DOI: https://doi.org/10.26593/jrsi.v12i2.6595.187-200

Jurnal Rekayasa Sistem Industri
Volume 12 No 2 - Oktober 2023
http://journal.unpar.ac.id/index.php/jrsi/index
ISSN 2339-1499 (*online*) – ISSN 0216-1036 (*print*)

Disampaikan : 3 April 2023
Direview : 5 April 2023
Diterima : 19 Mei 2023

Model Prediksi dengan Pembelajaran Mesin dalam Pemberian Program Beasiswa Kepada Calon Mahasiswa Baru Program S1 Di Sebuah Perguruan Tinggi Swasta.

Gideon Budiyanto[1], Dedy Suryadi[2]
[1,2)] Fakultas Teknologi Industri, Jurusan Teknik Industri, Universitas Katolik Parahyangan
Jl. Ciumbuleuit 94, Bandung 40141
Email: 8132101001@student.unpar.ac.id, dedy@unpar.ac.id

Abstract

Competition in the higher education, especially private higher education (PTS) in the digital era, is becoming increasingly tough. In order to achieve the number of prospective new students, various methods are used so that the target for admitting the number of new students can be achieved in each new academic year. Providing a scholarship program is one way to attract the prospective new students. The awarding of a scholarship program must consider various possibilities such as the seriousness or commitment of the prospective new student. Refusal to grant scholarship programs can occur and become an obstacle for achieving the target. The prediction model through machine learning using some variables such as high school's name, high school "category", province or area of high school located, focus of specialization in high school, high school's grade, type of parents income, and selected major of study in higher education. All of those variables will provides the probability values that will become an indicator that can be used to prioritize requests for scholarship program applications by taking into account the factors of acceptance or rejection from prospective students. Currently there is no measurement with accuracy of acceptance or rejection from prospective students. The purpose of this research is to build and compare machine learning models such as Logistic Regression, Artificial Neural Networks, Support Vector Machines, Decision Trees, Naïve Bayes, and K Nearest Neighbors so that a machine learning model is obtained that has the best predictions for awarding scholarship programs. The result of this research is that the Logistic Regression model has the highest model average accuracy value (62,05%) from training data compared to others. The highest accuracy of Logistic Regression model (62,29%) achieved based on the testing data. The highest AUC value (0,818) generated by Logistic Regression model which means the model is able to do the classification categorized "Good Classification" compare to other models.

Keywords: *machine learning, scholarship program, private higher education, prediction system*

Abstrak

Persaingan di dalam dunia pendidikan tinggi secara khusus Perguruan Tinggi Swasta (PTS) terutama di era digital menjadi semakin ketat. Dalam memperebutkan jumlah calon mahasiswa baru yang tersedia, berbagai cara dilakukan agar target penerimaan jumlah mahasiswa baru dapat tercapai. Pemberian program beasiswa adalah salah satu cara menjaring calon mahasiswa baru. Pemberian program beasiswa harus mempertimbangkan berbagai kemungkinan seperti keseriusan atau komitmen sedangkan penolakan pemberian program beasiswa dapat juga terjadi dan menjadi kendala pada akhir suatu periode Penerimaan Mahasiswa Baru (PMB). Model prediksi melalui pembelajaran mesin dengan beberapa atribut seperti asal sekolah SMA, "Kategori Sekolah" SMA, provinsi atau daerah asal SMA, jurusan saat SMA yang diambil, nilai akademik SMA, jenis pekerjaan orang tua, dan pilihan program studi atau jurusan yang akan diambil saat nanti berkuliah pada akhirnya dapat memberikan suatu indikator nilai peluang atau kemungkinan penerimaan atau penolakan program beasiswa dari seorang calon mahasiswa baru. Saat ini belum ada usaha untuk memprediksi secara sistematis terhadap penerimaan / penolakan program beasiswa. Tujuan penelitian ini adalah membangun dan membandingkan model pembelajaran mesin seperti *Logistic Regression, Artificial Neural Network,*

DOI: https://doi.org/10.26593/jrsi.v12i2.6595.187-200

Support Vector Machine, Decision Tree, Naïve Bayes, dan *K Nearest Neighbors* sehingga didapatkan satu model pembelajaran mesin yang memiliki prediksi yang terbaik terhadap pemberian program beasiswa. Dari hasil penelitian maka model *Logistic Regression* memiliki nilai akurasi rata-rata tertinggi (62,05%) saat melakukan pembelajaran model dengan data latihan dibandingkan dengan model lainnya. Akurasi model *Logistic Regression* memiliki nilai tertinggi terhadap data uji sebesar (62,29%) dan juga memiliki nilai AUC (0.818) yang berarti bahwa model dapat melakukan pengklasifikasian dengan baik terhadap kelompok pengambilan keputusan dibandingkan dengan model lainnya.

Kata kunci: pembelajaran mesin, program beasiswa, perguruan tinggi swasta, sistem prediksi.

Pendahuluan

Seiring dengan banyaknya jumlah Perguruan Tinggi Swasta atau dikenal dengan PTS di Indonesia, maka tingkat persaingan antara PTS menjadi semakin sengit. Berbagai perbaikan mutu dari sarana dan prasarana sebuah PTS menjadi modal untuk memiliki daya saing dalam berkompetisi antara sesama PTS. Berdasarkan data (PDDikti Kementerian Riset, Teknologi, dan Pendidikan Tinggi Republik Indonesia), terdapat peningkatan jumlah mahasiswa baru jenjang program sarjana S1 bagi PTS di beberapa provinsi tahun 2015 – 2020. Dan pada rentang tahun yang sama terjadi penurunan jumlah PTS itu sendiri. Data dapat dilihat pada Tabel 1. dan Tabel 2.

Tabel 1. Jumlah mahasiswa baru

Daerah	2017	2018	2019	2020
D.K.I. Jakarta	99,319	91,576	120,640	113,701
Jawa Barat	106,034	102,398	116,079	127,330
Banten	43,838	47,025	49,528	54,431
Jawa Tengah	61,604	61,908	73,762	77,269
Jogyakarta	42,337	42,381	50,206	45,693
Jawa Timur	89,852	88,899	112,754	108,842
Total	**442.984**	**434.187**	**522.969**	**527.266**

Tabel 2. Jumlah Perguruan Tinggi Swasta

Daerah	2015	2016	2017	2018	2019	2020
D.K.I. Jakarta	317	317	318	315	291	284
Jawa Barat	377	380	380	385	389	377
Banten	114	116	118	121	117	118
Jawa Tengah	248	253	256	271	263	259
Jogyakarta	106	107	108	106	106	103
Jawa Timur	326	329	328	320	337	328
Total	**1.488**	**1.502**	**1.508**	**1.518**	**1.503**	**1.469**

Momentum ini dimanfaatkan oleh sejumlah PTS untuk menjaring sebanyak mungkin calon mahasiswa baru. Ironisnya bahkan banyak PTS yang tidak dapat bersaing dalam memperebutkan jumlah calon mahasiswa baru yang ada dan berujung kepada tidak dapat berkembangnya PTS tersebut.

Direktorat Jenderal Pendidikan Tinggi, Riset, dan Teknologi (Ditjen Diktiristek) mengadakan akselerasi program penggabungan atau penyatuan Perguruan Tinggi Swasta (PTS) sebagai akibat banyaknya PTS yang tidak dapat berkembang dan tergolong PTS yang tidak sehat. (Herlina, 2021) Ditjen Diktiristek menargetkan pengurangan Perguruan Tinggi Swasta di Indonesia sebanyak 1.000 sampai dengan tahun 2024 dengan langkah program Merger antara sesama PTS yang tidak sehat.

PTS merancang berbagai strategi pemasaran untuk memastikan bahwa jumlah penerimaan mahasiswa baru atau dikenal dengan PMB mencapai jumlah yang telah ditargetkan oleh sebuah institusi pendidikan. Salah satu contohnya adalah penerapan *Marketing Mix Strategy* dalam meningkatkan PMB (Tukur at al., 2019). Beragam program beasiswa bagi calon mahasiswa baru diluncurkan agar menarik perhatian calon mahasiswa baru dan orang tua calon mahasiswa baru dalam memutuskan untuk melanjutkan pendidikan tingginya di sebuah PTS (Delima, 2019). Jumlah PMB di awal tahun pendidikan menjadi sebuah indikator untuk memastikan bahwa rencana program kerja baik secara operasional dan pengembangan institusi pendidikan secara umum dapat tercapai dari sisi keuangan. (Trusheim dan Rylee, 2011).

