Perbandingan Akurasi Model Pembelajaran Mesin untuk Prediksi Seleksi Masuk Perguruan Tinggi Negeri

http://dx.doi.org/10.28932/jutisi.v9i1.6126

Riwayat Artikel
Received: 26 Januari 2023 | *Final Revision*: 31 Januari 2023 | *Accepted*: 3 Februari 2023

Oktavianus Yopi Wardana ✉ #1, Mewati Ayub#2, Andreas Widjaja #3

Magister Ilmu Komputer, Universitas Kristen Maranatha
Jl. Surya Sumantri No.65, Bandung, 40164, Indonesia
[1]2179002@maranatha.ac.id
[2]mewati.ayub@it.maranatha.edu
[3]andreas.widjaja@it.maranatha.edu
✉Corresponding author: okta.yopi@gmail.com

Abstrak* — Seleksi Nasional Masuk Perguruan Tinggi Negeri (SNMPTN) masih menjadi salah satu jalur favorit bagi siswa SMA untuk melanjutkan pendidikan di Perguruan Tinggi Negeri (PTN). SNMPTN menggunakan nilai rapor semester 1 sampai dengan 5 untuk 6 mata pelajaran yang diinput di Pangkalan Data Sekolah dan Siswa (PDSS). Prediksi SNMPTN dapat dilakukan menggunakan model pembelajaran mesin dengan berbagai macam metode. Penelitian ini bertujuan untuk membuat model prediksi dengan menggunakan metode *Decision Tree* CART, *Gaussian Naïve Bayes* dan *Logistic Regression*, melakukan prediksi dan membandingkan tingkat akurasi dari model yang dibuat. Metodologi yang digunakan dalam penelitian ini adalah *Knowledge Discovery in Database* (KDD). Ini untuk mendapatkan pengetahuan yang berguna dari data. Dataset yang digunakan merupakan data nilai 6 mata pelajaran selama 5 semester dari angkatan 2015 hingga 2022. Evaluasi model menggunakan *Split Percentage Method* dan *K-Fold Cross Validation*. Hasil menunjukkan, skor akurasi untuk 3 model yang dibuat berbeda. *Logistic Regression* memiliki skor 0,82, diikuti oleh *Decision Tree* CART dengan skor 0,75 dan terakhir adalah *Gaussian Naïve Bayes* dengan skor 0,70. Hipotesis yang diajukan oleh peneliti sesuai dengan hasil yang didapatkan, bahwa model *Logistic Regression* memiliki skor akurasi lebih tinggi. Secara matematis, *Logistic Regression* tidak terlalu rumit jika dibandingkan dengan model lainnya. Untuk mendapatkan model yang sesuai dengan kebutuhan harus melibatkan iterasi melalui proses machine learning dan mencoba berbagai macam variasi.

***Kata kunci*— Akurasi; *Knowledge Discovery in Database*; Model; Prediksi; Seleksi Masuk.**

Accuracy's Comparison of Machine Learning Models for Predicting State College Admission Selection

Abstract — ***Seleksi Nasional Masuk Perguruan Tinggi Negeri (SNMPTN) is still one of the favorite admission routes for high school students to continue their education at Perguruan Tinggi Negeri (PTN). SNMPTN uses semester 1 to 5 report card scores for 6 subjects that are inputted in Pangkalan Data Sekolah dan Siswa (PDSS). Prediction of SNMPTN can be done using machine learning models with various methods. This study aims to create a predictive model using the Decision Tree CART, Gaussian Naïve Bayes and Logistic Regression methods, make predictions and compare the level of accuracy of the models made. The methodology used in this research is Knowledge Discovery in Database (KDD). This is to get useful knowledge from data. The dataset used is data on the scores of 6 subjects for 5 semesters from class 2015 to 2022. Model evaluation uses the Split Percentage Method and K-Fold Cross Validation. The results show that the accuracy scores for the 3 models are different. Logistic Regression has a score of 0.82, followed by Decision Tree CART with a score of 0.75 and finally Gaussian Naïve Bayes with a score of 0.70. The hypothesis put forward by the researcher is in accordance with the results obtained, that the Logistic Regression model has a higher accuracy score. Mathematically, Logistic Regression is not too complicated when compared to other models. To get a model that fits with needs must involve iterating through the machine learning process and trying various variations.***

Keywords— ***Accuracy; Knowledge Discovery in Database; Model; Prediction; Selection.***

I. PENDAHULUAN

Sebagian besar siswa lulusan SMA Santa Maria 1 Cirebon memutuskan untuk melanjutkan pendidikan ke perguruan tinggi. Perguruan Tinggi Swasta (PTS) menjadi tujuan sebagian besar siswa. Walau begitu, trend siswa yang berminat untuk melanjutkan ke Perguruan Tinggi Negeri (PTN) melalui jalur SNMPTN cukup tinggi. SNMPTN (Seleksi Nasional Masuk Perguruan Tinggi Negeri) menggunakan data dari nilai rapor siswa dari semester 1 sampai dengan 5 yang di input dalam PDSS (Pangkalan Data Sekolah dan Siswa). Sekolah dengan akreditasi A memiliki kesempatan untuk mendaftarkan 40% siswa terbaiknya di SNMPTN [1].

Ada banyak alasan spesifik yang melatarbelakangi siswa tersebut memilih PTN, diantaranya: faktor biaya kuliah yang murah, pilihan jurusan dan PTN favorit, prospek pasca lulus yang baik. Daya tampung total dari seluruh jurusan PTN di Indonesia tahun 2022 yaitu 122.651 kursi. Sementara itu pendaftar SNMPTN untuk tahun 2022 yaitu 612.049 siswa [2]. Dari 612.049 siswa pendaftar, hanya 20% yang dapat lolos SNMPTN. Ada beberapa hal yang dapat menyebabkan tidak lolos SNMPTN, diantaranya, atribut nilai dan yang tidak sesuai, kuota jurusan dan PTN yang terbatas serta linieritas jurusan SMA dengan jurusan di PTN. Target pencapaian siswa yang lolos SNMPTN di SMA Santa Maria ingin ditingkatkan. Untuk mencapai target tersebut butuh strategi-strategi khusus dalam pendampingan SNMPTN tahun selanjutnya. Salah satunya adalah perlu dibuatnya model prediksi SNMPTN.

Machine learning dapat memberikan solusi atas masalah klasifikasi [3]. Pada penelitian ini fokus utama klasifikasi adalah tentang siswa yang 'lolos' dan 'tidak lolos' SNMPTN. Kategori yang diprediksi merupakan binomial atau *binary classification*. Prediksi dibuat berdasarkan atribut-atribut yang terdapat di dalam dataset SNMPTN. Semua skenario yang muncul dari kombinasi ini menghasilkan sejumlah model dengan sejumlah asumsi, keuntungan, dan kerugian yang sesuai. Model klasifikasi yang dapat digunakan dalam prediksi diantaranya *Decision Tree, Naïve Bayes* dan *Logistic Regression*.

Salah satu fungsi utama dari *machine learning* adalah untuk melakukan generalisasi. Adanya *overfitting* dan *underfitting* menyebabkan *machine learning* tidak dapat menjalankan fungsi utamanya dalam hal generalisasi. *Overfitting* merupakan keadaan dimana model berusaha mempelajari seluruh detail yang ada di dalam data dan berusaha memasukkan semua data point ke dalam garis. Sedangkan *underfitting* merupakan keadaan dimana model *machine learning* tidak bisa mempelajari hubungan antara variabel dalam data serta memprediksi atau mengklasifikasi data point baru [3].

