ISSN: 2477-5126
e-ISSN:2548-9356

Analisis K-Nearest Neighbor Berdasarkan Forward Selection Untuk Prediksi Status Mahasiswa Non Aktif Pada STMIK Bani Saleh

Panca Indah Lestari[1*)], Miftah Andriansyah [2]
[1] Magister Sistem Informasi Bisnis, Universitas Gunadarma, Jakarta
[2] Universitas Gunadarma, Jakarta
[1]Jln. Margonda Raya No.100, Depok
[2]Jln. Margonda Raya No.100, Depok
email: [1]pancaindah24@gmail.com, [2] tugasmahasiswa.miftah@gmail.com

Abstract – One of the problems faced in managing student lecture activity data (AKM) is in determining the total credits and GPA for non-active students. In managing academic data into information as an aspect of decision making in determining student activity. Several factors such as Social Studies, Number of Semester Credits, GPA, Total Number of Credits, Fees and Student Status. Steps to prevent indications of non-active students need to analyze predictive patterns to determine the remaining student study period and produce accurate information and as predictive material to compare data per academic year against K-NN non-active students based on Forward Selection. Non-active student prediction research uses Rapid Miner testing on 342 student datasets, the quality of the accuracy value of K-Nearest Neighbor (k-3) is 93.55% and Forward Selection (k-3) is 99.39%. from the results of data analysis of students who will Drop Out of 1160 as a proposal for management in the next reporting period. then the research can be developed further in order to build an optimal k value by adding aspects of the classification of student status working or not working.
Keywords- Non-Active, *K-Nearest Neighbor* Algorithm, Forward Selection, and Rapid Miner.

Abstrak – Masalah yang dihadapi dalam pengelolaan data aktivitas kuliah mahasiswa (AKM) salah satunya dalam menentukan total SKS dan IPK pada mahasiswa non aktif. Dalam melakukan pengelolaan data akademik menjadi informasi sebagai aspek pengambilan keputusan dalam menentukan keaktifan mahasiswa. Beberapa faktor seperti IPS, Jumlah SKS Semester, IPK, Jumlah SKS Total, Biaya dan Status Mahasiswa. Langkah untuk mencegah indikasi mahasiswa non aktif perlu dilakukan analisis pola prediksi untuk menentukan sisa masa studi mahasiswa serta menghasilkan informasi yang akurat dan sebagai bahan prediksi untuk membandingkan data per tahun akademik terhadap mahasiswa non aktif K-NN berbasis Forward Selection. Penelitian prediksi mahasiswa non aktif menggunakan pengujian menggunakan Rapid Miner terhadap dataset mahasiswa sebanyak 342, menghasilkan nilai akurasi K-Nearest Neighbor (k-3) sebesar 93,55% dan Forward Selection (k-3) sebesar 99,39%. dari hasil analisis didapatkan data mahasiswa yang akan Drop Out sebesar 1160 sebagai usulan untuk manajemen pada periode pelaporan berikutnya. maka penelitian dapat dikembangkan lebih lanjut guna penentuan nilai k yang optimal dengan menambahkan aspek klasifikasi status mahasiswa bekerja atau tidak bekerja.
Kata Kunci – Non Aktif, Algoritma K-Nearest Neighbor, Forward Selection, dan Rapid Miner.

I. PENDAHULUAN

Pada era saat ini sistem data berbasis Teknologi ialah perihal yang sangat berarti buat keberlangsungan hidup organisasi ataupun bisnis. Sistem sebagai hal yang sangat dibutuhkan akibat terdapatnya kompleksitas yang besar dalam tiap organisasi bisnis. Tanpa data yang pas, tidak terdapat organisasi ataupun bisnis manapun baik dari segi aspek pemerintahan, pendidikan, maupun dunia usaha yang bisa mengambil langkah dengan benar dalam proses pengambilan keputusan. perkembangan sistem yang memberikan sebuah pemahaman keleluasaan dalam mengelola dan mencari informasi yang merupakan hal yang dinanti untuk dikelola sehingga menjadi bahan atau barometer dalam pengambilan keputusan.[1]