Pemberian program beasiswa kepada calon mahasiswa baru PTS dilakukan dengan beragam prosedur dan berbagai persyaratan. Lolosnya calon mahasiswa baru mendapatkan program beasiswa tidaklah menjamin bahwa calon mahasiswa baru pasti akan memanfaatkan kesempatan yang berharga tersebut. Hal ini juga menjadi permasalahan yang umum di setiap dunia pendidikan tinggi (Kanadpriya at al., 2019). Mencari kesempatan yang lebih baik di PTS lainnya untuk mendapatkan program beasiswa yang lebih atau diterimanya calon mahasiswa baru di Perguruan Tinggi Negeri menjadi pertimbangan tertentu sehingga pemberian program beasiswa menjadi tidak tepat sasaran dan tidak efektif dan efisien. Kosongnya sejumlah bangku yang

DOI: https://doi.org/10.26593/jrsi.v12i2.6595.187-200

sudah disiapkan dan mencari pengganti untuk mengisi kekosongan tersebut tidaklah dapat dengan cepat dan mudah dilakukan. Hal ini bisa diminimalisasi dan dihindari jika pemberian program beasiswa kepada calon mahasiswa baru benar-benar kepada yang membutuhkan dan serius untuk berkuliah.

Penelitian sistem prediksi dengan menggunakan pembelajaran mesin di dalam dunia institusi pendidikan tinggi secara umum digunakan untuk beberapa hal seperti:

1. *Enrollment* atau Penerimaan Mahasiswa Baru (PMB), jenis penelitian ini bertujuan membangun sistem prediksi gagal/lulusnya calon mahasiswa baru dari proses seleksi penerimaan mahasiswa baru.
   Data yang digunakan meliputi data PMB. Data PMB adalah data yang berhubungan dengan calon mahasiswa baru ketika mengikuti proses penerimaan mahasiswa, data-data tersebut berhubungan dengan latar belakang keluarga, latar belakang pendidikan, dan latar belakang nilai-nilai akademis beberapa mata pelajaran saat SMA (Basheer at al., 2019), (Nakhkob & Khadem, 2015), (Slim at al., 2018), (Harani & Prianto, 2020), (Alaka, 2017), dan (Kanadpriya at al. 2019).
2. *Student Performance* atau keberhasilan studi seorang mahasiswa, penelitian ini bertujuan membangun sistem prediksi untuk beberapa hal seperti:
   a. Kegagalan atau keberhasilan seorang mahasiswa dalam proses pembelajaran suatu mata kuliah tertentu. (Tomasevic at al., 2019), (Fernandes at al., 2018), (Ahmed at al., 2021). Penelitian lainnya yang bertujuan membangun sistem prediksi yang mengelompokan hasil kelulusan suatu mata kuliah berdasarkan indikator nilai Indeks Prestasi. (Yagci, 2022).
   b. Memprediksi pengelompokan seorang mahasiswa dalam menempuh proses pendidikan di suatu institusi pendidikan tinggi berdasarkan hasil Indeks Prestasi yang dicapai. (Sulastri at al 2021), (Acharya & Sinha, 2014), (Pumpuang at al., 2008)
   c. Memprediksi pengelompokan seorang mahasiswa setelah diselesaikannya proses pendidikan di suatu institusi pendidikan tinggi berdasarkan lama waktu yang dibutuhkan dimulai awal hingga lulus berkuliah. (Ploutz, 2018).
3. *Drop Out* atau kegagalan dalam studi sehingga mahasiswa haru meninggalkan institusi pendidikan tersebut karena satu dan lain hal (Kovacic, 2010), (Cardona dan Cudney, 2019), dan (Berens at al., 2019).
4. Model prediksi yang berbasiskan pembelajaran mesin digunakan untuk membantu proses pemberian program beasiswa kepada calon mahasiswa baru. Penelitian terhadap memprediksi jumlah mahasiswa baru dan kemudian mengoptimalkan pemberian beasiswa dari estimasi jumlah mahasiswa baru tersebut merupakan salah satu penelitian yang memulai menghubungkan pembelajaran mesin kepada program beasiswa. (Aulck at al., 2020). Penelitian terhadap memprediksi besaran nilai secara finansial beasiswa yang akan diterima oleh calon mahasiswa baru dari berbagai jenis beasiswa yang tersedia juga menjadi fokus penelitian (Delima, 2019). Sedangkan pada penelitian ini berfokus kepada memprediksi respons yang terjadi saat pemberian program beasiswa dilaksanakan, yaitu tidak lulus seleksi, dikategorikan “Tidak Lulus”, lulus seleksi akan tetapi melepas kesempatan tersebut, dikategorikan “Lepas’, dan yang terakhir adalah lulus seleksi dan menerima program beasiswa, dikategorikan “Terima”.

Belum banyaknya penelitian yang berfokus kepada pemberian beasiswa dan memprediksi penerimaan dari para penerima beasiswa menjadi perhatian utama dalam penelitian ini.

Model prediksi yang berbasiskan pembelajaran mesin dapat digunakan juga untuk membantu proses pemberian program beasiswa kepada calon mahasiswa baru. Penggunaan komputer untuk mengolah sejumlah variabel dan mempelajari pola dari data masa lampau untuk menghasilkan model prediksi yang akurat dalam waktu yang singkat sangat mudah untuk dicapai. Akan tetapi menjadi hal tersebut menjadi sulit jika prediksi tersebut dilakukan secara manual. Perumusan masalah yang dapat dibuat berdasarkan identifikasi masalah adalah bagaimana mengembangkan sistem prediksi dalam pemberian program beasiswa untuk calon mahasiswa baru agar tingkat penerimaan

DOI: https://doi.org/10.26593/jrsi.v12i2.6595.187-200

program beasiswa oleh calon mahasiswa baru dapat maksimal.

Landasan Teori

Logistic Regression (LR)

Regresi logistik adalah suatu pendekatan model matematika untuk menganalisis hubungan antara satu variabel terikat atau *dependent variable* yang bersifat dikotomi atau politomi dengan satu atau lebih variabel bebas atau *independent variable* yang bersifat ordinal, nominal maupun rasio. Maka nilai probabilitas atau nilai fungsi regresi logistik dituliskan sebagai berikut (Hosmer, 2013)

$$\pi(\mathbf{x}) = \frac{e^{\beta_0+\beta_1\mathbf{x}}}{1 + e^{\beta_0+\beta_1\mathbf{x}}} \qquad \text{Pers. 1}$$

$$g(\mathbf{x}) = ln\left[\frac{\pi(\mathbf{x})}{1 - \pi(\mathbf{x})}\right] = \beta_0 + \beta_1\mathbf{x} \qquad \text{Pers. 2}$$

Dimana :

$\pi(x)$:nilai probabilitas, nilai fungsi regresi logistik

β_0 : nilai intersepsi

β_1 : koefisien regresi dari variabel bebas x

g(x) : fungsi logit atau transformasi logit, jika nilai g(x) dari indeks nilai variabel bebas yang bervariasi dari -∞ hingga +∞ , maka nilai $\pi(x)$ juga bervariasi dari 0 hingga 1.

Support Vector Machine (SVM)

Konsep SVM dapat dijelaskan secara sederhana sebagai usaha mencari nilai *margin hyperplane* yang paling maksimum atau dikenal *the maximum marginal hyperplane*. *Hyperplane* ini berfungsi sebagai garis pemisah secara maksimal antara dua buah kelas. Jika garis *hyperplane* sudah ditemukan maka dapat diprediksi pengelompokan suatu data yang baru. Jika data bersifat *non linearly separable*, maka dapat menggunakan *non linear mapping* atau pemetaan secara *non linier* untuk mengubah data ke ruang dimensi yang lebih tinggi serta mencari nilai *margin hyperplane* yang paling maksimum di ruang dimensi yang lebih tinggi. (Aggarwal, 2015)

Suatu data D (X_1, y_1), (X_2, y_2), … ($X_{|D|}$, $y_{|D|}$), dimana X_i adalah suatu data latihan yang memiliki label yi dimana yi merupakan anggota dari dua buah kelas : +1 dan -1 , ($y_i \in \{+1, -1\}$. Fungsi garis pemisah *hyperplane* dituliskan : (Han, 2015)

$$\mathrm{W}.\mathrm{X} = \mathrm{b} + 0 \qquad \text{Pers. 3}$$

Dimana :

W : bobot vektor yang terdiri dari sejumlah n atribut W = (W_1, W_2, … , W_n)

b : nilai bias atau bobot tambahan

X : data pelatihan yang memiliki sejumlah n atribut X = (X_1, X_2, …. X_n)

Jika suatu data X (X_1, X_2) terdiri dari W (W_1, W_2) serta nilai b (bobot tambahan/bias) W_0, dan y_i adalah katagori kelas maka persamaan tersebut dituliskan :

$$\mathrm{W}_0 + \mathrm{W}_1\mathrm{X}_1 + \mathrm{W}_2\mathrm{X}_2 = 0 \qquad \text{Pers. 4}$$

Data yang terletak tepat atau diatas garis H_1 akan di kategorikan kelas +1 dan yang terletak tepat atau di bawah garis H_2 akan di kategorikan kelas -1. Sedangkan area yang berada diantara garis H_1 dan H_2 disebut *Support Vector*. Jarak dari garis *hyperplane* ke titik-titik di garis H_1 adalah 1/||W|| sedangkan ||W|| adalah jarak *Euclidean*. Hal yang sama berlaku untuk garis H_2 sehingga maksimal margin yang terjadi adalah 2/||W|| atau sama halnya jika meminimalkan nilai dari ||W||/2. Nilai bobot w hasil minimasi tersebut akan menjadi garis *hyperplane* yang paling optimal.