Untuk menghindari *overfitting* maupun *underfitting*, pemilihan model yang akan digunakan harus sesuai dengan karakteristik dataset yang digunakan. Jika model terlalu kompleks sedangkan data yang sederhana, maka dapat terjadi *overfitting*. Begitu pun sebaliknya, jika model terlalu sederhana sementara data yang digunakan terlalu kompleks, maka akan terjadi *underfitting*. Berdasarkan karakteristik data yang ada dan tujuan prediksi dalam penelitian ini, maka model *classification* seperti *Decision Tree* CART, *Gaussian Naïve Bayes* dan *Logistic Regression* merupakan model yang cocok diterapkan. Sehingga model prediksi yang dibuat dalam penelitian ini menggunakan metode *Decision Tree* CART, *Gaussian Naive Bayes* dan *Logistic Regression*.

Dalam penelitian Song dan Lu [4] *Decision Tree* merupakan metode penambangan data yang umum digunakan untuk membangun sistem klasifikasi berdasarkan beberapa kovariat atau untuk mengembangkan algoritma prediksi untuk variabel target. Metode ini mengklasifikasikan suatu populasi ke dalam segmen-segmen mirip cabang yang membentuk pohon terbalik dengan simpul akar, simpul dalam, dan simpul daun. Algoritma ini non-parametrik dan dapat secara efisien menangani kumpulan data yang besar dan rumit tanpa menerapkan struktur parametrik yang rumit.

Penelitian Pandey dan Sharma [5] menjelaskan untuk pembangunan model klasifikasi yang dapat memprediksi kinerja siswa, khususnya untuk cabang teknik, sebuah algoritma *Decision Tree* yang terkait dengan teknik data mining telah digunakan dalam penelitian. Sejumlah faktor dapat mempengaruhi kinerja siswa. Di sini beberapa faktor signifikan telah dipertimbangkan saat menyusun *Decision Tree* untuk mengklasifikasikan siswa menurut atribut (nilai) mereka.

Paper yang dibuat oleh Hamoud [6] mencantumkan dan membandingkan hasil penerapan tiga algoritma *Decision Tree* yang berbeda. Grafik *Decision Tree* dipengaruhi oleh jumlah atribut input dan atribut kelas akhir. Dua kelas utama (Kesuksesan Siswa/G3 Grade) dan Kemauan Belajar Pendidikan Tinggi (Tinggi) dipilih untuk membangun grafik pohon. Hasil menunjukkan J48 adalah algoritma pohon keputusan terbaik yang dapat digunakan sebagai peta jalan prediksi dan klasifikasi tindakan siswa.

Dalam penelitian Pattekari dan Parveen [7] pendukung keputusan dalam sistem prediksi penyakit jantung dikembangkan dengan menggunakan teknik *Naive Bayesian Classification*. Sistem mengekstrak pengetahuan tersembunyi dari database riwayat penyakit jantung. Ini adalah model yang paling efektif untuk memprediksi pasien dengan penyakit jantung. Model ini dapat menjawab pertanyaan yang rumit, masing-masing dengan kekuatannya sendiri berhubungan dengan kemudahan interpretasi model, akses ke informasi terperinci, dan akurasi.

Sebuah studi yang dilakukan Iman dan Tian [8] mengidentifikasi model klasifikasi *Naïve Bayes* dan *Logistic Regression* sebagai model prediktif yang baik untuk lulusan berdasarkan parameter kandidat. *Naïve Bayes* mengungguli *Logistic Regression* dengan selisih yang sangat kecil. Naive Bayes memiliki beberapa keunggulan yang menjadikannya model klasifikasi yang diterima secara luas, disukai oleh analis di banyak bidang. Beberapa kelebihannya adalah desain sederhana yang memungkinkan implementasi yang mudah dan cepat; tidak membutuhkan data pelatihan yang besar; dapat menangani variabel kontinu dan diskrit.

Dalam penelitian Hasanah, Sudiboyo dan Kurniawan [9], *Naïve Bayes* diterapkan untuk mencari nilai probabilitas terbesar pada setiap variabel. Variabel yang digunakan meliputi nilai rata-rata dari matematika, bahasa Indonesia, dan bahasa Inggris semester 1 s/d 5 serta prestasi siswa yang dilampirkan saat mendaftar dan prestasi sekolah. Hasilnya dengan *Naïve Bayes* mampu menghasilkan akurasi sebesar 83,3%.

Penelitian Putri dan Pudjiantoro [10] menggunakan metode *Naïve Bayes* dalam sistem yang dibangun untuk melakukan prediksi SNMPTN di sekolah XYZ. Hasil prediksi menunjukkan akurasi yang cukup tinggi. Manfaat dari penelitian ini dapat mempermudah bagian Bimbingan Konseling (BK) untuk mengarahkan siswa/i mendaftar PTN.

Studi yang dilakukan oleh Tolles dan Meurer [11] asosiasi yang ditemukan melalui model *Logistic Regression* cenderung memberikan wawasan tentang apa yang mungkin terjadi pada populasi serupa dari pasien di masa depan. Kombinasi tertentu dari karakteristik dan faktor pasien mungkin jarang terwakili dalam kumpulan data (misalnya, pasien muda dengan sepsis dan skala score koma Glasgow rendah tetapi tekanan darah dan laju pernapasan normal), dan perkiraan model untuk kematian di antara pasien tersebut harus dipertimbangkan dengan hati-hati. Prediktor yang termasuk dalam model *Logistic Regression* harus dipilih untuk menghindari redundansi dalam informasi yang mereka berikan (kolinearitas). Penting juga untuk mempertimbangkan kemungkinan bahwa nilai satu prediktor dapat mengubah efek yang lain (interaksi). Kedua situasi ini dapat mempengaruhi validitas model *Logistic Regression* yang dihasilkan.

Hasil penelitian oleh El Zeweidy dan Osman [12] menunjukkan bahwa skor *Decision Tree*, *Clustering*, dan *Naïve Bayes* sedikit lebih baik daripada tiga lainnya untuk himpunan {lulus, gagal} dan {sangat baik, sangat baik, baik, lulus, gagal, sangat buruk, tidak ada} sedangkan Association Rules, memiliki skor paling sedikit untuk kedua set. Hasil studi kasus ini memberikan wawasan tentang teknik untuk memprediksi kinerja siswa secara akurat dan membandingkan keakuratan algoritma data mining.

Pengujian pertama dalam penelitian Perez dan Perez [13] menunjukkan akurasi 73%, sedangkan eksperimen kedua menggunakan semua empat atribut teratas mencapai akurasi 84% dengan presisi kelas 80,46% dan ingatan kelas 83,33%; model tersebut menghasilkan hasil yang menarik. Menggunakan *Naïve Bayes* Classifier untuk mengembangkan model prediksi untuk penyelesaian program layak dilakukan. Ini dapat dimanfaatkan untuk menyediakan alat pendidikan yang kuat dalam menentukan kemungkinan siswa untuk menyelesaikan atau mempertahankan.