*) **penulis korespondensi**: Panca Indah Lestari
Email: pancaindah24@gmail.com

STMIK Bani Saleh adalah satu dari banyak perguruan tinggi swasta yang melakukan pengelolaan data akademik menjadi informasi sebagai aspek pengambilan keputusan. Serta berusaha mengadakan layanan informasi lebih baik, efektif dan akurat baik di dalam maupun diluar perguruan tinggi. Pada perjalanannya di temukan hambatan, dalam pengelolaan data AKM seperti menentukan total SKS dan IPK pada mahasiswa non aktif. Mahasiswa merupakan salah satu aspek penting dalam evaluasi keberhasilan penyelenggaraan program studi pada suatu perguruan tinggi. Penilaian keberhasilan studi mahasiswa tersebut dapat digunakan sebagai masukan penting dalam pengambilan keputusan.

STMIK Bani Saleh mengolah data aktivitas belajar mengajar seperti data mahasiswa, dosen, kelas, KRS, nilai dan AKM. Sebagai bahan pelaporan pada PDDIKTI mengenai aktivitas kuliah mahasiswa terutama mahasiswa non aktif, status aktivitas kuliah mahasiswa harus dilaporkan per semester di halaman http://pddiktiadmin.kemdikbud.go.id/admin/kemahasiswaan/data-mahasiswa. PMPK Republik Indonesia Nomor 3 Tahun 2020 Pasal 17 nomor 1 tentang masa serta beban belajar

penyelenggaraan program pendidikan adalah batas masa studi, program diploma tiga 5 tahun dengan beban belajar mahasiswa minimum 72 SKS dan program sarjana 7 tahun dengan beban belajar mahasiswa minimum 108 SKS.3.[2]

Sesuai panduan akademik mahasiswa Non-Aktif adalah mahasiswa tidak terdaftar sebagai mahasiswa aktif atau tidak melakukan pengisian Form Registrasi Mahasiswa (FRS) pada beberapa semester tanpa seizin program studi. Mahasiswa non aktif tidak bisa menggunakan hak-haknya sebagai mahasiswa sebagaimana mahasiswa aktif. Sehingga timbul prediksi faktor penyebab terjadinya status tidak jelas, diantaranya Dikeluarkan, Mutasi, Mengundurkan Diri, Putus Kuliah, Wafat, Lulus dan Dropout (DO).[3]

Manajemen perlu mengidentifikasi dan melakukan antisipasi terhadap mahasiswa berstatus "tidak diharapkan", dalam mengetahui faktor munculnya permasalahan tersebut dilakukan proses identifikasi masalah yaitu analisis pola prediksi berdasarkan data dari Pusat Komputer (PUSKOM), mengenai faktor–factor seperti menentukan total SKS dan IPK yang mempengaruhi munculnya mahasiswa berstatus non aktif dalam kualitas

II. METODE PENELITIAN

A. Pengumpulan Data

[4]Teknik pengumpulan data pada riset ini melalui klasifikasi data primer, data yang diperoleh dari dengan metode survei dan wawancara untuk pengambilan data secara langsung ke lapangan untuk memperoleh data populasi ataupun sampel data yang dihasilkan adalah sebuah data mahasiswa selama 3 tahun dengan tahun ajaran 2017-2018, tahun 2018-2019, serta tahun 2019-2020 ± 18.611 kemudian ditentukan sampel sebesar 342 berdasarkan tabel Isaac Michael, serta dalam pengumpulan data menggunakan alat ukur (instrumen) penelitian melalui Algoritma K- Nearest Neighbor dan Forward Selection serta dibantu dengan aplikasi Rapid Miner untuk mendapatkan akurasi yang tepat.