K-Nearest Neighbor (KNN)

K-Nearest Neighbor (KNN) adalah suatu metode mencari suatu pola dari sebanyak k data latihan yang berjarak lebih dekat dengan data yang tidak diketahui. Sejumlah k data latihan ini adalah k *nearest neighbor* atau k dari tetangga terdekat. Kedekatan ini didefinisikan sebagai jarak antara 2 titik dimana menggunakan metode *Euclidian*. (Han, 2015).

$$\mathrm{dist}(\mathrm{X}_1, \mathrm{X}_2) = \sqrt{\sum_{i=1}^{n} (\mathrm{x}_{1i} - \mathrm{x}_{2i})^2} \qquad \text{Pers. 5}$$

Dimana :

X_1 : data *training*

X_2 : data *testing*

X_{1i} : data *training* ke-i

X_{2i} : data *testing* ke-i

Metode KNN mempunyai langkah-langkah sebagai berikut: (Gorunescu, 2011)

DOI: https://doi.org/10.26593/jrsi.v12i2.6595.187-200

a. Hitung jarak (*Euclidean Distance*) antara data yang akan dievaluasi dengan data dari semua pelatihan.
b. Tentukan nilai parameter K
c. Urutkan jarak yang terbentuk (mulai dari jarak terdekat).
d. Tentukan jarak terdekat sampai urutan ke K.
e. Pasangkan kelas yang bersesuaian.
f. Menentukan jenis kelas dari data yang akan dievaluasi berdasarkan metode *voting* atau mayoritas jenis kelas yang ada.

Naive Bayes (NB)

Metode *Naive Bayes* merupakan sebuah pengklasifikasian probabilistik bersyarat yang menghitung sekumpulan probabilitas dengan menjumlahkan frekuensi dan kombinasi nilai dari *dataset* yang diberikan. Algoritma menggunakan teorema *Bayes* mengasumsikan semua atribut independen atau tidak saling ketergantungan yang diberikan oleh nilai pada variabel kelas (Aggarwal, 2015). Teorema *Bayes* seperti berikut ini:

$$p(Y|X) = \frac{p(X|Y)\,.\,p(Y)}{p(X)} \qquad \text{Pers. 6}$$

Dimana :
X : Data dengan kelas yang belum diketahui.
Y : Hipotesis data X dari suatu kelas spesifik.
p(Y|X): Probabilitas hipotesis Y berdasar kondisi X (*Posterior Probability*)/(*Posterior*).
p(Y): Probabilitas hipotesis Y (*Prior Probability*).
p(X|Y):Probabilitas X berdasarkan kondisi pada hipotesis Y.
p(X) : Probabilitas X (*Evident*).
Denominator dari teorema *Bayes* yaitu p(X) dapat diabaikan karena bersifat tetap

Menurut Aggarwal (Aggarwal, 2015), teorema ini dapat diperjelas sebagai berikut:

$$p(Y = C_k|X_1 \dots X_d) = \frac{p(Y = C_k)\prod_j p(X_j|Y = C_k)}{\sum_i p(Y = y_i)\prod_j p(X_j|Y = y_i)} \qquad \text{Pers. 7}$$

Dimana :
$Y = C_k$: Y adalah variabel diskrit yang terdiri dari 1 sampai k banyaknya kelas
$X_1 \dots X_d$: Data dengan jumlah kelas sebanyak d
Karena denominator bersifat tetap maka dapat diabaikan. Sehingga pengklasifikasian atau penentuan nilai Y didasarkan atas argumen yang bernilai maksimum dari sejumlah k kelas.

$$Y \leftarrow \arg\max_{C_k} p(Y = C_k)\prod_j p(X_j|Y = C_k) \qquad \text{Pers. 8}$$

Decision Trees (DT)

Pohon Keputusan atau dikenal dengan *Decision Trees* digunakan untuk memprediksi dari kelompok suatu objek yang berdasarkan berbagai katagori atau kelas. Metode ini memberikan kelebihan dalam visualisasi berbentuk pohon yang secara sistematis merangkum pengelompokan yang terjadi. (Gorunescu, 2011).

Proses sebuah atribut yang akan dipilih untuk menjadi *node* atau titik awal dari puncak sebuah pohon keputusan disebut juga dengan *Splitting Criterion* (kriteria pemisahan) dimana proses mencari nilai terbaik dari kriteria pemisahan untuk menghasilkan partisi yang lebih kecil dari sebuah sekelompok data tergantung dengan aturan atau cara yang digunakan untuk memisahkan atribut tersebut.

Berikut ini beberapa jenis pengukuran terhadap proses kriteria pemisahan (Han, 2015):

1. Metode ID3 menggunakan *Information Gain* atau "Penguatan Informasi" sebagai cara dalam mengukur atribut dalam proses "Kriteria Pemisahan". Sebuah Atribut yang memiliki nilai tertinggi dari *Information Gain* dipilih sebagai atribut yang akan dipisahkan untuk menjadi sebuah titik awal (*Node*) dari sebuah Pohon Keputusan. Atribut ini membutuhkan informasi yang minimum dalam mengklasifikasikan sebuah kelompok data, atau dikenal dengan istilah *impurity* atau ketidakmurnian yang minimal dari sebuah partisi data.

   Informasi yang dibutuhkan untuk membuat klasifikasi dari sekelompok data dirumuskan (Han, 2015):

$$\text{Info}(D) = -\sum_{i=1}^{m} p_i \log_2(p_i) \qquad \text{Pers. 9}$$

   Dimana :
   D : sekelompok data
   p_i : probabilitas *non zero* (yang bukan bernilai nol) dari suatu kelas C_i dan

DOI: https://doi.org/10.26593/jrsi.v12i2.6595.187-200

diestimasikan $|C_i/D|$ / $|D|$ yang dijumlahkan dari sebanyak m kelas
$\log_2$: fungsi logaritma berbasis nilai 2
Info(D) : *Entropy* dari D

Ketika proses "Kriteria Pemisahan" diharapkan menghasilkan sebuah partisi yang memiliki kemurnian yang terbaik untuk mewakili sekelompok data, akan tetapi partisi tersebut tidak terlepas dari *impurity* (ketidakmurnian). *Impurity* atau ketidakmurnian berarti masih ada sejumlah informasi yang dibutuhkan dari sebuah atribut ketika proses partisi selesai untuk mendekati kesamaan dengan sekelompok data. Kebutuhan informasi tersebut dirumuskan (Han, 2015):

$$\mathrm{Info}_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \mathrm{Info}(D_j) \qquad \text{Pers. 10}$$

Dimana :
$\mathrm{Info}_A(D)$: sejumlah informasi yang masih dibutuhkan untuk mengklasifikasi sekelompok data D berdasarkan atribut A.

$\frac{|D_j|}{|D|}$: bobot dari partisi ke-j

Semakin kecil nilai Info_A (D) mengandung arti semakin murni sebuah partisi berdasarkan atribut yang dipilih.