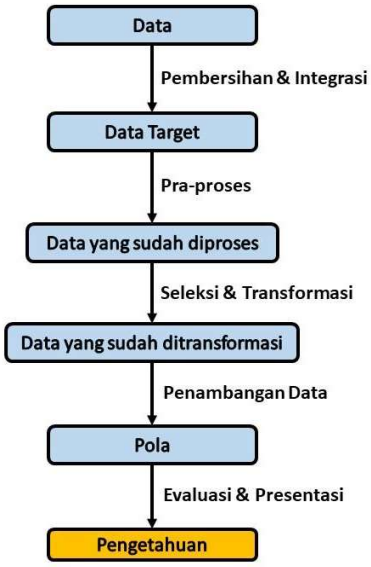
Rumusan masalah dalam penelitian ini adalah tentang bagaimana tingkat akurasi model prediksi SNMPTN dengan menggunakan metode *Decision Tree* CART, *Gaussian Naïve Bayes* dan *Logistic Regression*? Dan bagaimana perbandingan akurasi dari ketiga model tersebut? Adapun tujuan penelitian ini adalah untuk membuat model prediksi, melakukan prediksi dan membandingkan akurasi model prediksi SNMPTN menggunakan metode *Decision Tree* CART, *Gausssian Naïve Bayes* dan *Logistic Regression*. Untuk menghindari *overfitting* maupun *underfitting*, pemilihan model yang akan digunakan harus sesuai dengan karakteristik dataset yang digunakan. Berdasarkan karakteristik data yang ada dan tujuan prediksi dalam penelitian ini, maka model *classification* seperti *Decision Tree* CART, *Gaussian Naïve Bayes* dan *Logistic Regression* merupakan model yang cocok diterapkan. Hasil penelitian yang menunjukkan prediktor-prediktor apa saja yang dominan akan digunakan sebagai masukan bagi sekolah dalam pendampingan siswa-siswinya mengikuti SNMPTN.

II. METODE PENELITIAN

Metodologi penelitian adalah cara ilmiah yang digunakan dalam upaya untuk menemukan atau mendapatkan data demi tujuan atau kegunaan tertentu. Metodologi yang digunakan dalam penelitian ini adalah *Knowledge Discovery in Database*

(KDD) [14]. KDD sendiri merupakan ekstraksi implisit dari informasi yang sebelumnya tidak diketahui serta berpotensi mendapatkan pengetahuan berguna yang didapatkan dari data [5].

Proses penemuan pengetahuan dalam database atau *Knowledge Discovery in Database* (KDD) tentu saja memerlukan data. Data yang digunakan adalah data yang sudah terpisah dengan data operasional. Adapun tahapan dalam melakukan KDD seperti yang tercantum pada Gambar 1.



Gambar 1. Langkah-Langkah yang Menyusun Proses KDD.

Secara terperinci, langkah kerja dalam penelitian ini mencakup:

1) Pembersihan dan Integrasi Data

Dalam langkah ini dilakukan pengumpulan data dan data integration yang telah dikumpulkan.

- Pembersihan data (*Data cleaning*) bertujuan untuk menghilangkan noise dan data yang tidak konsisten.
- Penggabungan data (*Data integration*) bertujuan untuk menggabungkan data-data yang berasal dari berbagai sumber.

2) Seleksi dan Transformasi

Langkah selanjutnya adalah data selection dan transformation

- Seleksi data (*Data selection*) berfungsi untuk mengambil data dari database yang relevan dengan tugas analisis.
- Transformasi data (*Data transformation*) berfungsi mentransformasikan dan mengkonsolidasikan data ke dalam bentuk yang sesuai untuk penambangan dengan melakukan operasi ringkasan atau agregasi.

3) Penambangan Data

Penambangan data (Data mining) diterapkan untuk mengekstrak pola data.

4) Evaluasi dan Presentasi

- Evaluasi pola (*Pattern evaluation*) untuk mengidentifikasi pola-pola yang benar-benar menarik yang mewakili pengetahuan berdasarkan ukuran-ukuran yang menarik.
- Presentasi pengetahuan (*Knowledge presentation*) digunakan untuk memvisualisasikan dan merepresentasikan pengetahuan yang ditambang.

5) Pembuatan Model

Dalam langkah ini, model dibuat menggunakan metode *Decision Tree* CART, *Gaussian Naïve Bayes* dan *Logistic Regression*. Data yang digunakan menggunakan data nilai pelamar SNMPTN SMA Santa Maria 1 Cirebon dari angkatan 2015 hingga 2022. Kelas pada model prediksi berupa binomial yaitu lolos (*yes*) dan tidak lolos (*no*).

6) Evaluasi dan Interpretasi Model

Evaluasi model menggunakan *Split Percentage Method* dan *K-Fold Cross Validation Method*.

III. HASIL DAN PEMBAHASAN

A. Dataset Penelitian

Dataset yang digunakan memiliki 193 rows dan 41 columns. Dataset diambil dari tahun 2015 hingga 2022. Keterangan-keterangan kolom yang ada pada dataset dalam penelitian ini dijelaskan dalam tabel 1.

TABEL 1
KETERANGAN KOLOM PADA DATASET

No	Columns	Detail	Possible Value
1.	nis	Nomor Induk Siswa	15160001 - 19200136
2.	class	Jurusan di SMA	IPA (1), IPS (2)
3.	year	Tahun Lulus	2015 s/d 2022
4.	math-1 s/d math-5	Nilai Matematika Semester 1 s/d 5	0 - 100
5.	indo-1 s/d indo-5	Nilai B. Indonesia Semester 1 s/d 5	0 - 100
6.	eng-1 s/d eng-5	Nilai B. Inggris Semester 1 s/d 5	0 - 100
7.	fiso-1 s/d fiso-5	Nilai Fisika/Sosiologi Semester 1 s/d 5	0 - 100
8.	kiko-1 s/d kiko-5	Nilai Kimia/Ekonomi Semester 1 s/d 5	0 - 100
9.	bige-1 s/d bige-5	Nilai Biologi/Geografi Semester 1 s/d 5	0 - 100
10.	tot_scor	Total Skor Nilai Rapor	0 - 3000
11.	sch_acr	Akreditasi Sekolah	4 (A), 3 (B), 2 (C), 1 (D)
12.	snmptn_adms	Pendaftaran SNMPTN	1 (yes), 0 (no)
13.	colg_rank	Ranking PTN Pilihan	1 - 101
14.	quota	Kuota Jurusan PTN Pilihan	6 – 228
15.	appl	Total Pelamar Jurusan PTN Pilihan	6 – 3745
16.	linier	Linieritas Jurusan SMA dan PTN	1 (yes), 0 (no)
17.	snmptn	Status SNMPTN	1 (yes), 0 (no)

Jumlah total angkatan siswa SMA Santa Maria 1 Cirebon yang digunakan adalah 8 tahun. Dimulai dari tahun 2015 sampai dengan tahun 2022. Total siswa dari 8 tahun tersebut adalah 961 siswa. Total siswa yang masuk melalui jalur SNMPTN sebanyak 193 siswa. Data-data dari 193 siswa ini yang dijadikan sebagai dataset. Persentase jumlah siswa pelamar SNMPTN per jumlah total siswa 8 angkatan adalah 20,08%.

