diperlukan. Pada tahap ini juga *dataset* akan diberikan label agar dapat di klasifikasi pada perhitungan berikutnya. Ada beberapa tahapan-tahapan yang akan dilakukan dalam praprosesing *dataset* pada penelitian ini, diantaranya adalah :

1. Pembersihan Data

Pada tahapan ini, dilakukan pembersihan data yang bertujuan untuk membuat data mentah lebih layak digunakan. Dikarenakan setiap data yang diperoleh terkadang masih memiliki kolom-kolom yang tidak diperlukan, *missing value* dan sebagainya.

2. Pelabelan Data

Tahap ini merupakan tahapan lanjutan yang dilakukan dengan memberikan label pada data. Pelabelan pada data bertujuan untuk menjadi pedoman terhadap fitur yang akan di hitung nantinya.

3. Normalisasi Data

Tujuan data di normalisasi adalah agar data dapat di hitung menggunakan metode yang digunakan. Karena fitur-fitur yang digunakan pada *dataset* masih memiliki nilai yang tidak seimbang, maka perlu dilakukan normalisasi agar data tersebut menjadi seimbang dan dapat di olah [19]. Normalisasi ini juga menggunakan metode Z-Score dengan teorema sebagai berikut :

$$Z = \frac{X - \mu}{\sigma} \quad (1)$$

X : nilai asli
 μ : rata-rata dari data
 σ : standar deviasi dari data
 Z : nilai standar (z-score)

2.3 Klasifikasi Data

Klasifikasi merupakan salah satu teknik dalam pengolahan data yang bekerja dengan cara objek yang dipergunakan dibagi menjadi kelas-kelas dengan jumlah kelas sesuai dengan yang diinginkan. Klasifikasi dapat menciptakan suatu pola yang dapat memisahkan tiap-tiap kelas data yang bertujuan guna menentukan objek yang tergolong ke dalam kategori tertentu dilihat dari perilaku serta atribut dari kelompok yang telah didefinisikan [20]. Tahapan klasifikasi ini dimulai dari pembagian data menjadi data *train* dan *test*. Data *train* digunakan sebagai proses pembangunan model untuk melakukan prediksi/peramalan. Sedangkan data *test* adalah proses pengujian untuk melihat performa atau keakuratan data. Menghitung nilai *error* dan menghitung akurasi pengklasifikasian. Setelah membagi data, maka langkah berikutnya adalah melatih pemodelan yang kemudian diprediksi. Setelah diprediksi maka akan terlihat hasil klasifikasi yang dilakukan dengan menampilkan nilai akurasi penelitian.

Metode Regresi Logistik Multinomial memiliki dua kegunaan utama, pertama untuk melakukan prediksi keanggotaan kelompok di mana variabel dependen (respon) memiliki sifat politomi (lebih dari dua kategori). Kedua, metode regresi logistik dapat mengidentifikasi factor kesuksesan yang paling berpengaruh [22]. Pada penelitian ini, model akan menormalisasi dataset yang masih memiliki nilai tidak setara pada setiap kolomnya. Dengan menormalisasi 7 fitur atau kelas sehingga dari hasil normalisasi tadi akan dilakukan pemodelan sehingga melihat kecocokan terhadap label yang telah dibuat. Probabilitas bersyarat sehingga $P(y = j|x) = \pi_j(x)$, $j = 0, 1, 2$ persamaannya :

$$P(Y = 0|x) = \frac{1}{1 + e^{g_1(x)} + e^{g_2(x)}} \quad (2)$$

$$P(Y = 1|x) = \frac{1}{1 + e^{g_1(x)} + e^{g_2(x)}} \quad (3)$$

$$P(Y = 2|x) = \frac{1}{1 + e^{g_1(x)} + e^{g_2(x)}} \quad (4)$$

Adapun bentuk persamaan regresi *logistic* multinomial jika diubah ke dalam fungsi logit adalah sebagai berikut :

$$g_1(x) = \ln \left[\frac{P(Y=1|x)}{P(Y=0|x)} \right] = \beta_{10} + \beta_{11}x_1 + \beta_{12}x_2 + \dots + \beta_{1p}x_p \quad (5)$$

$$g_2(x) = \ln \left[\frac{P(Y=2|x)}{P(Y=0|x)} \right] = \beta_{20} + \beta_{21}x_1 + \beta_{22}x_2 + \dots + \beta_{2p}x_p \quad (6)$$