Information Gain (penguatan informasi) di rumuskan (Han, 2015):

$$\mathrm{Gain}(A) = \mathrm{Info}(D) - \mathrm{Info}_A(D) \qquad \text{Pers. 11}$$

Gain(A) menunjukkan seberapa banyak informasi yang didapatkan jika memilih atribut A. Dan pengurangan informasi yang dikehendaki yang disebabkan dari atribut A. Sebuah atribut A yang memiliki nilai Gain(A) paling besar dipilih untuk menjadi kriteria pemisahan untuk membentuk titik awal (*node*) dari sebuah pohon keputusan. Dengan kata lain atribut A akan menjadi atribut pengelompokan yang terbaik karena memiliki sejumlah informasi yang dibutuhkan secara maksimal atau kekurangan sejumlah informasi yang minimum dalam rangka merepresentasikan sekelompok data (Han, 2015).

2. Metode CART, metode ini menggunakan GINI Index dalam mengukur *impurity* (ketidakmurnian) suatu kelompok data.

$$\mathrm{Gini}(D) = 1 - \sum_{i=1}^{m} p_i^2 \qquad \text{Pers. 12}$$

Dimana :
D : sekelompok data
p_i : probabilitas *non zero* (yang bukan bernilai nol) dari suatu kelas C_i dan diestimasikan $|C_i/D|/|D|$. Dijumlahkan dari sebanyak m kelas
Gini(D) : "Gini" dari D

Dengan memilih atribut yang memiliki nilai paling maksimal dari Δ Gini(A), yang dirumuskan (Han 2015):

$$\Delta \mathrm{Gini}(A) = \mathrm{Gini}(D) - \mathrm{Gini}_A(D) \qquad \text{Pers. 13}$$

nilai Gini A(D) yang memiliki nilai yang paling minimum akan menghasilkan Δ Gini(A) secara maksimal.

Artificial Neural Network (ANN)

Suatu jaringan saraf tiruan terdiri dari sejumlah lapisan dan simpul yang berbeda untuk tiap-tiap lapisannya. Jenis *layer*/lapisan suatu jaringan saraf terdiri dari (Gorunescu, 2011):

1. *Input Layer*: terdiri dari unit-unit simpul yang berperan sebagai input proses pengolahan data pada jaringan
2. *Hidden Layer*: terdiri dari unit-unit simpul yang dianalogikan sebagai lapisan tersembunyi dan berperan sebagai lapisan yang meneruskan respons dari input.
3. *Output Layer*: terdiri dari unit-unit simpul yang berperan memberikan solusi dari data input.

Saraf tiruan diibaratkan sebagai mesin komputasional yang mengubah input menjadi output. Salah satu jenis unit saraf buatan yaitu Sigmoidal Unit dimana unit tersebut terdiri dari 2 komposisi yaitu (Aggarwal, 2015):

1. *Net Value Function* (ξ) , fungsi ini menggabungkan parameter input (x) dan bobot (w) dari masing-masing input atau dikenal dengan "*weighted sum*" menjadi *Net Value* (v) yang di tuliskan sebagai : v = ξ (x,w)

DOI: https://doi.org/10.26593/jrsi.v12i2.6595.187-200

$$v = \sum_{i=1}^{d} x_i\, w_i + w_0 \qquad \text{Pers. 14}$$

Dimana :
x = parameter input $(1, x_1, \cdots, x_d)$
w = bobot $(w_0, w_1, \cdots, w_d)$ dan w_0 : bias

2. *Activation Function* atau *Squashing Function* (ϕ), fungsi ini mengubah *Net Value* (v) menjadi nilai output (o), dimana *activation function* menggunakan metode sigmoid atau tanh yang dituliskan sebagai :

$$o = \phi(v) = \frac{1}{1+\exp(-v)} \qquad \text{Pers. 15}$$

Error atau kesalahan adalah perbedaan dari nilai output hasil jaringan saraf tiruan dengan target nilai yang diberikan berdasarkan jarak *Euclidean* atau dikenal dengan *Mean Square Error* (MSE), dituliskan (Aggarwal, 2015):

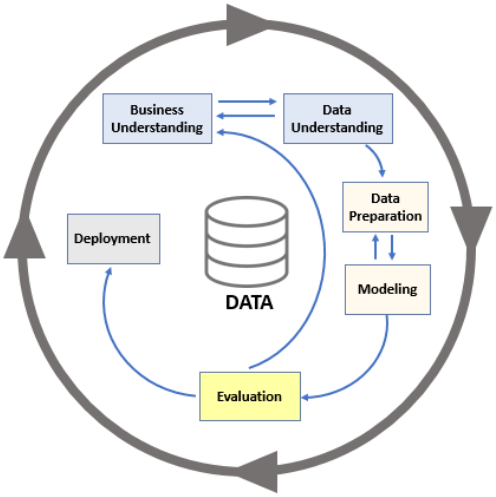
$$E_{mse}(x; w) = \sum_{k=0}^{K} (o_k - t_k)^2 \qquad \text{Pers. 16}$$

Dimana :
E_{mse} : *Error* dari *mean square error*
x : parameter *input*
w : parameter bobot
o : nilai *output*
t : target nilai *output*
k : jumlah *output* unit

Proses menghasilkan nilai output yang dimulai dari input terhadap suatu rangkaian jaringan saraf tiruan disebut dengan *Forward Propagation* atau perambatan maju. (Aggarwal, 2015). *Back Propagation* atau perambatan mundur adalah suatu metode yang bertujuan meminimalisasi *Error* dari perbedaan *output target* dengan *output* hasil *Forward Propagation* dengan cara dilakukannya penyesuaian terhadap setiap parameter bobot dan bias menggunakan *Gradient Descent*. (Witten, 2011) Secara matematik *back propagation* adalah penyesuaian bobot berdasarkan perambatan fungsi turunan *Error* terhadap masing-masing bobot dan biasnya (Aggarwal, 2015).

Metodologi Penelitian

Metodologi Penelitian Dalam melakukan penelitian, penggunaan prosedur baku dalam penggalian data yang biasa disebut dengan CRISP-DM (Cross Industry Standard Process for Data Mining). CRISP-DM terdiri dari enam tahapan untuk dipakai dalam menghasilkan model sistem prediksi yang akurat. Prosedur ini dapat dilihat pada Gambar 1.



Gambar 1. Prosedur CRISP-DM

1. Pemahaman Bisnis

Pemahaman bisnis meliputi memahami beragam jenis beasiswa dan proses administrasi sebuah PTS yang dibutuhkan jika seorang calon mahasiswa baru mengajukan program beasiswa tersebut.

2. Pemahaman Data

Pada tahap ini dilakukan pemahaman terhadap data yang akan diolah. Atribut data yang dipakai adalah nama asal sekolah SMA, provinsi atau daerah asal SMA (PD1 : Jawa Barat, PD2 : Jabodetabeka, PD3 : Jawa Tengah, PD4 : Jawa Timur dan Indonesia Timur, PD5 : Sumatera dan Kalimantan), kategori sekolah SMA (KS1 : Sekolah SMA Swasta “Katolik”, KS2 : Sekolah SMA Swasta “Kristen”, KS3 : Sekolah SMA Negeri, KS4 : Sekolah SMA Swasta lainnya) , nilai akademis mata pelajaran SMA, jurusan SMA (IPA dan IPS) yang diambil, pilihan program studi jenjang S1 di perguruan tinggi yang akan diambil, jenis pekerjaan orang tua, dan status program beasiswa. Status program beasiswa berarti ada

DOI: https://doi.org/10.26593/jrsi.v12i2.6595.187-200

3 kemungkinan dari calon mahasiswa baru setelah mengajukan beasiswa, yaitu status "Tidak Lulus" (tidak lulus seleksi penerimaan program beasiswa), bagi yang lulus seleksi maka akan terdiri dari status "Terima" (calon mahasiswa baru menerima program beasiswa), dan status "Lepas" (calon mahasiswa melepas kesempatan menerima program beasiswa).

3. Persiapan data

Pada tahap ini pengolahan data berupa:

A. Transformasi data

Transformasi setiap atribut data yang berbentuk kategorikal ke dalam bentuk biner (nilai 0 atau 1). Setiap kategori dalam sebuah atribut akan diidentifikasi dan diberi nilai 1 jika sesuai dengan kategori yang ada dan 0 jika tidak sesuai dengan kategori yang ada. Sedangkan atribut nilai rapor dilakukan normalisasi dengan cara mengubah ke dalam nilai antara 0 sampai dengan 1. Hasil transformasi data dapat dilihat pada Tabel 3.