B. Exploratory Data Analysis

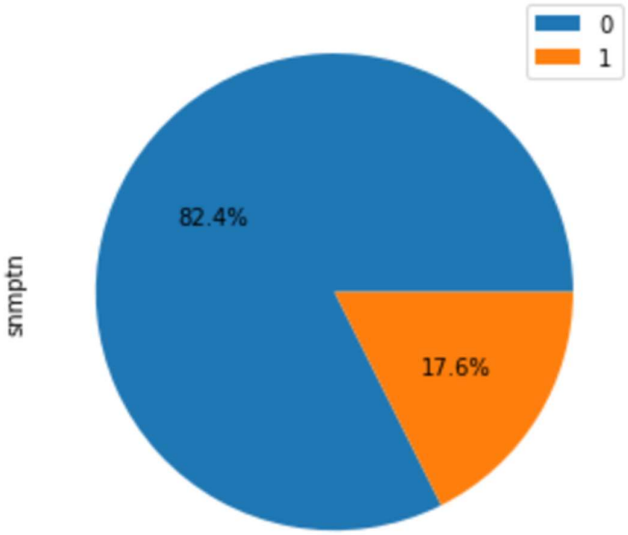
Dalam penelitian ini dilakukan *Exploratory Data Analysis* untuk mendapat pengetahuan-pengetahuan dari dataset yang digunakan. Beberapa hal yang ditemukan dari dataset dijelaskan pada tabel 2.

TABEL 2
REKAP DATA PELAMAR BERDASAR PTN PILIHAN

No	PTN	Siswa Pelamar	Linieritas Jurusan		Total	Status SNMPTN		Total
			Ya	Tidak		Ya	Tidak	
1.	ITB	46	37	9	46	10	36	46
2.	UNPAD	31	29	2	31	8	23	31
3.	UGM	25	25	0	25	0	25	25
4.	UNSOED	22	22	0	22	2	20	22
5.	UI	19	17	2	19	4	15	19
6.	UNNES	14	14	0	14	2	12	14
7.	UNDIP	8	7	1	8	2	6	8
8.	UPI	8	6	2	8	2	6	8
9.	UB	3	3	0	3	0	3	3
10.	IPB University	3	3	0	3	2	1	3
11.	UNY	3	3	0	3	1	2	3
12.	UNAIR	2	2	0	2	0	2	2
13.	UNS	2	2	0	2	1	1	2
14.	ISI Yogyakarta	2	1	1	2	0	2	2
15.	ITS	1	1	0	1	0	1	1
16.	UNHAS	1	1	0	1	0	1	1
17.	UNSRAT	1	1	0	1	0	1	1

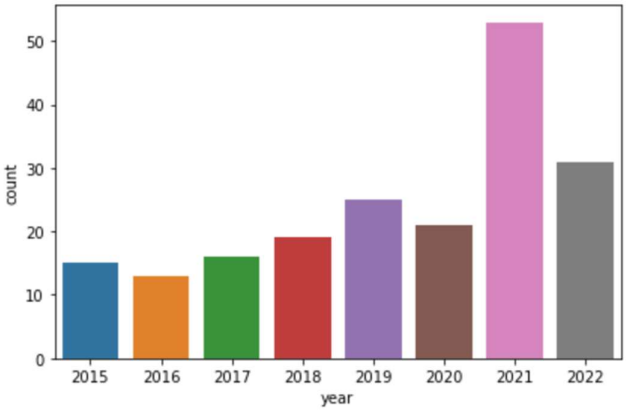
18.	UNUD	1	1	0	1	0	1	1
	Total Siswa	193	176	17	193	34	159	193

Berdasarkan linieritas jurusan di SMA dan PTN pilihan, dari 193 siswa pelamar, terdapat 176 siswa yang linier dengan jurusan PTN yang dipilih. Persentasenya ada di angka 91,2%.



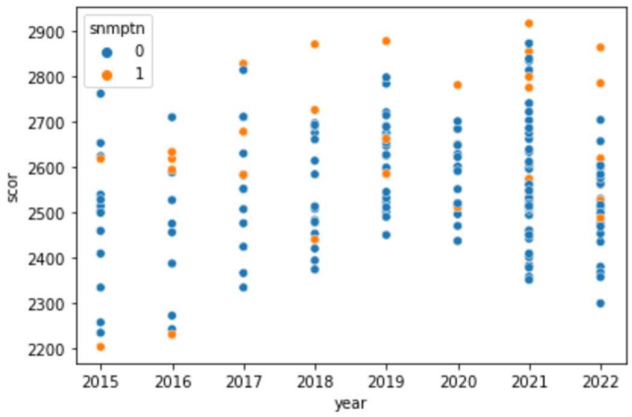
Gambar 2. Persentase Siswa Yang Lolos SNMPTN

Gambar 2 menjelaskan status diterima atau tidaknya SNMPTN. Dari total 193 pelamar, terdapat 34 siswa yang berhasil lolos SNMPTN. Jika dihitung persentase siswa lolos SNMPTN adalah sebesar 17,6%.



Gambar 3. Grafik Jumlah Siswa Pelamar SNMPTN

Jika melihat jumlah siswa pelamar SNMPTN per tahunnya dari total siswa 8 angkatan, dapat dilihat grafik batang pada gambar 3. Dari total jumlah siswa sebanyak 961, terdapat 193 siswa yang melamar. Jumlah tertinggi ada pada angkatan 2021 yaitu sebanyak 53 siswa dan yang paling rendah ada pada angkatan 2016 yaitu sebanyak 13 siswa. Rata-rata jumlah pelamar per tahunnya yaitu 24,1. Untuk perolehan nilai total (score) mata pelajaran yang dijadikan indikator untuk SNMPTN beragam. Sebaran nilai total siswa dapat dilihat melalui gambar 4 berikut.

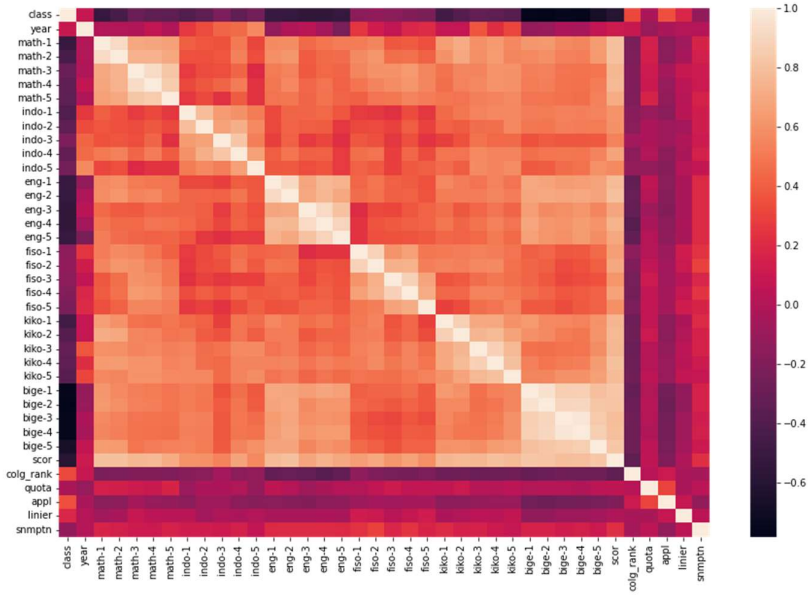


Gambar 4. Grafik Total Skor Siswa Pelamar SNMPTN Per Tahun

Skor tertinggi untuk 8 angkatan dipegang oleh siswa dengan NIS 181910032 (angkatan 2021) yaitu sebesar 2918 poin. Sedangkan skor terendah dipegang oleh siswa dengan NIS 121310042 (angkatan 2015) yaitu sebesar 2204 poin. Jika melihat grafik di atas, ada hal menarik yang dapat diperhatikan. Sebaran siswa yang lolos SNMPTN memiliki nilai yang beragam, bahkan siswa dengan nilai rendah pun, ada yang berhasil lolos SNMPTN. 2 siswa yang lolos SNMPTN dengan nilai yang rendah yaitu: 1) NIS 121310042 (angkatan 2015) yaitu sebesar 2204 poin, dan 2) NIS 131410092 (angkatan 2016) yaitu sebesar 2231 poin.