Fungsi *likelihood* dibangkitkan dengan membuat tiga variabel biner dengan kode 0 atau 1 untuk memperlihatkan keanggotaan kelompok dari sebuah pengamatan. Kategori variabel dilakukan seperti berikut: apabila $Y=0$ sehingga $Y_0=1$, $Y_1=0$, dan $Y_2=0$; apabila $Y=1$ sehingga

$Y_0 = 0$, $Y_1 = 1$, dan $Y_2 = 0$; apabila $Y = 2$ sehingga $Y_0 = 0$, $Y_1 = 0$, dan $Y_2 = 1$. Oleh karena itu, berapapun nilai Y , jumlah dari variabel tersebut [21].

2.4 Interpretasi Data

Pada tahap ini akan disajikan hasil akhir yang diperoleh pada penelitian ini. Hasil akhir pada penelitian ini disajikan dalam bentuk output evaluasi model pemodelan menggunakan data pelatihan dan juga *output* dalam bentuk *confusion matrix*. Hasil yang diperoleh juga berupa akurasi klasifikasi yaitu 99%. Sedangkan pada model pengujian juga mendapatkan akurasi sebanyak 99%. Perbedaan nilai data latih dan uji terletak pada *recall* dan *f1-score*. Dimana *recall* pada data latih yaitu 95% sedangkan pada data uji yaitu 88%. *F1-score* data latih adalah 97% sedangkan pada data uji yaitu 92%. Untuk *precision* keduanya sama yaitu 99%. Untuk *confusion matrix* yang digunakan juga berdimensi 5:5. Perbandingan ini diperoleh berdasarkan jumlah label yang digunakan yaitu sebanyak 5 label.

Dalam penentuan hasil akurasi, dilakukan sebuah perbandingan hasil untuk menemukan nilai tersebut. Hasil prediksi data dengan nilai sebenarnya kemudian dibagi dua akan menjadi hasil dari nilai akurasi. Hasil analisis terdapat penyimpangan yang signifikan (bias) tingkat akurasi rendah yang disebabkan oleh adanya kesalahan sistematis. Ketelitian hasil analisis yang baik dengan nilai deviasi rendah. Semakin tinggi akurasi hasil analisis, semakin rendah RSD (*Relative Standard Deviation*) tidak terjadi bias [23]. Dan untuk rumus perhitungan akurasi, presisi, *recall* dan *f1-score* adalah dibawah ini :

$$Accuracy = \frac{\text{Jumlah Prediksi Benar}}{\text{Total Prediksi}} \quad (7)$$

TP = True Positif

TN = True Negatif

FP = False Positif

FN = False Negatif

$$Precision = \frac{TP}{TP+FP} \quad (8)$$

$$Recall = \frac{TP}{TP+FN} \quad (9)$$

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (10)$$

3. HASIL DAN PEMBAHASAN

3.1 Pengumpulan Data

Ditahap ini dilakukan pengumpulan data yang diperoleh dari pihak prodi Sistem Informasi di kampus UINSU Medan. Jenis data ini merupakan data sekunder. Hasil dari pengumpulan data tersebut menunjukkan jumlah data sebanyak 1943 yang diperoleh dari mahasiswa SI Angkatan 2015 – 2024. Berikut ini adalah tampilan *dataset* yang akan digunakan untuk klasifikasi :

Tabel 1. Tabel *Dataset* Penelitian

no	nim	tahun_masuk	lama_studi	ips	ipk	skssemester	total_sks	biaya_kuliah	label
1	621...43	2015	18	0	0	0	0	2500000	3
2	721...01	2015	18	0	0,2	0	30	2500000	3
3	721...02	2015	18	0	3,54	0	151	2500000	1
4	721...03	2015	18	0	3,63	0	148	2500000	1
5	721...04	2015	18	0	3,49	0	144	2500000	1
...

1939	702....22	2024	0	0	0,00	0	0	3724000	1
1940	702....23	2024	0	0	0,00	0	0	3724000	1
1941	702....24	2024	0	0	0,00	0	0	3724000	1
1942	702....25	2024	0	0	0,00	0	0	3724000	1
1943	702....26	2024	0	0	0,00	0	0	4581000	1

3.2 Praprosesing Data

Setelah data dikumpulkan, berikutnya data dilakukan praprosesing. Tahapan praprosesing pada penelitian ini diantaranya adalah pembersihan data, pelabelan dan normalisasi data.