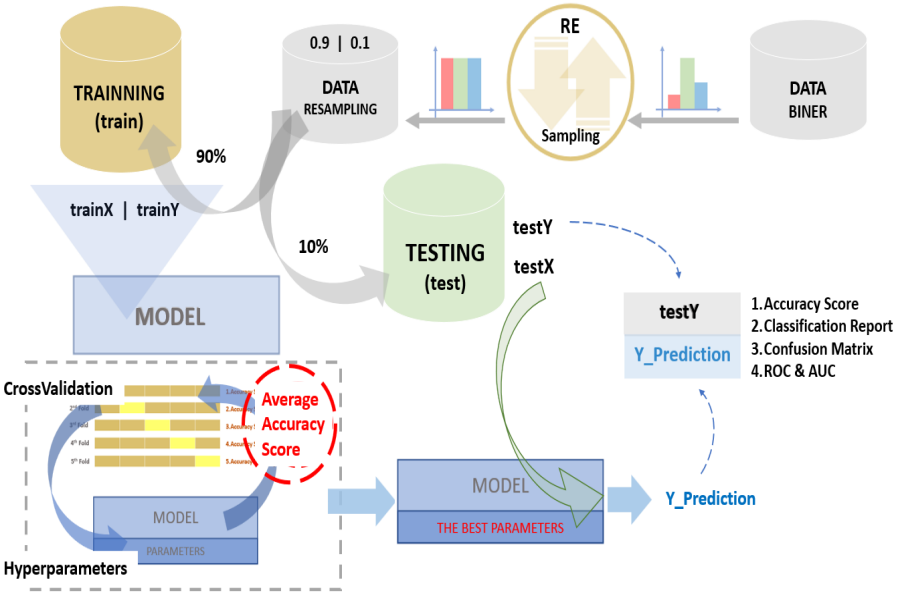
Gambaran umum tahapan setelah data biner dihasilkan dapat dilihat pada Gambar 2.

B. Resampling

Tiga kategori pengelompokan penerimaan program beasiswa ("Tidak Lulus", "Terima", dan "Lepas") memiliki jumlah data yang tidak seimbang antara 1 kategori dengan yang lainnya. Pada tahap ini jumlah data untuk ke 3 kategori tersebut hendak diseimbangkan agar sama jumlahnya. Tujuan utama dari resampling ini agar hasil pembelajaran model menjadi lebih maksimal. Metode resampling menggunakan SMOTE (*Synthetic Minority Oversampling Technique*), dimana jumlah data dari ke 3 kategori "Status" menjadi sama mengikuti kategori yang memiliki jumlah data terbanyak. Teknik SMOTE ini menambahkan data dari kelompok minoritas dengan cara melakukan interpolasi jarak dari data minoritas yang saling berdekatan dengan memanfaatkan metode *K-Nearest Neighbors* (Indrawati, 2020), dan (Chawla, 2002). Gambaran umum tahapan resampling data dapat dilihat pada Gambar 2.

Tabel 3. Hasil transformasi data

No	STATUS	PD1	PD2	KS1	KS2	JUR	PRO1	PRO2
1	Tidak Lulus	0	0	0	1	1	0	0
2	Terima	1	0	1	0	1	0	0
3	Tidak Lulus	1	0	0	1	1	0	1
4	Terima	1	0	0	1	0	0	0
5	Terima	0	1	1	0	0	1	0
...	...	...	...	...	...	...	...	...
6201	Terima	1	0	0	0	1	0	0
6202	Terima	1	0	0	1	1	0	0
6203	Terima	0	1	0	1	0	1	0
6204	Tidak Lulus	0	1	0	1	0	0	1
6205	Lepas	1	0	0	0	1	0	0



Gambar 2. Persiapan data dan pembuatan model

DOI: https://doi.org/10.26593/jrsi.v12i2.6595.187-200

C. Data *training* dan data *testing.*
Data *training* atau data latihan yang dipakai untuk proses pembelajaran model dialokasikan sebanyak 90%, sedangkan data testing atau data uji sebanyak 10% dari data hasil resampling. Gambaran umum tahapan persiapan data dapat dilihat pada Gambar 2.

4. Pembuatan Model

Model yang dibentuk akan melalui proses pencarian parameter terbaik atau dikenal dengan *hyperparameter*. Setiap kandidat parameter akan divalidasi, penggunaan teknik *cross validation* dengan menggunakan metode k-fold dengan nilai cv=5 atau k=5 untuk setiap kandidat parameter di setiap modelnya. Model yang digunakan untuk proses pembelajaran mesin adalah *Logistic Regression* (LR) , *Decision Tree* (DT), *Naïve Bayes* (NB), *K Nearest Neighbor* (KNN), *Support Vector Machine* (SVM), dan *Artificial Neural Network* (ANN). Model akan diberikan data pembelajaran dalam mencari parameter terbaik dan model terbaik dipilih dimana memiliki akurasi rata-rata model yang paling tinggi saat dilakukan *cross validation*. Ke-6 model yang sudah terbentuk akan diberikan data *testing* atau data uji untuk mengetahui kinerja masing-masing model. Evaluasi kinerja model terhadap data uji diukur melalui akurasi dan nilai AUC. Gambaran umum tahapan pembuatan model dapat dilihat pada Gambar 2.

5. Evaluasi

Pada tahap evaluasi akan dipilih model terbaik dari berdasarkan nilai akurasi rata-rata saat pembuatan model berdasarkan data latih (*data training*) dan dilanjutkan dengan membandingkan kinerja model tersebut terhadap data uji (*data testing*), yaitu dengan mengevaluasi nilai AUC. Evaluasi selanjutnya adalah mengevaluasi nilai koefisien yang paling berpengaruh dalam model prediksi untuk menentukan apakah seorang mahasiswa akan menerima program beasiswa dan tidak menolaknya.

Hasil dan Pembahasan

Dari data *sampling* sejumlah 6.206 data diambil dengan menggunakan 10 program studi Program Sarjana S1 dari salah satu perguruan tinggi swasta, maka terlihat pengelompokan terhadap atribut keputusan “Status” bersifat tidak seimbang sehingga

Setelah data atribut keputusan “Status” sudah seimbang maka dimulailah proses pembelajaran model melalu data latihan sebanyak 90% dari total data yang ada, sedangkan 10% data sisanya akan di pakai untuk pengujian model yang sudah terbentuk. Pembelajaran model menggunakan metode *Grid Search* yaitu mencari parameter yang terbaik dari setiap model dari beberapa kandidat parameter yang sudah didefinisikan sebelumnya dan setiap kandidat parameter dilakukan validasi terharap nilai akurasi rata-rata yang dihasilkan melalui proses *cross validation* (Pedregosa, 2011). Kandidat parameter yang memiliki nilai akurasi rata-rata tertinggi akan dipilih menjadi parameter yang terbaik untuk suatu model pembelajaran.

Model *Logistic Regression* menggunakan kandidat parameter nilai C=0.001 sampai dengan C=100 dengan memecah menjadi 10 kandidat nilai C (C=0.001, C=0.003, C=0.01, C=0.04, C=0.1, C=0.5, C=2, C=7, C=27, C=100). Model *Decision Tree* menggunakan kandidat parameter terdiri dari 2 jenis *criterion* yaitu *gini* dan *entropy*, sedangkan masing-masing *criterion* akan disimulasi dengan nilai *Max Depth* sebanyak 6 nilai (4, 6, 8, 10, 12, dan 14). Sehingga akan ada total 12 kandidat. Model *Naïve Bayes* menggunakan rentang nilai alpha dari alpha=0.0001 sampai dengan alpha=100 dengan memecah 10 kandidat nilai alpha (alpha =0.0001, alpha=0.0004, alpha=0.002, alpha=0.01, alpha=0.04, alpha=0.2, alpha=1, alpha=4, alpha=21, alpha=100). Model K Nearest Neighbors menggunakan jenis pembobotan *weights* yaitu *uniform*, sedangkan masing-masing nilai *n_neighbors* akan disimulasi sebanyak 5 kandidat (3, 7, 11, 17, dan 21).

Model *Support Vector Machine* (SVM) menggunakan parameter nilai C yang disimulasikan sebanyak 4 nilai (0.001, 0.1, 1, dan 10).

Model *Artificial Neural Network* (ANN) menggunakan beberapa kandidat parameter yang terdiri dari nilai *alfa* (0.5, dan 0.005) dan nilai *learning_rate_init* (0.5, dan 0,001), sedangkan *hidden_layer_sizes* terdiri dari 2 kandidat dimana setiap kandidat terdiri 3 lapisan dan setiap lapisan memiliki sejumlah unit-unit simpul (*node*/*neuron*) : (200, 150, 100) dan (90, 50, 30). Total akan ada 8 kandidat yang akan divalidasi nilai rata-rata terbaiknya.