Setelah ditelusuri siswa dengan NIS 121310042 memilih jurusan yang non-linier dan masuk ke jurusan FSRD (Fakultas Seni Rupa dan Desain) ITB. Walau dengan nilai yang rendah, portofolio berupa hasil karya seni dapat memperbesar peluang masuk ke jurusan yang diinginkan. Sedangkan siswa dengan NIS 131410092 memilih jurusan yang linier tapi dengan peminat yang rendah, yaitu Sastra Rusia UNPAD.

Pada gambar 5 dapat dilihat *Correlation Matrix* dari tiap kolom prediktor yang ada di dalam dataset. Warna putih menunjukkan nilai 1,0 sedangkan warna hitam menunjukkan nilai -1,0.



Gambar 5. *Correlation Matrix* Dari Tiap Kolom Predikor

Nilai tepat 1,0 dalam suatu *correlation coefficient* adalah tanda bahwa ada hubungan positif sempurna antara kedua variabel. Untuk peningkatan positif pada satu variabel, ada juga peningkatan positif pada variabel kedua. Sementara itu, nilai -1,0 berarti terdapat hubungan negatif sempurna antara kedua variabel tersebut. Hal ini menunjukkan bahwa variabel bergerak

berlawanan arah. Untuk kenaikan positif pada satu variabel, ada penurunan pada variabel kedua. Jika korelasi antara dua variabel adalah 0, maka tidak ada hubungan linier di antara keduanya.

C. Komparasi Model Machine Learning

Model *Machine Learning* yang dibuat ada 3, yaitu: *Decision Tree* CART, *Gaussian Naïve Bayes* dan *Logistic Regression*. Dengan menggunakan bahasa pemrograman *Python 3* ketiga model tersebut dibuat. Karena dataset yang digunakan sudah dalam bentuk numerikal, sehingga *library* yang diimpor adalah *numpy* dan *pandas* untuk memulai membuat model. Ada dua metode evaluasi model yang digunakan, yaitu: 1) *Split Percentage Method*, dan 2) *K-Fold Cross Validation* dengan 5 *fold*. Untuk *Split Percentage Method*, pembagian data training dan data testing dibagi menjadi 3 metode seperti yang tercantum pada tabel 3 berikut.

TABEL 3
TRAINING DAN TESTING DATA PADA *SPLIT PERCENTAGE METHOD*

Metode	Data Training	Data Testing
90 – 10	90%	10%
80 – 20	80%	20%
70 – 30	70%	30%

Dan hasil evaluasi dari ketiga model tersebut dengan menggunakan *Split Percentage Method* dapat dilihat di tabel 4 berikut.

TABEL 4
KOMPARASI AKURASI MODEL MENGGUNAKAN *SPLIT PERCENTAGE METHOD*

Metode	Data Training	Data Testing	*Decision Tree*	*Naïve Bayes*	*Logistic Regression*
90 – 10	90%	10%	0,85	0,80	1,00
80 – 20	80%	20%	0,77	0,67	0,85
70 – 30	70%	30%	0,72	0,67	0,69
Rata-rata			**0,78**	**0,71**	**0,85**

Model *Logistic Regression* memiliki nilai rata-rata akurasi paling tinggi jika dibanding lainnya. Dapat dilihat berdasar tabel 4, jika Data testingnya dinaikan, maka akurasi model akan memiliki kecenderungan menurun, hal ini berlaku untuk ketiga model. Nilai akurasi paling tinggi ada pada model *Logistic Regression* pada metode dengan data testing 10% dengan skor 1,00. Akurasi merupakan rasio prediksi benar (positif dan negatif) dengan keseluruhan data. Nilai 1,00 menandakan bahwa prediksi benar (positif dan negatif) dalam hal ini terkait prediksi 'lolos' dan 'tidak lolos' SNMPTN dengan data aktual (kenyataan) adalah sama. Evaluasi model juga menggunakan *K-Fold Cross Validation* dengan menggunakan 5 *fold* (K=5). Komparasi akurasi model menggunakan *K-Fold Cross Validation* dapat dilihat pada tabel 5.

TABEL 5
KOMPARASI AKURASI MODEL MENGGUNAKAN *K-FOLD CROSS VALIDATION*

Fold	Data Training	Data Testing	*Decision Tree*	*Naïve Bayes*	*Logistic Regression*
1	154 (80%)	39 (20%)	0,79	0,64	0,72
2	154 (80%)	39 (20%)	0,77	0,62	0,77
3	154 (80%)	39 (20%)	0,64	0,82	0,82
4	155 (80%)	38 (20%)	0,61	0,74	0,79
5	155 (80%)	38 (20%)	0,76	0,66	0,87
Rata-rata			**0,71**	**0,69**	**0,79**

Pemilihan nilai 'K' harus sedemikian rupa sehingga model tidak mengalami varian dan bias tinggi. Dalam kebanyakan kasus, pilihan K biasanya 5 atau 10, tetapi tidak ada aturan formal. Namun, nilai K bergantung pada ukuran dataset. Runtime dari algoritma *cross validation* dan biaya komputasi dengan nilai K yang besar. Penggunaan K=5 dan K=10 tidak menunjukkan perbedaan signifikan. Semakin besar nilai K tidak memperlihatkan perbedaan yang berarti. Hal itu sangat jelas terlihat berbeda jika dibandingkan dengan K=1 atau K=3. Nilai K=5 ini dapat dibilang sangat ideal untuk diterapkan dalam evaluasi model di penelitian ini. Rasio data training dengan data testing pada K-Fold Cross Validation yaitu 80% : 20%.

Dengan menggunakan *K-Fold Cross Validation*, model *Logistic Regression* memiliki rata-rata akurasi paling tinggi di angka 0,79 diikuti oleh *Decision Tree* CART di angka 0,71 dan *Gaussian Naïve Bayes* di angka 0,69. Dapat dilihat pola yang sama baik menggunakan *Split Percentage Method* maupun *K-Fold Cross Validation*, yaitu urutan akurasi model dari yang

paling tinggi: 1) *Logistic Regression*, 2) *Decision Tree* CART, dan 3) *Gaussian Naïve Bayes*. Secara detail untuk rata-rata akurasi model menggunakan 2 metode evaluasi dapat dilihat pada tabel 6.

TABEL 6
RATA-RATA AKURASI MODEL MENGGUNAKAN *SPLIT PERCENTAGE METHOD* DAN *K-FOLD CROSS VALIDATION*

Method	*Decision Tree*	*Naïve Bayes*	*Logistic Regression*
Split Percentage Method	0,78	0,71	0,85
K-Fold Cross Validation	0,71	0,69	0,79
Total Average	**0,75**	**0,70**	**0,82**

Model *Logistic Regression* memiliki rata-rata akurasi paling tinggi jika dibandingkan dengan model yang lainnya.