B. Metode Pemodelan

Metode-metode yang digunakan dalam penelitian diantaranya:

1. Algoritma *K-Nearest Neighbor*

[5]Metode klasifikasi yang mempunyai cara kerja sederhana dalam algoritma ini dibandingkan dari algoritma klasifikasi lainnya,namun mampu menghasilkan prediksi akurasi yang tidak burukdengan menghitung kedekatan antara kasus baru dan lama dengan berdasarkan pada kecocokan bobot dari sejumlah fitur yang ada. Data set mahasiswa dengan klasifikasi dalam menganalisis status mahasiswa non aktif memiliki 6 faktor utama yaitu : Indek Prestasi Semester (IPS), Jumlah SKS Semester, Indek Prestasi Kumulatif (IPK), Jumlah SKS Total, Biaya dan Status Mahasiswa. Dengan Klasifikasi voting yang paling banyak antara klasifikasi objek yang sudah ditentukan. Menentukan pencarian jarak terdekat digunakan rumus Euclidean Distance.[6]Untuk perhitungan rumus dari K-Nearest Neighbor dapat dijelaskan pada persamaan sebagai berikut:

$$d_i = \sum_{i=1}^{p} \sqrt{(x_{2i} + x_{1i})^2}$$

Dimana: p = Dimensi; i = Atribut; x_1= Data sampel; x_2=Data uji; d = Jarak.

2. Forward Selection

Pemodelan yang diawali empty model, berikutnya satu per satu perubah dimasukkan agar kriteria tertentu terpenuhi. Penggunaan Forward Selection yaitu untuk memaksimalkan hasil akurasi fitur guna data tidak relevan harus dihapus agar tidak mengurangi nilai akurasi. [7]

• Dalam memilih variable dependen yang paling berkorelasi dengan dan jika signifikan dimasukkan ke dalam model. Dengan menentukan predictor variable tidak dalam model, pilih satu satu variable dengan nilai p- value terkecil dan kurang dari taraf nyata.

Koefisiensi regresi a ditentukan dengan menggunakan rumus dibawah ini:

$$a = \frac{(\Sigma Y_1)(\Sigma X_{1^2}) - (\Sigma X_1)(\Sigma X_1 Y_1)}{n\,\Sigma X_{1^2} - (\Sigma X_{1^2})}$$

Koefisien b ditentukan dengan menggunakan rumus :

$$b = \frac{n(\Sigma X_1 Y_1) - (\Sigma X_1 Y_1)}{n\,\Sigma X_{1^2} - (\Sigma X_1)^2}$$

• Untuk mengukur kekuatan hubungan antar variable predictor X dan response Y, dimana dilakukan analisis korelasi yang hasilnya dinyatakan oleh suatu bilangan yang dikenal dengan koefisien korelasi. Biasanya analisis regresi sering dilakukan bersama-sama dengan analisis korelasi (r), sebagai berikut:

$$r = \frac{n\sum_{i=1}^{n} X_i Y_i - (\sum_{i=1}^{n} X_i)(\sum_{i=1}^{n} Y_i)}{\sqrt{[n\sum_{i=1}^{n} X_{i^2} - (\sum_{i=1}^{n} X_i)^2][n\sum_{i=1}^{n} Y_{i^2} - (\sum_{i=1}^{n} Y_i)^2}}$$

• Meregresikan variabel respons Y, dengan *predictor* Xa, dijumlah dengan setiap predictor selain dari Xa serta *predictor* yang lain. Setelah itu tentukan model dengan nilai R^2 yang paling tinggi, misal memiliki tambahan *predictor* Xb, misal model pada persamaan:

$$Y = b_0 + b_a x_a + b_x x_b$$

• *Predictor* terpilih Xb artinya memiliki Fsequensial yang paling tinggi. Formula Fsequensial untuk Xb adalah:

$$F_{seq} = R(\beta_0\beta_0\beta_0)MSE\ db$$

Dimana pada persamaan nilai Fsequensial untuk Xb juga bisa didapat dengan cara mengkuadratkan statistik uji T *predictor* Xb.