1. Pembersihan Data

Tahapan awal pada praprosesing data ini dengan melakukan pembersihan pada kolom-kolom dataset yang tidak diperlukan. Proses ini dilakukan secara tradisional pada Microsoft Excel. Pada penelitian ini, hanya 7 fitur yang akan digunakan untuk klasifikasi dari yang awalnya berjumlah lebih dari 20 kolom. Penghapusan ini dilakukan karena kolom-kolom tersebut tidak berpengaruh pada penelitian ini.

2. Pelabelan Data

Langkah berikutnya adalah memberikan label pada setiap *dataset*. Pelabelan ini didasari dari kolom “status” pada *dataset* mentah yang masih dipertahankan sebagai acuan pelabelan data. Langkah ini juga masih dilakukan secara tradisional yaitu melalui *Microsoft Excel*. Pada kolom “Status”, terdapat 5 kategori yaitu “aktif”, “cuti”, “lulus”, “keluar”, “nonaktif”. Untuk mahasiswa yang berstatus “aktif” maka akan tergolong dalam label 0. Sedangkan mahasiswa “cuti”, akan digolongkan ke dalam label 2 atau 3.

Penentuan label 2 atau 3 tersebut berdasarkan kategori “tahun_masuk” dan “total_sks”. Jika seorang mahasiswa memiliki total sks ≤ 100 dan merupakan mahasiswa Angkatan 2019 kebawah, maka mahasiswa tersebut tergolong dalam label “berpotensi DO” atau label 3. Sedangkan mahasiswa yang memiliki total sks ≥ 100 dan merupakan Angkatan 2020 keatas, maka tergolong dalam label “berpotensi lulus” atau label 2. Untuk mahasiswa Angkatan 2015, 2016, 2017 dan 2018, jika memiliki total sks ≥ 130 maka tergolong dalam label 2. Untuk mahasiswa Angkatan 2019, jika memiliki total sks ≥ 110 juga tergolong dalam label 2.

Selanjutnya untuk mahasiswa yang berstatus “lulus” maka digolongkan dalam label “lulus” atau label 1. Dan untuk mahasiswa yang berstatus “keluar” dan “nonaktif”, akan digolongkan dalam label “DO” atau label 4. Untuk mahasiswa yang berstatus “aktif” namun merupakan Angkatan 2020 kebawah, maka akan dilakukan pelabelan berdasarkan hasil inferensi pada mahasiswa yang berstatus “cuti” tadi.

3. Normalisasi Data

Tahapan terakhir praprosesing dimana dataset setelah di bersihkan dan di beri label, maka dilakukan normalisasi. Proses normalisasi dilakukan pada *Google Colab* dengan menggunakan metode *Z-Score* dan *library Sklearn*. Sebelum dinormalisasi, *dataset* terlebih dahulu di *split* melalui *Google Colab*. *Split* data yang digunakan adalah 80% : 20% dengan data latih sebanyak 1554 dan data uji sebanyak 389. Dibawah ini merupakan tampilan dari *dataset* setelah di praprosesing :

	tahun_masuk	lama_studi	lpa	lpb	skor_materi	total_sks	biaya_kuliah	label
0	1.215532	1.217457	1.048874	0.489528	1.098665	1.223337	0.045500	1.048868
1	-0.343333	0.342253	-0.096386	0.582730	-0.824513	0.752231	0.356424	0.852151
2	-0.732779	0.732180	-0.096386	0.577489	-0.824513	0.752779	0.356424	0.852151
3	1.215532	1.217457	1.048874	0.489528	1.098665	1.223337	0.045500	1.048868
4	0.528932	-0.527525	1.248866	0.455399	1.325686	-0.542223	-1.181523	-1.049569
...
1545	0.549541	-0.547575	1.017021	0.425581	0.528810	0.732287	-2.185531	0.852151
1660	0.435321	0.437602	0.774906	0.425549	1.669746	0.143711	1.166551	1.049069
1661	0.143039	0.142253	0.830065	0.253777	0.054915	1.223337	2.610641	1.773811
1992	0.528932	-0.527525	1.175188	0.547827	1.588262	-0.557588	-1.181523	-1.049569
1993	0.549541	-0.547575	0.889120	0.510290	0.388023	0.552342	-0.380128	0.852151