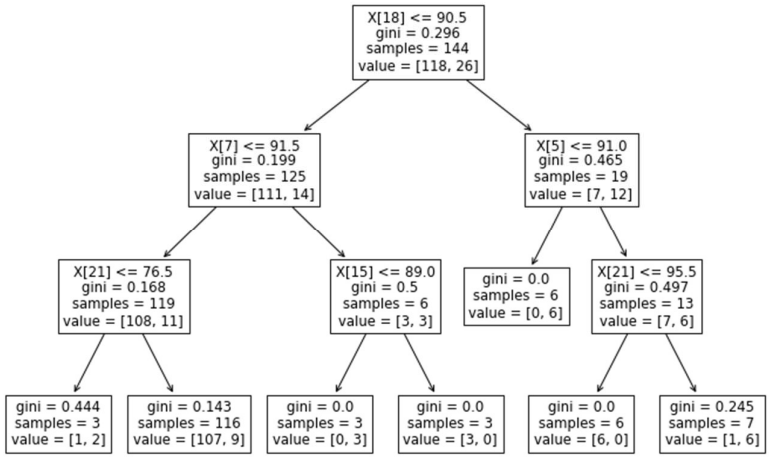
Pada tabel 7 dapat dilihat secara detail mengenai *Classification Report* dari masing-masing model. Kategori 0 menunjukkan kategori negatif sedangkan kategori 1 menunjukkan kategori positif.

TABEL 7
SKOR UNTUK PRECISION, RECALL DAN F1 MASING-MASING MODEL

Method	Kategori	Precision Score			Recall Score			F1-Score		
		DT	G-NB	LR	DT	G-NB	LR	DT	G-NB	LR
90 - 10	0	1,00	1,00	1,00	0,80	0,80	1,00	0,89	0,89	1,00
80 - 20	0	0,84	0,88	0,86	0,94	0,70	0,97	0,89	0,78	0,91
70 - 30	0	0,89	0,94	0,85	0,80	0,67	0,78	0,85	0,78	0,82
Average		0,91	0,94	0,90	0,85	0,72	0,92	0,88	0,82	0,91

Method	Kategori	Precision Score			Recall Score			F1-Score		
		DT	G-NB	LR	DT	G-NB	LR	DT	G-NB	LR
90 - 10	1	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
80 - 20	1	0,00	0,23	0,50	0,00	0,50	0,17	0,00	0,32	0,25
70 - 30	1	0,17	0,23	0,00	0,29	0,71	0,00	0,21	0,34	0,00
Average		0,06	0,15	0,17	0,10	0,40	0,06	0,07	0,22	0,08

Pohon yang dibangun dengan algoritma *Decision Tree* CART dapat dilihat pada gambar 6.

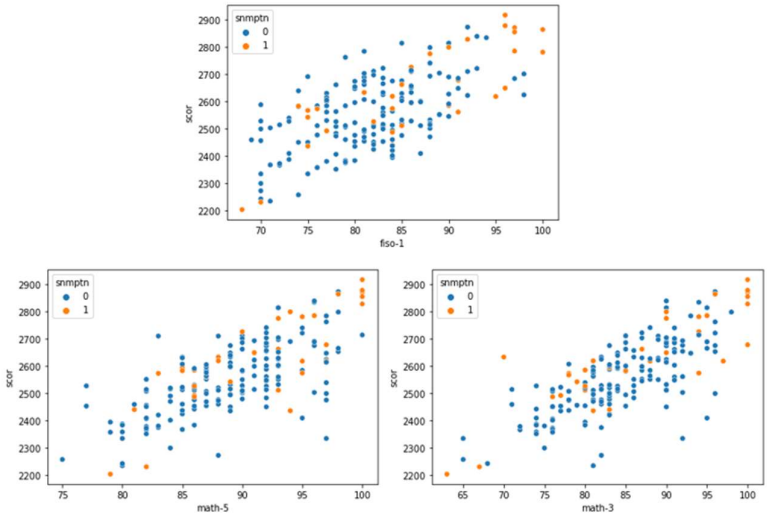


Gambar 6. Visualisasi Pohon Keputusan Yang Dibangun

Dari gambar 6 dapat dilihat *node* utama dari pohon keputusan yang dibangun merupakan X[18] yang merupakan kolom fiso-1 sebagai prediktor dengan nilai gini sebesar 0,296. Node dibawahnya yaitu X[7] untuk kolom math-5 dan X[5] untuk kolom math-3 dengan masing-masing nilai gini sebesar 0,199 dan 0,465.

Klasifikasi dimulai dari *root node* (*depth* 0, *top node*) dan memeriksa apakah nilai X[18] (fiso-1) lebih kecil sama dengan 90,5. Jika *true*, pindah ke *child root node* sebelah kiri (*depth* 1, kiri). Jika *false*, pindah ke *child root node* sebelah kanan (*depth* 1, kanan). Klasifikasi selanjutnya dilihat di bagian kiri dan memeriksa apakah X[7] (math-5) lebih kecil sama dengan 91,5. Jika *true*, pindah ke *child root node* sebelah kiri (*depth* 2, kiri). Jika *false*, pindah ke *child root node* sebelah kanan (*depth* 2, kanan).

Untuk level yang sama, klasifikasi selanjutnya dilihat di bagian kanan dan memeriksa apakah X[5} (math-3) lebih kecil sama dengan 91. Jika *true*, pindah ke *child root node* sebelah kiri (*depth* 3, kiri). Jika *false*, pindah ke *child root node* sebelah kanan (*depth* 3, kanan). Hal ini berlaku untuk langkah selanjutnya sampai ditemukan *leaf node* yang tidak dapat dibagi lagi. Berdasarkan pohon keputusan yang dibangun, akan dilihat hubungan antara kolom skor total (scor) dengan kolom yang menjadi *root node* (fiso-1) dengan 2 *node* di bawahnya (math-5 dan math-3) pada gambar 7.



Gambar 7. Grafik Hubungan Antara Skor Total dengan Nilai Fiso-1, Math-5 dan Math-3

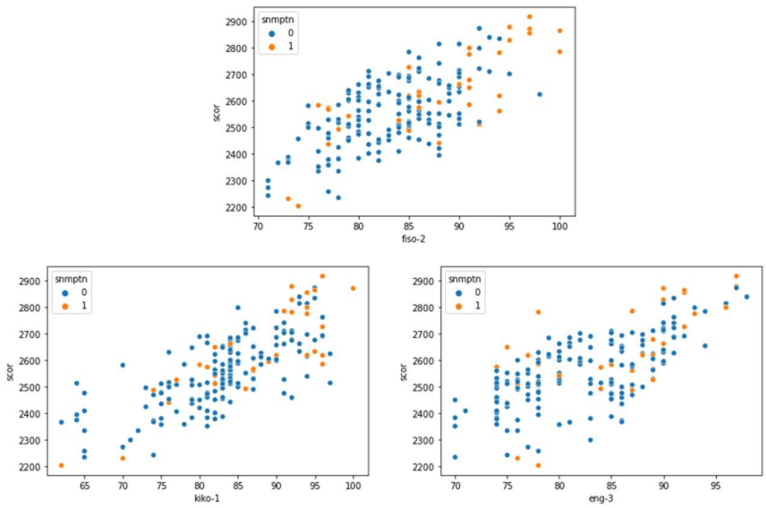
Jika melihat gambar 7 dapat dilihat adanya hubungan berbanding lurus antara total skor nilai siswa dengan nilai dari kolom fiso-1, math-5 dan math-3 sebagai prediktor. Total skor naik, begitupun dengan nilai fiso-1, math-5 dan math-3.