$$T = \frac{\sum d_1}{\sqrt{\frac{N\sum d_{1^2} + \sum d_{1^2}}{N-1}}}$$

ISSN: 2477-5126
e-ISSN:2548-9356

- Proses tersebut diulang hingga didapatkan $F_{sequensial}$l>F_{in}
  $F_{in} = F(1, v, a_{in})$jadi mode terbaik yang didapatkan adalah model yang tidak memiliki predictor dengan Fsequensial<F_{in}

$$F = \frac{S^2 terbesar}{S^2 terkecil}$$

kemudian dilanjutkan dengan menghitung $F_{sequensial}$:

$$F_{sequensial} = (a; \frac{dk\ (B)}{dk\ (A)})$$

3. Rapid Miner
   Aplikasi Rapid Miner sangat mempermudah mendapatkan nilai akurasi yang tinggi berdasarkan algoritma yang dihitung selain itu, dalam hampir semua kasus Rapid miner merupakan alat bantu yang sangat mengurangi gangguan dalam data masukan dengan menghilangkan fitur yang tidak diperlukan. Model yang kurang rentan terhadap over fitting dan menjadi lebih sederhana yang membuatnya lebih tahan banting perubahan data kecil. Selain itu, kesederhanaan banyak meningkatkan pemahaman model. Algoritma evolusioner adalah teknik yang ampuh untuk pemilihan fitur. dimana tidak terjebak dalam optimal lokal pertama seperti pemilihan maju atau eliminasi mundur. Ini mengarah pada model yang lebih akurat. Alat algoritma ini mengirimkan solusi Pareto penuh dalam waktu proses yang sama. Dengan menggunakan antara kompleksitas dan keakuratan data apakah benar-benar layak untuk menambahkan tambahan ini fitur Pendekatan pemilihan fitur multi-tujuan tanpa pengawasan memiliki keunggulan lain dalam teknik pengelompokan.[8]

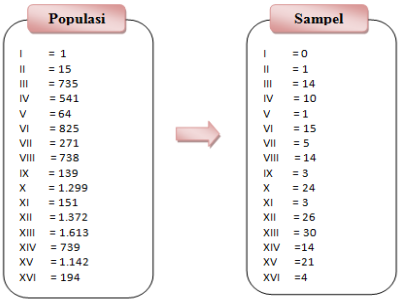
III. HASIL DAN PEMBAHASAN

Dalam penelitian ini menggunakan dataset mahasiswa dengan populasi selama 3 tahun terakhir terdiri dari tahun 2017, 2018, 2019 dengan semester ganjil dan genap dari seluruh jurusan dengan jumlah ± 18.611. Dalam menganalisis status mahasiswa non aktif dengan 6 faktor seperti yang tertera pada tabel berikut :

Tabel 2
Perhitungan Dengan Melibatkan Sampel Data Set

No	Data ke-	IPS	Jumlah sks Semester	IPK	Jumlah sks Total	Biaya Kuliah	Status Mahasiswa	Jarak
1	1	-	-	-	-	-	NON-AKTIF	10,961
2	2	3.41	22.00	3.52	46	-	AKTIF	8,756
3	3	1.47	19.00	1.89	132	4,515,000	AKTIF	2.124,853
4	4	-	-	1.86	133	-	NON-AKTIF	3,73
5	5	-	-	0.51	110	-	NON-AKTIF	3,391
6	6	-	-	1.47	38	-	NON-AKTIF	9,064
7	7	-	-	-	-	-	NON-AKTIF	10,961
8	8	3.38	8.00	3.12	114	-	AKTIF	2,925
9	9	-	20.00	-	20	6,050,000	AKTIF	2.459,675
10	10	-	-	-	15	-	NON-AKTIF	10,255
11	11 (Objek Baru)	4.00	2	4.00	120.00	-	?	