DOI: https://doi.org/10.26593/jrsi.v12i2.6595.187-200

Bentuk topologi ANN terdiri dari : 1 *Input layers* dengan jumlah neuron sebanyak 125 sesuai dengan jumlah variabel bebas dalam model, 3 *hidden layers* dengan masing-masing jumlah neuron dari hasil parameter terbaik adalah (200,150, dan 100), dan 1 *output layers* dengan jumlah neuron sebanyak 3 sesuai dengan hasil akhir pengelompokan yang terdiri dari 3 kelas. Hasil pembentukan model yang menggunakan parameter terbaik serta akurasi rata-rata dapat dilihat pada Tabel 4.

Tabel 4. Hasil pembentukan model dan parameter terbaik

Model	Parameter	Akurasi rata-rata model
Logistic Regression (LR)	C=27	62,05%
Decision Tree (DT)	Criterion='gini' max depth=14	56,66%
Naïve Bayes (NB)	Alfa=1.0	51,46%
K Nearest Neighbor (KNN)	Weight='uniform' n_neighbors=3	60,43%
Support Vector Machine (SVM)	C=10	61,31%
Artificial Neural Network (ANN)	Alfa=0,5 learning rate=0,001 hidden layer size= (200,150,100)	60,69%

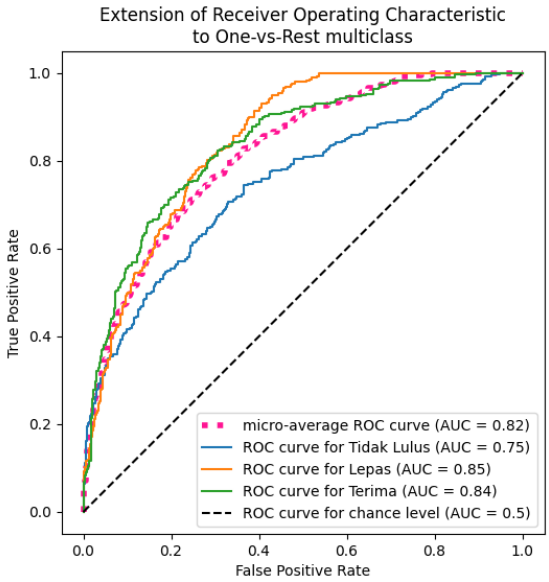
Dari Tabel 4. Terlihat bahwa model *Logistic Regression* memiliki nilai akurasi rata-rata yang paling tinggi dibandingkan dengan model pembelajaran mesin lainnya. Penerapan model ini dalam dunia nyata dengan akurasi rata-rata model (62%) artinya jika model memprediksi ada 100 calon mahasiswa baru yang akan melepaskan kesempatan program beasiswa maka sebanyak kurang lebih 62 orang secara tepat dapat diprediksi dan sisanya sebanyak 38 orang ada kemungkinan model salah dalam memprediksi. Akurasi rata-rata ini juga berlaku untuk kelompok calon mahasiswa baru yang akan menerima program beasiswa maupun mereka yang tidak lulus dalam seleksi program beasiswa. Ke-6 model yang sudah terbentuk kemudian diuji dengan data uji atau data testing yang sudah didefinisikan sebelumnya. Hasil uji tersebut di evaluasi berdasarkan nilai akurasi dan nilai AUC.

Nilai akurasi berarti hasil prediksi model terhadap data aktual dari data uji sedangkan nilai AUC (*Area Under the ROC Curve*) yaitu luasan daerah yang berada dibawah grafik ROC (*Receiver Operating Characteristic*). ROC adalah sebuah grafik yang nilai perbandingan grafik 2 dimensi yang menggambarkan *True Positive* (TP) pada sumbu y dan *False Positive* (FP) pada sumbu x. (Gorunescu, 2011). Hasil kinerja model dapat dilihat pada Tabel 5.

Tabel 5. Hasil kinerja data *testing.*

Model	Akurasi Test	Nilai AUC
Logistic Regression (LR)	62,29%	0,818
Decision Tree (DT)	59,96%	0,735
Naïve Bayes (NB)	50,06%	0,688
K Nearest Neighbor (KNN)	60,85%	0,768
Support Vector Machine (SVM)	62,18%	0,807
Artificial Neural Network (ANN)	60,73%	0,792

Model *Logistic Regression* memiliki nilai akurasi test yang lebih baik dari model lainnya demikian juga nilai AUC model *Logistic Regression* memiliki nilai yang paling tinggi dibandingkan dengan model lainnya. Rentang nilai AUC : 0 sampai dengan 1. Semakin besar nilai AUC maka model dapat secara akurat membedakan pengelompokan yang terjadi. (Gorunescu, 2011). Grafik ROC dan nilai AUC dari model *Logistic Regression* dapat dilihat pada Gambar 3.



Gambar 3. Grafik ROC dan nilai AUC

Koefisien yang terbentuk dari model *Logistic Regression* terdiri atas 3 bagian sesuai dengan 3 kategori kelompok pengambilan keputusan. Dari Tabel 6. Dapat dilihat koefisien kategori kelompok "Status: Lepas", "Status: Terima", dan "Status: Tidak Lulus".

Koefisien positif di dalam sebuah kelompok pengklasifikasian (contohnya "Status: Lepas") artinya atribut tersebut memberikan pengaruh "menarik" model untuk masuk menjadi kelompok pengklasifikasian, dalam hal ini "Status: Lepas". Besar kekuatan tarik tergantung dari besar nilai koefisien positif. Koefisien positif yang nilainya paling tinggi

DOI: https://doi.org/10.26593/jrsi.v12i2.6595.187-200

adalah atribut NMAT (Nilai Matematika) di kelompok pengambilan keputusan "Status: Lepas". Artinya semakin tinggi nilai Matematika maka kemungkinan akan masuk ke dalam kelompok "Status: Lepas" atau melepas program beasiswa semakin besar. Sedangkan koefisien negatif di dalam sebuah kelompok pengklasifikasian (contohnya "Status: Lepas") artinya atribut tersebut memberikan pengaruh paling besar bagi model untuk "mendorong" model agar tidak masuk menjadi bagian dari "Status: Lepas". Besar kekuatan dorong tergantung dari besar nilai absolut koefisien negatif tersebut. Koefisien negatif yang nilai absolut paling besar dari "Status: Lepas" adalah atribut SCH108 ("Sekolah 108"), artinya seorang calon mahasiswa baru yang berasal dari "Sekolah 108" memiliki kecenderungan probabilitas yang tinggi untuk tidak masuk ke dalam kelompok "Status: Lepas". Persamaan Model Regresi Logistik secara lengkap adalah sebagai berikut:

Persamaan 17 untuk "Status: Lepas"
Persamaan 18 untuk "Status: Terima"
Persamaan 19 untuk "Status: Tidak Lulus"

$$g_{LEPAS}(\mathbf{x}) = \beta_{0\ LEPAS} + \beta_{1\ LEPAS}(\mathbf{x}) \quad \text{Pers. 17}$$

Dimana :

$g_{LEPAS}(\mathbf{x})$: fungsi logit untuk "Status: Lepas"

$\beta_{0\ LEPAS}$: nilai intersep "Status: Lepas" : 2.130

$\beta_{1\ LEPAS}$: koefisien regresi "Status: Lepas" dari variabel bebas **x.** Lihat Tabel 6.

$$g_{TERIMA}(\mathbf{x}) = \beta_{0\ TERIMA} + \beta_{1\ TERIMA}(\mathbf{x}) \quad \text{Pers. 18}$$

Dimana :

$g_{TERIMA}(\mathbf{x})$: fungsi logit untuk "Status: Terima"

$\beta_{0\ TERIMA}$: nilai intersep "Status: Terima" : -22.732

$\beta_{1\ TERIMA}$: koefisien regresi "Status: Terima" dari variabel bebas **x.** Lihat Tabel 6.

$$g_{TIDAK\ LULUS}(\mathbf{x}) = \beta_{0\ TIDAK\ LULUS} + \beta_{1\ TIDAK\ LULUS}(\mathbf{x}) \quad \text{Pers. 19}$$

Dimana :

$g_{TIDAK\ LULUS}(\mathbf{x})$: fungsi logit untuk "Status: Tidak Lulus"

$\beta_{0\ TIDAK\ LULUS}$: nilai intersep "Status: Tidak Lulus" : 20.601

$\beta_{1\ TIDAK\ LULUS}$: koefisien regresi "Status: Tidak Lulus" dari variabel bebas **x.** Lihat Tabel 6.