Pada tabel 8 dapat dilihat skor koefisien dari *Logistic Regression*.

TABEL 8
SKOR KOEFISIEN PADA *LOGISTIC REGRESSION*

No.	Col.	Score	No.	Col.	Score	No.	Col.	Score	No.	Col.	Score
1	fiso-2	0,096	11	math-5	0,022	21	eng-1	0,001	31	bige-3	-0,021
2	kiko-1	0,044	12	indo-4	0,018	22	sch_acr	0,000	32	fiso-3	-0,024
3	eng-3	0,042	13	colg_rank	0,013	23	snmptn_adms	0,000	33	math-2	-0,029
4	fiso-1	0,040	14	indo-3	0,008	24	bige-2	0,000	34	indo-5	-0,042
5	fiso-4	0,039	15	math-1	0,007	25	appl	-0,001	35	math-3	-0,055
6	indo-1	0,029	16	quota	0,007	26	class	-0,002	36	bige-5	-0,057
7	eng-2	0,028	17	bige-4	0,006	27	fiso-5	-0,002	37	kiko-5	-0,058
8	eng-5	0,026	18	scor	0,005	28	year	-0,007	38	kiko-4	-0,061
9	indo-2	0,025	19	linier	0,003	29	bige-1	-0,014	39	math-4	-0,070
10	eng-4	0,024	20	kiko-3	0,002	30	kiko-2	-0,016			

Berdasarkan nilai koefisien pada Logistic Regression di tabel 8 dapat dilihat grafik yang menjelaskan hubungan antara skor total (scor) dengan 3 kolom yang memiliki koefisien paling tinggi, yaitu fiso-2, kiko-1 dan eng-3 pada gambar 8 berikut.



Gambar 8. Grafik Hubungan Antara Skor Total dengan Nilai Fiso-2, Kiko-1 dan Eng-3

Pada gambar 8 dapat dilihat adanya hubungan yang berbanding lurus antara total skor nilai siswa dengan nilai dari kolom fiso-2, kiko-1 dan eng-3. Total skor naik, begitupun dengan nilai fiso-2, kiko-1 dan eng-3.

D. Diskusi dan Evaluasi

Sebanyak 193 siswa dari 8 angkatan menjadi peserta SNMPTN dari SMA Santa Maria 1 Cirebon. Dari 193 siswa yang mendaftar terdapat 34 siswa yang lolos. Ini berarti dari total siswa pelamar, hanya 17,6% yang lolos (Gambar 2). Rata-rata jumlah pelamar tiap tahunnya yaitu ± 24 siswa. Sementara itu rata-rata untuk jumlah siswa yang lolos SNMPTN tiap tahunnya adalah ± 4 siswa. Skor total nilai siswa yang tertinggi untuk yang lolos SNMPTN adalah 2918, sedangkan skor total nilai siswa yang terendahnya adalah 2204 dengan NIS 121310042. Sementara itu untuk skor total nilai siswa yang tertinggi untuk yang tidak lolos SNMPTN adalah 2874, sedangkan skor total nilai siswa yang terendahnya adalah 2235 (Gambar 4).

Terjadi hal yang menarik disini, dimana siswa dengan skor yang rendah (2204) bisa lolos SNMPTN jika dibandingkan dengan siswa dengan skor yang lebih tinggi (2874). Siswa dengan NIS 121310042 yang memiliki skor terendah tersebut memilih jurusan FSRD (Fakultas Seni Rupa dan Desain) ITB. Ini menunjukkan berarti siswa dengan skor tinggi sekalipun dapat menjadi tidak lolos SNMPTN jika memilih jurusan dan PTN yang memiliki tingkat persaingan tinggi. Pemilihan jurusan dan PTN yang tepat mampu memperbesar peluang untuk dapat lolos SNMPTN.

Studi dilakukan untuk membandingkan 3 model yang dibuat menggunakan metode *Decision Tree* CART, *Gaussian Naïve Bayes* dan *Logistic Regression*. Ketiganya merupakan model klasifikasi yang dapat digunakan untuk memprediksi. Model dibuat menggunakan bahasa pemrograman *Python*. Evaluasi model menggunakan metode *Split Percentage Method* dan *K-Fold Cross Validation*. Hasil menunjukkan *Logistic Regression* memiliki skor akurasi paling tinggi jika dibandingkan dengan *Decision Tree* CART dan *Gaussian Naïve Bayes* (Tabel 6).

Algoritma *Decision Tree* CART menerapkan metode Indeks Gini untuk memulai pemisahan biner. Indeks Gini beroperasi pada variabel target kategori dalam hal "lolos" atau "tidak lolos" dan hanya melakukan pemisahan biner. Indeks Gini bervariasi antara nilai 0 dan 1, di mana 0 menyatakan kemurnian klasifikasi, yaitu semua elemen milik kelas tertentu atau hanya ada satu kelas di sana. Dan 1 menunjukkan distribusi acak elemen di berbagai kelas. Nilai Indeks Gini 0,5 menunjukkan pemerataan elemen pada beberapa kelas.

Indeks Gini digunakan untuk analisis skenario *real time*, dan data *real* yang diambil dari analisis waktu nyata. Dalam banyak definisi disebut sebagai ketidakmurnian data atau bagaimana data didistribusikan. Jadi dapat dihitung data mana yang kurang atau lebih banyak mengambil bagian dalam pengambilan keputusan. Berdasarkan indeks gini yang terdapat pada gambar 6, 3 kolom prediktor utama dari pohon keputusan yang dibangun adalah X[18] yaitu fiso-1, X[7] yaitu math-5 dan X[5] yaitu math-3. Ketiga prediktor ini memiliki korelasi yang signifikan terhadap kelas “lolos” atau “tidak lolos” – nya SNMPTN. Berdasarkan gambar 7 tentang grafik hubungan antara skor total dengan nilai fiso-1, math-5 dan math-3 menunjukkan hubungan yang berbanding lurus. Semakin tinggi nilai siswa dari ketiga kolom tersebut, makin tinggi pula skor total yang dimilikinya.

Naive Bayes mengasumsikan bahwa semua prediktor (atau fitur) bersifat independen satu sama lain, hal ini jarang terjadi dalam kehidupan nyata. Padahal variabel-variabel prediktor dalam kenyataannya ada *dependency* yang dapat dilihat pada gambar 5 tentang *Correlation Mateix* dari tiap kolom prediktor. Selain itu, dilihat dari adanya asumsi bahwa variabel-variabel prediktornya berdistribusi normal (*Gaussian*) yang dalam kenyataannya tidak semua variabel prediktor berdistribusi normal. Ini membatasi penerapan algoritma dalam kasus penggunaan di dunia nyata. Algoritma ini juga menghadapi 'masalah frekuensi nol' di mana ia memberikan probabilitas nol ke variabel kategori yang kategorinya dalam kumpulan data uji tidak tersedia dalam kumpulan data pelatihan. Hal ini yang mengakibatkan *Naïve Bayes* memiliki skor akurasi yang paling rendah jika dibandingkan dengan model lainnya.