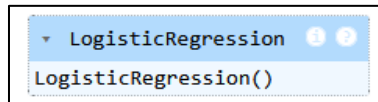
Gambar 2. Normalisasi Data Pelatihan

	tahun_masuk	lama_studi	lpa	lpb	skor_materi	total_sks	biaya_kuliah	label
0	1.215532	1.217457	0.969066	0.562730	1.098665	1.223337	0.045500	1.048868
1	0.528932	-0.527525	1.063050	0.521270	1.590352	-0.537585	-1.067523	-1.049569
2	1.215532	1.217457	1.248866	0.489528	1.325686	1.223337	0.045500	1.048868
3	0.549541	-0.547575	1.111101	0.625580	0.688510	0.756231	2.683534	0.852151
4	1.175188	1.175188	0.889120	0.510290	0.388023	0.552342	0.045500	1.048868
...
284	1.215532	1.217457	0.969066	0.562730	1.098665	1.223337	0.045500	1.048868
285	-0.732779	0.732180	-0.096386	0.577489	-0.824513	0.752287	1.253116	-2.185531
286	-0.732779	0.732180	-0.096386	0.577489	-0.824513	0.752287	1.185286	0.852151
287	0.549541	-0.547575	1.248866	0.621720	0.688510	0.756231	-0.880186	0.852151
288	0.528932	-0.527525	1.175188	0.547827	1.588262	-0.557588	-1.181523	-1.049569

Gambar 3. Normalisasi Data Pengujian

3.3 Klasifikasi Dataset

Gambar 2 dan 3 merupakan hasil dari normalisasi *dataset* untuk model pelatihan dan pengujian. Kedua model ini akan dilakukan klasifikasi untuk melihat hasil akhirnya. Langkah berikutnya adalah melakukan pemodelan menggunakan metode Regresi Logistik :



Gambar 4. Tampilan Model Regresi Logistik Pada Google Colab

Setelah membuat pemodelan regresi logistik pada *Google Colab*, selanjutnya melakukan prediksi nilai untuk memperoleh evaluasi model pelatihan dan juga pengujian beserta perhitungan nilai akurasi pada data latih dan data uji.

Evaluasi Model Pelatihan:					
Akurasi: 0.99					
		precision	recall	f1-score	support
0	1.00	1.00	1.00	1.00	622
1	1.00	1.00	1.00	1.00	290
2	0.98	1.00	0.99	0.99	497
3	0.98	0.95	0.96	0.96	126
4	1.00	0.91	0.90	0.90	13
accuracy				0.99	1554
macro avg		0.99	0.97	0.97	1554
weighted avg		0.99	0.98	0.98	1554
Matriks Konfusi :					
[[622 0 0 0 0]					
[0 290 1 0 0]					
[0 0 497 0 0]					
[0 0 7 121 0]					
[0 0 0 2 13]]					

Gambar 5. Output Visualisasi Model Pelatihan

$$Accuracy = \frac{\text{Jumlah Prediksi Benar}}{\text{Total Prediksi}} = \frac{622 + 290 + 497 + 121 + 13}{1554} = \frac{1543}{1554} = 0,9931$$

Klasifikasi Pengujian:				
	precision	recall	f1-score	support
0	1.00	1.00	1.00	169
1	1.00	1.00	1.00	81
2	0.97	1.00	0.99	101
3	0.97	0.92	0.94	35
4	1.00	0.50	0.67	7
accuracy			0.99	389
macro avg	0.99	0.88	0.92	389
weighted avg	0.99	0.92	0.95	389
Matriks Konfusi:				
[[169 0 0 0 0]				
[0 81 0 0 0]				
[0 0 101 0 0]				
[0 0 3 35 0]				
[0 0 0 1 1]]				