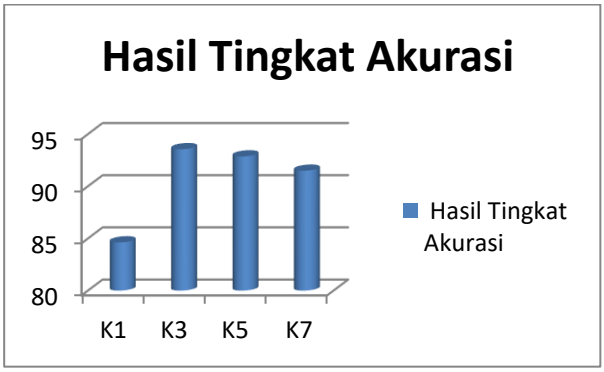
hasil akurasi yang diperoleh dari perhitungan menggunakan algoritma *K-Nearest Neighbor* (k-NN) dengan *Forward Selection* dengan sampel



Gambar 1 Hasil Perhitungan Sampel dengan Tingkat Kesalahannya 5%

Maka dapat disimpulkan penentuan hasil sampel adalah jumlah sampel awal sebanyak 18.211 setelah dihitung dengan teknik DSRS dengan tingkat kesalahan 5% menggunakan table yang dikembangkan *Isaac* dan *Michael*[9] jumlah sampel dari populasi 18.211 adalah 20.000 sampel dengan tingkat 5% adalah 342 sampel populasi yang proporsional dengan 17 cluster yang digunakan.

1. Algoritma *K-Nearest Neighbor*
   Data berdasarkan nilai K, dan untuk menentukan label yang frekuensinya paling sering diantara k training records yang paling dekat dengan objek berdasarkan dataset mahasiswa, Untuk mengetahui tingkat akurasi yang didapatkan pada proses klasifikasi menggunakan algoritma K-Nearest Neighbor tanpa seleksi fitur Forward Selection tersebut maka akan digunakan nilai k=1, k=3, k=5, k=7. berikut merupakan hasil nilai akurasi yang didapatkan.



Grafik 1 Hasil Tingkat Akurasi berdasarkan K

ISSN: 2477-5126
e-ISSN:2548-9356

ISSN: 2477-5126
e-ISSN:2548-9356

Table 3 Hasil Akurasi Tiap Nilai K

Tingkat K	Hasil Tingkat Akurasi
K1	84,62
K3	93,55
K5	92,86
K7	91,49

Berdasarkan table 3 diatas nilai k yang mempunyai tingkat akurasi yang paling tinggi adalah k=3 dengan nilai sebesar 93,55% dengan dilanjutkan k=5 dengan nilai 92,86%, k=7 sebesar 91,49%, dan k=1 sebesar 84,62%.

Gambar 2 Hasil Nilai Akurasi Algoritma KNN Untuk K=3

Berdasarkan hasil perhitungan dengan algoritma k-NN menghasilkan maka diperoleh hasil akurasi dengan tetangga terdekat di K3 dengan tingkat ke akurasian 93,55 % dari 4 percobaan dengan tingkat k1, k3, k5, dan K7.

2. *Forward Selection*
Selanjutnya dilanjutkan dengan proses perhitungan Forward Selection dalam metode ini dari variabel-variabel yang telah dilakukan seleksi terlebih dahulu untuk dapat menentukan variabel dependent dan independent. [10]Untuk dapat menghitung persamaan regresi linier:

$$Y = a + bX$$

$$Y = 57{,}64 + 298{,}35\ x\ 817{,}19 = 243{,}868.29$$

mengukur kekuatan hubungan variable predictor X dan response Y, dilakukan analisis korelasi yang hasilnya dinyatakan dengan koefisien korelasi. Persamaan koefisien korelasi (r):

$$r = \frac{1.152.308{,}96}{2{,}20547} = 522{,}479$$

selanjutnya dengan *Stepwise Regression* (regresi bertatar)[11]. Dari matriks diatas variable x dengan nilai tinggi korelasi nya dengan variable respon Y, r= 5,22. Dengan demikian variable x yang dimasukan ke dalam persamaan regresi.