Tabel 6. Koefisien regresi logistik

KOEF. STATUS	LEPAS	TERIMA	TIDAK LULUS	KOEF. STATUS	LEPAS	TERIMA	TIDAK LULUS
PD1	-1.216	2.848	-1.632	SCH31	-0.749	2.469	-1.720
PD2	-1.102	2.456	-1.354	SCH32	-0.628	3.842	-3.214
PD3	-1.223	2.582	-1.359	SCH33	0.037	1.450	-1.487
PD4	-1.047	2.306	-1.260	SCH34	-0.881	2.513	-1.632
PD5	-1.192	2.672	-1.480	SCH35	-1.269	2.832	-1.563
KS1	-1.210	2.657	-1.447	SCH36	-0.823	0.996	-0.173
KS2	-1.077	2.914	-1.837	SCH37	-1.684	3.226	-1.541
KS3	-1.008	2.708	-1.700	SCH38	-0.615	2.256	-1.641
KS4	-1.217	2.918	-1.701	SCH39	-0.822	2.053	-1.231
JUR	0.003	0.169	-0.172	SCH40	-0.770	3.333	-2.563
NBIS	1.918	1.354	-3.272	SCH41	-0.534	1.720	-1.186
NMAT	5.604	3.789	-9.393	SCH42	-1.382	3.353	-1.972
NFIS	-0.941	2.280	-1.340	SCH43	-0.644	2.613	-1.969
PRO1	-2.520	3.993	-1.473	SCH44	-1.381	3.265	-1.883
PRO2	-1.950	4.387	-2.437	SCH45	-2.002	2.551	-0.549
PRO3	-1.931	4.350	-2.419	SCH46	-0.653	3.141	-2.487
PRO4	-2.115	4.338	-2.223	SCH47	-1.180	2.834	-1.654
PRO5	-1.871	4.420	-2.549	SCH48	-0.190	2.177	-1.987
PRO6	-1.007	2.550	-1.544	SCH49	-0.696	2.949	-2.253
PRO7	-1.590	4.910	-3.320	SCH50	-0.667	2.655	-1.989
PRO8	-0.784	2.783	-1.999	SCH51	-0.463	1.416	-0.953
PRO9	-1.493	5.084	-3.591	SCH52	-0.794	3.051	-2.257
PRO10	-1.320	4.654	-3.335	SCH53	-0.939	2.620	-1.681
PA1	0.000	0.000	0.000	SCH54	-1.027	3.117	-2.091
PA2	-1.319	3.358	-2.040	SCH55	-1.125	3.270	-2.145
PA3	-1.091	3.047	-1.955	SCH56	-1.117	1.932	-0.816
PA4	-1.531	3.213	-1.682	SCH57	-0.666	2.349	-1.683
PA5	0.000	0.000	0.000	SCH58	-0.896	2.373	-1.477
PA6	0.000	0.000	0.000	SCH59	-1.032	2.155	-1.123
PA7	0.000	0.000	0.000	SCH60	-1.655	2.933	-1.277
PA8	0.000	0.000	0.000	SCH61	-1.550	2.193	-0.644
PA9	-1.205	3.118	-1.912	SCH62	-1.250	2.553	-1.303
PA10	-1.412	3.376	-1.964	SCH63	-1.918	2.815	-0.897
PA11	-1.272	3.120	-1.848	SCH64	-1.165	2.854	-1.689
PA12	0.060	-0.283	0.223	SCH65	-0.604	3.036	-2.432
PA13	0.000	0.000	0.000	SCH66	-1.789	2.912	-1.123
PA14	0.000	0.000	0.000	SCH67	-0.822	2.714	-1.891
PA15	0.000	0.000	0.000	SCH68	-0.657	3.281	-2.624
PA16	-1.569	3.434	-1.866	SCH69	-1.696	2.753	-1.057
PA17	-0.908	2.820	-1.912	SCH70	-0.899	2.898	-1.999
PA18	-1.396	3.367	-1.971	SCH71	-1.340	2.855	-1.515
PA19	0.000	0.000	0.000	SCH72	-0.966	2.344	-1.378
PA20	-1.435	3.245	-1.811	SCH73	-1.468	3.013	-1.545
PI1	0.000	0.000	0.000	SCH74	-1.162	2.219	-1.056
PI2	-1.415	3.375	-1.960	SCH75	-0.953	1.975	-1.022
PI3	-1.239	3.194	-1.955	SCH76	-1.360	2.510	-1.150
PI4	-1.284	3.175	-1.891	SCH77	-0.243	2.533	-2.289
PI5	0.000	0.000	0.000	SCH78	-0.883	2.668	-1.785
PI6	0.000	0.000	0.000	SCH79	-1.421	3.044	-1.623
PI7	0.000	0.000	0.000	SCH80	-1.370	2.833	-1.463
PI8	0.000	0.000	0.000	SCH81	-0.888	1.974	-1.086
PI9	-1.318	3.062	-1.744	SCH82	-1.375	3.076	-1.701
PI10	-1.383	3.249	-1.867	SCH83	-0.999	3.073	-2.074
PI11	-1.915	2.996	-1.081	SCH84	-1.476	2.906	-1.429
PI12	0.000	0.000	0.000	SCH85	-0.962	2.143	-1.181
PI13	0.000	0.000	0.000	SCH86	-1.088	2.350	-1.262
PI14	0.000	0.000	0.000	SCH87	-0.789	2.471	-1.683
PI15	0.000	0.000	0.000	SCH88	-0.671	1.996	-1.325
PI16	-1.469	3.197	-1.728	SCH89	-0.852	1.705	-0.853
PI17	-1.642	3.195	-1.553	SCH90	-0.404	2.202	-1.798
PI18	-1.427	3.161	-1.734	SCH91	-2.116	2.256	-0.140
PI19	0.000	0.000	0.000	SCH92	-0.847	2.534	-1.688
PI20	-1.416	3.257	-1.841	SCH93	-1.123	2.336	-1.213
SCH1	-1.011	2.416	-1.406	SCH94	-1.259	2.385	-1.127
SCH2	-1.789	2.930	-1.142	SCH95	-1.292	2.619	-1.327
SCH3	-0.867	3.449	-2.582	SCH96	-1.573	2.890	-1.317
SCH4	-1.626	3.758	-2.132	SCH97	-2.302	3.400	-1.098
SCH5	-0.770	2.687	-1.916	SCH98	-0.627	2.081	-1.454
SCH6	-1.107	2.201	-1.094	SCH99	-2.388	2.734	-0.346
SCH7	-0.985	2.948	-1.963	SCH100	-0.648	1.944	-1.296
SCH8	-0.950	2.370	-1.420	SCH101	-1.052	2.281	-1.229
SCH9	-1.675	3.863	-2.189	SCH102	-1.852	2.804	-0.952
SCH10	-0.503	1.664	-1.161	SCH103	-2.011	3.477	-1.466
SCH11	-0.961	2.861	-1.900	SCH104	-0.753	2.652	-1.899
SCH12	-0.743	2.539	-1.796	SCH105	-0.460	3.584	-3.125
SCH13	-0.946	2.666	-1.719	SCH106	0.980	-1.304	0.324
SCH14	-1.021	2.515	-1.493	SCH107	-1.711	2.767	-1.056
SCH15	-0.515	3.226	-2.711	SCH108	-3.079	5.110	-2.031
SCH16	-0.786	2.798	-2.012	SCH109	-1.130	2.741	-1.611
SCH17	-0.977	2.974	-1.997	SCH110	-1.056	1.885	-0.829
SCH18	-0.993	4.072	-3.079	SCH111	0.083	1.642	-1.725
SCH19	-0.852	2.983	-2.131	SCH112	-0.961	2.790	-1.829
SCH20	-0.849	1.151	-0.302	SCH113	-0.956	2.205	-1.249
SCH21	-0.242	2.079	-1.838	SCH114	-1.279	2.712	-1.433
SCH22	-0.636	2.428	-1.792	SCH115	-1.772	3.281	-1.510
SCH23	-0.238	2.211	-1.972	SCH116	-1.991	2.857	-0.866
SCH24	-1.333	2.683	-1.350	SCH117	-0.982	2.150	-1.167
SCH25	-1.274	2.516	-1.243	SCH118	-2.133	3.117	-0.983
SCH26	-0.586	2.279	-1.692	SCH119	-0.866	2.833	-1.967
SCH27	-1.210	1.883	-0.672	SCH120	-1.638	2.709	-1.070

DOI: https://doi.org/10.26593/jrsi.v12i2.6595.187-200

Kesimpulan dan Saran

Program pemberian beasiswa kepada calon mahasiswa baru untuk jenjang sarjana S1 di sebuah perguruan tinggi swasta dapat dimodelkan dengan pembelajaran mesin menggunakan model *Logistic Regression*. Akurasi rata-rata pembelajaran model dari data latihan / data *training* adalah 62.05% sedangkan akurasi model terhadap data uji / data testing adalah 62.29% dan kinerja model *Logistic Regression* terhadap kemampuan mengklasifikasikan kelompok pengambilan keputusan yang direpresentasikan dengan nilai AUC=0.818. Berdasarkan pengkategorian klasifikasi nilai AUC (Gorunescu, 2011) dapat dilihat di Tabel 7.