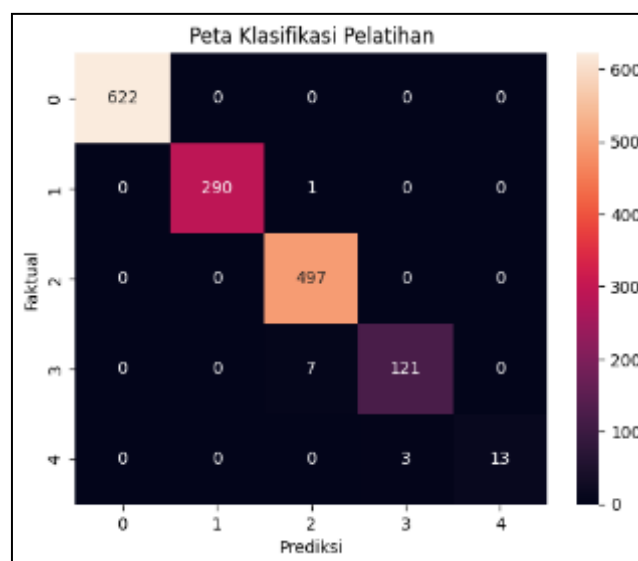
Gambar 6. Output Visualisasi Model Pengujian

$$Accuracy = \frac{\text{Jumlah Prediksi Benar}}{\text{Total Prediksi}} = \frac{169 + 81 + 101 + 33 + 1}{389} = \frac{385}{389} = 0,9897$$

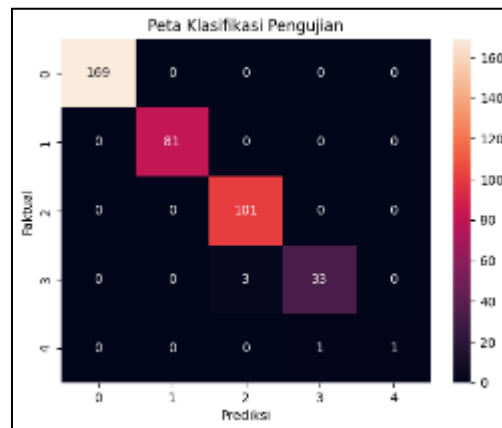
Berdasarkan gambar 5 dan 6, dapat dilihat hasil evaluasi kedua model yaitu pelatihan dan pengujian juga dilakukan perhitungan akurasi secara manual untuk melihat ketepatan hasil dalam perhitungan manual terhadap model yang didapat. Keduanya memiliki akurasi yang sama yaitu 99%. Namun pada *recall* dan *f1-score* jumlahnya sedikit berbeda yaitu pada model pelatihan, *recall* bernilai 95% dan *f1-score*nya 97%. Sedangkan pada model pengujian, *recall* bernilai 88% dan *f1-score* 92%. Namun yang hanya menjadi patokan pada penelitian ini adalah model pelatihan.

3.4 Interpretasi Data

Dari hasil pelatihan dan pengujian yang telah dilakukan hingga menemukan hasil evaluasi model, hasil tersebut kemudian di visualisasikan pada sebuah confusion matrix. Confusion matrix yang digunakan berdimensi 5 : 5. Dimensi ini diperoleh dari jumlah label yang dimiliki. Berikut ini adalah output dari confusion matrix data pelatihan dan pengujian :



Gambar 7. Confusion Matrix Model Pelatihan

Gambar 8. *Confusion Matrix* Model Pengujian

Dari hasil uji coba yang telah dilakukan, dapat dilihat hasil akhir pada penelitian ini yaitu pada model pelatihan, *confusion matrix* menunjukkan hasil klasifikasi berupa nilai pada label 0 sebanyak 622. Sedangkan pada label 1 yaitu mahasiswa lulus sebanyak 290. Pada label 2 yaitu berpotensi lulus, jumlahnya lumayan banyak yaitu 497. Pada label 3 yaitu berpotensi *dropout*, jumlahnya mencapai 121. Dan terakhir yaitu label 4 *dropout* menunjukkan nilai sebanyak 13. Jumlah ini sangat sedikit untuk rentang data akademik mahasiswa prodi Sistem Informasi dari angkatan tahun 2015 sampai 2024. Ini merupakan hasil yang sangat baik, dimana akurasi penelitian mencapai 99% dengan jumlah *dropout* yang sangat sedikit pada saat diklasifikasi yaitu 13 data. sedangkan yang berpotensi *dropout* sebanyak 121 data.