	Df	*SS*	*MS*	*F*	*Significance F*
Regression	1	44395.74616	44395.74616	22.64441451	2.88706E-06
Residual	340	666590.593	1960.560568		
Total	341	710986.3392			

Nilai F-parsial terkecil dan tidak nyata, sehingga harus dikeluarkan dari model. dengan rumus diatas sebagai berikut:

$$F = (22{,}6 : (342 - 817{,}2) = 0{,}07$$

Dari nilai F-parsial, baik pada taraf nyata 5% maupun taraf nyata 10%, ternyata variable X yang terkecil dan tidak signifikan, sehingga harus dikeluarkan dari persamaan. Maka variable yang dipilih adalah X dan demikian model persamaan baru $Y = f(X, X2)$

$$Y = 0{,}07 + 817.2 + 243{,}9$$

$$R^2 = \frac{\sum x_1 x_2}{(\sum x_1^2)(\sum x_1^2)} = \frac{715.108.887{,}72}{1.172.309.652} = 0{,}61$$

Berdasarkan hasil yang sudah didapat dari perhitungan K-Nearest Neighbor akan dilakukan perbandingan dengan klasifikasi nilai k=1, k=3, k=5, dan k=7.sehingga

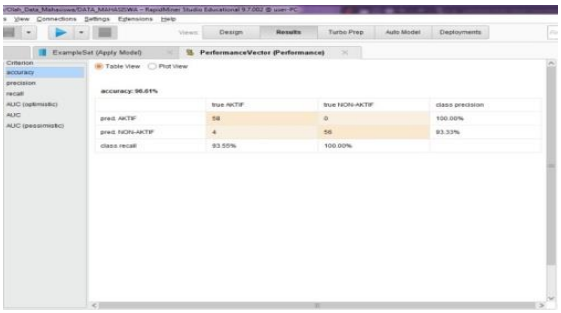
RUMUS ^0.5

	NON-AKTIF	AKTIF	Class Precision (%)
NON-AKTIF	58.00	-	100.00
AKTIF	4.00	58.00	6.45
Class Recall (%)	93.55	-	

$$Akurasi = \frac{58}{58 + 62} x100 = 93{,}55\ \%$$

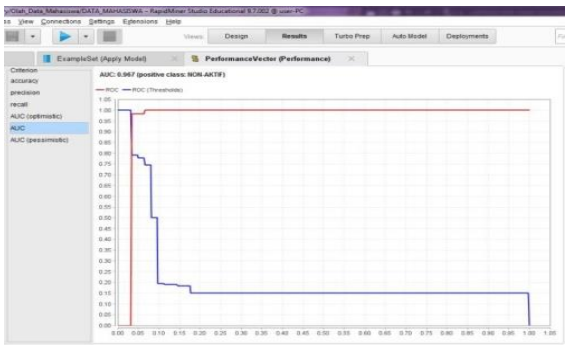
memperoleh hasil akurasi sebesar 99,39% dengan diawali perhitungan regresi linier yang memperoleh nilai a = 57.64, b = 298,35 maka dihasilkan nilai Y sebesar 243,868.29 kemudian dilanjutkan dengan menghitung nilai r = 522,479 maka menghasilkan nilai F = 22,6 sehingga dari proses regresi tersebut menghasilkan nilai R2 dengan tingkat ke akurasian sebesar 99,39% .

3. Perhitungan *Rapid Miner*
   a. Hasil akurasi yang diperoleh dalam pengujian dengan menggunakan algoritma K-Neighbor sebesar 93,55%.



Gambar. 3 Hasil Akurasi dengan pengujian rapidminer

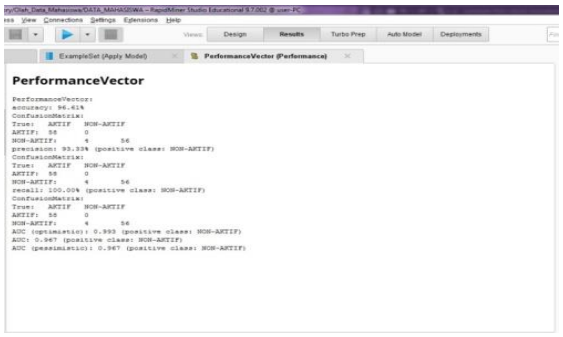
b. Hasil Dalam penelitian ini menggunakan [12] alat ukur untuk mengevaluasi model yang digunakan AUC untuk mengevaluasi model yang diusulkan. Berdasarkan hasil penelitian dapat disimpulkan sebagai Keakuratan kinerja untuk memprediksi status mahasiswa dievaluasi oleh pengukuran evaluasi yang akan digunakan untuk mengukur akurasi model yang diusulkan seperti gambar dibawah ini:



Gambar 4 AUC K- Nearest Neighbor

Pada gambar diatas menjelaskan bahwa nilai yang dihasilkan dalam analisis menggunakan rapid miner 9.7 berdasarkan data yang sudah diolah sebesar AUC: 0.967% (positive class: NON- AKTIF)

c. Dari hasil akurasi AUC maka dapat disimpulkan nilai akurasi performance actor sebagai berikut:



Gambar 5 Performance actor K-NN

IV. KESIMPULAN

Hasil penelitian disimpulkan sebagai berikut:

a. Dalam menganalisis status mahasiswa non aktif memiliki 6 faktor utama yaitu : IPS, Jumlah SKS Semester, IPK, Jumlah SKS Total, Biaya dan Status Mahasiswa.
b. Dalam pengujian menggunakan menggunakan Rapid Miner terhadap dataset mahasiswa sebanyak 342, menghasilkan nilai akurasi *K-Nearest Neighbor* (k-3) sebesar 93,55% dan Forward Selection (k-3) sebesar 99,39%.
c. Hasil analisis tetangga terdekat (K3) berdasarkan dataset mahasiswa memiliki akurasi sebesar +0.97 % pada tabel *Kurva Area Under Curve* (AUC), didapatkan data mahasiswa yang akan Drop Out sebesar 1160 sebagai usulan untuk manajemen pada periode pelaporan berikutnya.

UCAPAN TERIMA KASIH

Ucapan terima kasih penulis kepada pihak yang membantu ataupun memberikan dukungan terkait dengan penelitian yang dilakukan seperti bantuan fasilitas penelitian, dan lainnya.

DAFTAR PUSTAKA

[1] E. K. Natakusumah, “Perkembangan Teknologi Informasi di Indonesia,” *Pus. Penelit. Inform. Bandung*, 2002.
[2] M. Pendidikan, D. A. N. Kebudayaan, and R. Indonesia, “jdih.kemdikbud.go.id,” 2020.
[3] SPMI STMIK Bani Saleh, *Panduan Akademik STMIK Bani Saleh*. 2016.
[4] Sugiyono, *Metodologi Penelitian Pendidikan (Pendidikan Kualitatif, Kuantitatif)*. Bandung: Alfabeta, 2010.
[5] E. T. Kusrini, *Algoritma Data Mining*. Yogyakarta: ANDI Yogyakarta, 2009.
[6] Daniel T Larose, *Data Mining Methods Models*. Canada: by John Wiley, 2006.
[7] I. Conference, “November 3 - 4, 2020,” no. Icic, 2020.
[8] R. Sanjaya and F. Fitriyani, “Prediksi Bedah Toraks Menggunakan Seleksi Fitur Forward Selection dan K-Nearest Neighbor,” *J. Edukasi dan Penelit. Inform.*, vol. 5, no. 3, p. 316, 2019, doi: 10.26418/jp.v5i3.35324.
[9] Sugiyono, “Metodologi Penelitian pendidikan pendekatan kuantitatifa, kualitatif dan R&D,” *Univ. Pendidik. Indones.*, vol. 1, no. Metodologi Penelitian, pp. 1–58, 2010.
[10] Yuliara and I Made, “Regresi Linier Sederhana,” *Fisika*, pp. 7–41, 2016.
[11] R. B. Bendel and A. A. Afifi, “Comparison of stopping rules in forward ‘stepwise’ regression,” *J. Am. Stat. Assoc.*, vol. 72, no. 357, pp. 46–53, 1977.
[12] L. Porte *et al.*, “Evaluation of a novel antigen-based rapid detection test for the diagnosis of SARS-CoV-2 in respiratory samples,” *Int. J. Infect. Dis.*, vol. 99, pp. 328–333, 2020, doi: 10.1016/j.ijid.2020.05.098.