Tabel 7. Nilai AUC dan klasifikasi

AUC Score	Classification Category
0,9 – 1	Excellent Classification
0,8– 0,89	Good Classification
0,7 – 0,79	Fair Classification
0,6 – 0,69	Poor Classification
< 0,6	Failure Classification

Maka nilai AUC=0.818 terkategori "*Good Classification*" atau model *Logistic Regression* dapat dengan baik melakukan pengklasifikasian terhadap kelompok pengambilan keputusan. Untuk penelitian selanjutnya nilai akurasi rata-rata model dan nilai akurasi model terhadap data uji dapat ditingkatkan lebih lagi terutama dalam 2 hal : pertama, *missing value* atau kelengkapan data dalam setiap atribut dapat diminimasi sehingga akurasi pembelajaran model dapat meningkat. Kedua, penambahan atribut yang dipakai seperti "Penghasilan Orang Tua". (Slim at al., 2018), (Harani dan Prianto, 2020), dan (Aulck at al., 2020) secara signifikan mempengaruhi akurasi dalam pembelajaran model. Atribut lainnya seperti apakah calon penerima beasiswa pernah di interview? dan apakah calon penerima beasiswa pernah berkunjung dalam rangka program pengenalan kampus? juga dipakai oleh penelitian lainnya dalam membangun model prediksi. (Kanadpriya at al., 2019), (Alaka, 2017), dan (Aulck at al., 2020).

Daftar Pustaka

Aggarwal, C.C. (2015), *Data Classification Algorithms and Applications*, CRC Press Taylor & Francis Group, Watson Research Center Yorktown Heights, New York, USA.

Ahmed, D.M., Abdulazeez, A.M., Zeebaree, D.Q., dan Ahmed, F.Y.H., (2021), *Predicting University's Students Performance Based on Machine Learning Techniques*, IEEE International Conference on Automatic Control and Intelligent Systems (I2CACIS 2021), Malaysia, June 26.

Alaka, B.O., (2017). *A Dimensional student enrollment prediction model: case of Strathmore University*, Master Degree Thesis, Strathmore University.

Aulck, L., Nambi, D., dan West, J., (2020), *Increasing Enrollment by Optimizing Scholarship Allocations Using Machine Learning and Genetic Algorithms*, Proceedings of The 13th International Conference on Educational Data Mining (EDM 2020).

Basheer, M.Y.I., Mutalib, S., Hamid, N.H.A., Rahman, S.A., dan Malik, A., (2019). *Predictive analytics of university student intake using supervised methods*, IAES International Journal of Artificial Intelligence (IJ-AI) Vol. 8, No. 4, December 2019, pp. 367~374.

Berens, J., Schneider, K., Görtz, S., Oster, S., dan Burghoff, J., (2019), *Early Detection of Students at Risk - Predicting Student Dropouts Using Administrative Student Data from German Universities and Machine Learning Methods*, Journal of Educational Data Mining, Vol. 11, No. 3.

Cardona, T.A., dan Cudney, E.A., (2019). *Predicting Student Retention Using Support Vector Machines*, 25th International Conference on Production Research Manufacturing Innovation: Cyber Physical Manufacturing August 9-14, 2019, Chicago, Illinois (USA).

Chawla, N.V., Bowyer, K.W., Hall, L.O., dan Kegelmeyer, W.P., (2002), "SMOTE: synthetic minority over-sampling technique," Journal of artificial intelligence research, 321-357, 2002.

Delima, A.J.P. , (2019). *Predicting Scholarship Grants Using Data Mining Techniques*, International Journal of Machine Learning and Computing, Vol.9, No.4.

Fernandes, E., Holanda M., Marcio Victorinom M., Borges, V., Carvalho, R., dan Erven, G., (2018). *Educational data mining: Predictive analysis of academic*

performance of public school students in the capital of Brazil, Elsevier Journal of Business Research

Gorunescu, F. (2011). "*Data Mining : Concepts, Models and Techniques*", Springer-Verlag Berlin Heidelberg

Hamers, Y., (2017). *Predicting student enrollment Logistic regression on attended marketing events*, Master Degree Thesis, Tilburg University.

Han, J., Kamber, M., dan Pei, J., (2015), *Data Mining Concepts and Techniques.* 3rd ed. The Morgan Kaufmann series in data management systems.

Harani, N.H., dan Prianto, C., (2020). *Penerapan algoritma Adaboost guna menentukan pola masuknya calon mahasiswa. Journal Transformtika*, Vol.18, No.1, July 2020, pp. 123 – 132

Herlina, N. (2021*)."Ditjen Diktiristek Akselerasi Program Penggabungan atau Penyatuan PTS"*. (https://dikti.kemdikbud.go.id/kabar-dikti/kabar/ditjen-diktiristek-akselerasi-program-penggabungan-atau-penyatuan-pts, diakses 15 Oktober 2022).

Hosmer, D.W., Lemeshow, S., dan Sturdivant, R.X. (2013), *Applied Logistic Regression.* 3rd ed. John Wiley & Sons, Inc., Hoboken, New Jersey

Indrawati, A., Subagyo, H., Sihombing, A., dan Afandi, S. (2020), *Analyzing The Impact Of Resampling Methode For Imbalanced Data Text In Indonesian Scientific Articles Categorization*, BACA: Jurnal Dokumentasi Dan Informasi, baca.v41i2.563

Kanadpriya, B., Treena, B., Buckmire, R., dan Nishu, L., (2019), *Predictive Models of Student College Commitment Decisions Using Machine Learning*, MDPI Journal Data. 2019, 4, 65.

Kovacic, Z.J., (2010). *Predicting student success by mining enrolment data*, Proceedings of Informing Science & IT Education Conference (InSITE), 19-24 June 2010, Cassino, Italy.

Nakhkob, B., dan Khadem, M., (2015). *Predicted Increase Enrollment in Higher Education Using Neural Networks and Data Mining Techniques*, Journal of Computer Research and Development.

PDDikti Kementerian Riset, Teknologi, dan Pendidikan Tinggi Republik Indonesia, (2015 ~ 2020). "*Statistik Pendidikan Tinggi2015~2020*".(https://pddikti.kemdikbud.go.id, diakses Juli 2022)

Pedregosa, F., Varoquaux, G. dan Gramfort, A. (2011),"*Scikit-learn: Machine Learning in Python*",(https://scikitlearn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html, diakses 1 Maret 2023).

Ploutz, E.C., (2018). *Machine Learning Applications in Graduation Prediction at the University of Nevada*, Las Vegas, Master Degree Thesis, University of Nevada.

Slim, A., Hush, D., & Ojah, T., dan Babbitt, T., (2018), *Predicting Student Enrollment Based on Student and College Characteristics*, Proceedings of the 11th International Conference on Educational Data Mining, July 15-18, 2018, Buffalo, NY USA

Tomasevic, N., Gvozdenovic, N., dan Vranes, S., (2019), *An Overview And Comparison Of Supervised Data Mining Techniques For Student Exam Performance Prediction*. Elsevier Journal.

Trusheim, D., dan Rylee, C. (2011), *Predictive modeling: linking enrollment and budgeting*. Planning for Higher Education, 40(1):12, 2011.

Tukur, M.A., Abubakar, L.A., dan Sayuti, O.A., (2019), "*Marketing Mix and Students Enrolment in Private Universities in Kwara State Nigeria".* Makerere Journal of Higher Education.

Witten, I.H., Frank, E., Hall, M.A. (2011). Data Mining : *Practical Machine Learning Tools and Techniques*, 3rd ed The Morgan Kaufmann series in data management systems

Yagci, M. (2022). *Educational data mining: prediction of students' academic performance using machine learning algorithms,* Springer Open Journals.

DOI: https://doi.org/10.26593/jrsi.v12i2.6595.187-200

This page is intentionally left blank.