4. KESIMPULAN

Dari hasil penelitian yang telah dilakukan, maka dapat diambil kesimpulan yaitu program studi Sistem Informasi pada Universitas Islam Negeri Sumatera Utara dapat menekan angka *dropout* pada prodi tersebut. Terbukti dari hasil klasifikasi yang telah dilakukan, dimana pada hasil model pelatihan yang telah dilakukan memberikan hasil pada label 4 hanya 13 data dari 1554 dataset yang diperoleh saat *splitting* data untuk data latih. Dengan akurasi sebesar 99%, akurasi tersebut hampir mendekati sempurna. Sedangkan untuk label 3, jumlah yang berhasil diklasifikasi sebanyak 121 data. ini menunjukkan kinerja antar mahasiswa dan dosen sudah sangat baik sehingga kasus *dropout* hanya sedikit dari banyaknya mahasiswa yang terdaftar. Sedangkan *dataset* yang digunakan merupakan data akademik dari mahasiswa angkatan 2015 sampai 2024. Namun untuk label 3 yaitu berpotensi *dropout*, masalah tersebut dapat di evaluasi ulang kembali agar kasus mahasiswa yang berpotensi *dropout* juga mengalami penurunan.

DAFTAR PUSTAKA

- [1] Jaedi, B. Supriyono and M. R. K. Muluk, "Evaluasi Kebijakan Masa Studi di Perguruan Tinggi (Studi di Fakultas Ilmu Administrasi Universitas Brawijaya)," *Jurnal Ilmiah Administrasi Publik (JIAP)*, vol. 5, no. 3, pp. 268-276, 2019.
- [2] R. Yendra, L. Marifni and I. Suryani, "Klasifikasi Data Mining Untuk Seleksi Penerimaan Calon Pegawai Negeri Sipil Tahun 2017 Menggunakan Metode Naïve Bayes," *Jurnal Sains Matematika dan Statistika*, vol. 6, no. 1, pp. 65-78, 2020.
- [3] R. H. Situngkir and P. Sembiring, "Analisis Regresi Logistik Untuk Menentukan Faktor-Faktor Yang Mempengaruhi Kesejahteraan Masyarakat Kabupaten/Kota Di Pulau Nias," *Jurnal Matematika dan Pendidikan Matematika (FARABI)*, vol. 6, no. 1, pp. 25-31, 2023.

- [4] F. Arofah and A. Sofro, "Penerapan Regresi Logistik Multinomial Untuk Analisis Model Tingkat Depresi Pada Lansia," *Jurnal Ilmiah Matematika (MATHunesa)*, vol. 10, no. 1, pp. 84-93, 2022.
- [5] R. Syahrani, Suhartono and S. Zaman, "Regresi Logistik Multinomial untuk Prediksi Kategori Kelulusan Mahasiswa," *JISKA(Jurnal Informatika Sunan Kalijaga)*, vol. 8, no. 2, pp. 102-111, 2023.
- [6] Nurmalitasari and E. Purwanto, "Prediksi Performa Mahasiswa Menggunakan Model Regresi Logistik," *Jurnal Derivat*, vol. 9, no. 2, pp. 145-152, 2022.
- [7] N. Ulinnuha and A. Fanani, "Klasifikasi Status Drop Out Mahasiswa Menggunakan Naïve Bayes dengan Seleksi Fitur Information Gain," *Techno.COM*, vol. 22, no. 4, pp. 1014-1025, 2023.
- [8] Salmawati, Yuyun and Hazriani, "Klasifikasi Mahasiswa Berpotensi Drop Out Menggunakan Algoritma Naive Bayes dan Decision Tree," *Jurnal Ilmiah Ilmu Komputer*, vol. 8, no. 2, pp. 108-114, 2022.
- [9] M. P. Agustini, A. F. Supianto and W. Purnomo, "Aplikasi Data Mining untuk Memprediksi Mahasiswa Berpotensi Drop Out menggunakan Algoritme K-Nearest Neighbor (K-NN)," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, vol. 3, no. 6, pp. 6230-6237, 2019.
- [10] A. Ramadhani, R. F. Noor, D. Vernanda and T. Herdiawan, "Klasifikasi Mahasiswa Berpotensi Drop Out Menggunakan Algoritma C4.5 di Politeknik Negeri Subang," *urnal TEKNO KOMPAK*, vol. 18, no. 1, pp. 101-112, 2024.
- [11] V. Tamuntuan, Kusriani and Kusnawi, "Perbandingan Kinerja Algoritma Support Vector Machine dan Neural Network Backpropagation untuk Mengklasifikasikan Mahasiswa Berpotensi Dropout," *Building of Informatics, Technology and Science (BITS)*, vol. 6, no. 2, pp. 847-855, 2024.
- [12] R. Kamsurya, M. A. M. Hayat and R. Y. Bakti, "Klasifikasi Mahasiswa Berpotensi Putus studi menggunakan Algoritma Naive Bayes pada Fakultas Teknik Unismuh Makassar," *Arus Jurnal Sains dan Teknologi (AJST)*, vol. 2, no. 2, pp. 511-517, 2024.
- [13] D. Sinaga, E. J. Solaiman and F. J. Kaunang, "Penerapan Algoritma Decision Tree C4.5 Untuk Klasifikasi Mahasiswa Berpotensi Drop out Di Universitas Advent Indonesia," *Jurnal TelKa*, vol. 11, no. 2, pp. 167-173, 2021.
- [14] U. F. Laili, C. Umatin and M. U. Ridwanulloh, "Analisis Potensial Drop Out Mahasiswa Dengan K-Means++ Clustering Dalam Upaya Peningkatan Kualitas Iain Kediri," *Paedagoria : Jurnal Kajian, Penelitian dan Pengembangan Kependidikan*, 14 (2), pp. 145-153, 2023.
- [15] J. A. Putra and B. Septian, "Penerapan Algoritma Naive Bayes dalam Memprediksi Tingkat Mahasiswa Drop Out pada Program Studi Informatika," *Digital Transformation Technology (Digitech)*, vol. 4, no. 2, pp. 1227-1235, 2024.
- [16] S. Fitriana and dkk, "Prediksi Siswa Putus Sekolah dan Keberhasilan Akademik Menggunakan Machine Learning," *The Indonesian Journal of Computer Science*, vol. 13, no. 6, pp. 10207-10220, 2024.
- [17] Ardiansyah, Risnita and M. S. Jailani, "Teknik Pengumpulan Data Dan Instrumen Penelitian Ilmiah Pendidikan Pada Pendekatan Kualitatif dan Kuantitatif," *IHSAN : Jurnal Pendidikan Islam*, vol. 1, no. 2, pp. 1-9, 2023.
- [18] A. A. A. Daniswara and I. K. D. Nuryana, "Data Preprocessing Pola Pada Penilaian Mahasiswa Program Profesi Guru," *(JINACS) Journal of Informatics and Computer Science*, vol. 5, no. 1, pp. 97-100, 2023.

- [19] A. C. Kamilla, N. Priyani, R. Priskila and V. H. Pranatawijaya, "Analisis Sentimen Film Agak Laen Dengan Kecerdasan Buatan : Text Mining Metode Naïve Bayes Classifier," *JATI (Jurnal Mahasiswa Teknik Informatika)* , vol. 8, no. 3, pp. 2923-2928, 2024.
- [20] N. Widiastuti, A. Hermawan and D. Avianto, "Implementasi Metode Naive Bayes Untuk Klasifikasi Data Blogger," *JUPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)*, vol. 8, no. 3, pp. 985-994, 2023.
- [21] S. S. Dewi, Resmawan and L. O. Nashar, "Analisis Regresi Logistik Multinomial dengan Metode Bayes untuk Identifikasi Faktor-Faktor Terjadinya Diabetes Melitus," *Journal of Mathematics: Theory and Applications (JOMTA)*, vol. 5, no. 2, pp. 51-60, 2023.
- [22] Y. Malik, "Akurasi dan Presisi Analisis Kadar Nikel (Ni) pada Sampel Nikel Laterit Menggunakan X-Ray Fluorescence Spectrometry (XRF)," *(SAINS) Jurnal Kimia dan Pendidikan Kimia*, vol. 12, no. 2, pp. 87-94, 2023.
- [23] A. Ermilian and K. Nugroho, "Perancangan Model Deteksi Potensi Siswa Putus Sekolah Menggunakan Metode Logistic Regression Dan Decision Tree," *Jurnal Informatika: Jurnal pengembangan IT*, vol. 9, no. 3, pp. 281-295, 2024.