Koefisien pada *Logistic Regression* menunjukkan bahwa kolom fiso-2 (Fisika & Sosiologi Semester 2) memiliki skor paling tinggi jika dibandingkan dengan kolom lainnya. Sementara itu untuk kolom colg_rank (Ranking PTN), quota (Kuota Jurusan), scor (Total Skor) dan linier (Linieritas Jurusan) ada pada urutan 13, 16, 18 dan 19 (Tabel 8). Koefisien di regresi menyatakan hubungan searah atau berbalik antara variabel independen dan variabel dependen. Jika koefisien bernilai positif, maka hubungan variabel independen dan variabel dependen adalah hubungan searah. Artinya jika terjadi peningkatan nilai X maka akan terjadi peningkatan di Y atau variabel dependennya.

Kolom fiso-2 yang merupakan prediktor yang memiliki nilai koefisien *Logistic Regression* yang paling tinggi yaitu 0,096 dilihat pada tabel 8. Hal ini menjelaskan jika nilai fiso-2 makin tinggi maka nilai peluang untuk 'lolos' SNMPTN juga semakin besar. Ini ditunjukkan oleh angka positif yang menunjukkan hubungan yang berbanding lurus (korelasi positif). Sementara itu kolom math-4 memiliki nilai koefisien *Logistic Regression* yang paling rendah yaitu -0,070 dilihat pada tabel 8. Ini menandakan hubungan yang berbanding terbalik (korelasi negatif). Hal ini menjelaskan jika nilai math-4 makin tinggi, maka nilai peluang untuk 'lolos' SNMPTN semakin kecil. Dengan tidak mengesampingkan kolom prediktor lainnya, prediktor-prediktor yang memiliki koefisien *Logistic Regression* dengan nilai positif besar dapat dijadikan fokus sekolah dalam mendampingi siswa mengikuti SNMPTN. Sekolah dapat mendorong siswa-siswi untuk dapat mendapatkan nilai besar pada prediktor-prediktor bernilai positif besar seperti fiso-2, kiko-1 dan eng-3.

Sama halnya dengan *Gaussian Naïve Bayes*, *Logistic Regression* pun memiliki asumsi bahwa variabel-variabel prediktornya saling *independent* satu sama lain. Namun hal yang membedakannya dengan *Gaussian Naïve Bayes* adalah tidak mempengaruhi akurasi, dan hanya berpengaruh terhadap penentuan koefisien regresi-nya. Hal ini dapat diantisipasi dan juga dapat dikurangi dengan cara menambah ukuran sampel [15].

IV. SIMPULAN

Penelitian ini memiliki tujuan utama untuk membandingkan model prediksi masuk SNMPTN dari tiga metode yang berbeda. Hal ini dilakukan untuk melihat metode mana yang memiliki tingkat akurasi paling tinggi. Hasil menunjukkan model *Logistic Regression* memiliki skor rata-rata paling tinggi jika dibanding dengan *Decision Tree* CART dan *Gaussian Naïve Bayes*. Secara berurutan dari yang paling tinggi, skor akurasi dari tiga model yang dibuat adalah: 1) *Logistic Regression*, 2) *Decision Tree* CART, dan 3) *Gaussian Naïve Bayes*. Akurasi sendiri merupakan rasio prediksi benar (positif dan negatif) dengan keseluruhan data. Akurasi menjawab pertanyaan tentang berapa persen siswa yang benar diprediksi lolos dan lolos dari keseluruhan siswa pelamar SNMPTN. Berdasar pohon keputusan yang dibangun ada 3 kolom prediktor yang memiliki skor gini besar yaitu fiso-1, math-5 dan math-3. Sedangkan berdasar koefisien *Logistic Regression,* 3 kolom prediktor yang memiliki skor positif besar yaitu fiso-2, kiko-1 dan eng-3. Fiso-1 maupun fiso-2 merupakan nilai Fisika (untuk jurusan IPA) atau nilai Sosiologi (untuk jurusan IPS) sama-sama memiliki skor tertinggi dilihat dari gini maupun koefisien regresi-nya. Sehingga dapat dikatakan untuk memperbesar peluang masuk SNMPTN, maka nilai untuk mapel-mapel tersebut harus besar. Saran yang dapat diberikan untuk sekolah adalah: 1) sekolah dapat fokus untuk melakukan pendampingan dalam hal menaikkan nilai untuk prediktor-prediktor yang bernilai besar, 2) sekolah dapat mengintegrasikan model prediksi ke dalam program yang berisi sistem prediksi yang dapat dibangun untuk melihat peluang siswa memilih jurusan dan PTN tertentu dan rekomendasi jurusan serta PTN lain yang memiliki peluang lebih besar.

Hipotesis yang diajukan oleh peneliti sesuai dengan hasil yang didapatkan, bahwa model *Logistic Regression* memiliki skor akurasi lebih tinggi. Secara matematis, *Logistic Regression* tidak mengasumsikan bahwa variabel-variabel prediktor-nya berdistribusi normal (*Gaussian*). Kecepatan dalam pemrosesan data juga baik walaupun dengan volume data besar. Pemecahan masalah dan koreksi kesalahan juga lebih mudah karena perhitungannya tidak terlalu rumit. Hal ini yang menyebabkan *Logistic Regression* memiliki nilai akurasi yang lebih tinggi dibanding model lainnya.

UCAPAN TERIMA KASIH

Ucapan terima kasih saya ucapkan kepada SMA Santa Maria 1 Cirebon, Yayasan Santo Dominikus dan Program Studi Magister Ilmu Komputer Universitas Kristen Maranatha yang telah mendukung terselesaikannya penelitian ini.

DAFTAR PUSTAKA

[1] Kemendikbud, "LTMPT - Lembaga Tes Masuk Perguruan Tinggi," Kementerian Pendidikan dan Kebudayaan Republik Indonesia, 2021. [Online]. Available: https://www.ltmpt.ac.id/?mid=4. [Accessed 16 May 2022].

[2] Y. Ramadhani, "Tirto.id," PT Tirta Adi Surya, 29 March 2022. [Online]. Available: https://tirto.id/resmi-pengumuman-snmptn-2022-120643-siswa-diterima-ini-cara-cek-gqod. [Accessed 16 May 2022].

[3] D. Whitenack, Machine Learning With Go, Birmingham: Packt, 2017.

[4] Y.-y. Song and Y. Lu, "Decision Tree Methods: Applications for Classification and Prediction," *Shanghai Archives of Psychiatry,* vol. 27, no. 2, 2015.

[5] M. Pandey and V. K. Sharma, "A Decision Tree Algorithm Pertaining to The Student Performance Analysis and Prediction," *International Journal of Computer Applications,* vol. 61, no. 13, pp. 1-5, 2013.

[6] A. Hamoud, "Selection of Best Decision Tree Algorithm for Prediction and Classification of Students' Action," *American International Journal of Research in Science, Technology, Engineering & Mathematics,* vol. 16, no. 1, pp. 26-32, 2016.

[7] S. A. Pattekari and A. Parveen, "Prediction System For Heart Disease Using Naive Bayes," *International Journal of Advanced Computer and Mathematical Sciences,* vol. 3, no. 3, pp. 290-294, 2012.

[8] A. Iman and X. Tian, "A Comparison of Classification Models in Predicting Graduate," *Journal of Higher Education Theory and Practice,* vol. 21, no. 7, pp. 219-230, 2021.

[9] H. Hasanah, N. A. Sudibyo and E. Kurniawan, "Prediksi Jurusan Pada Seleksi Nasional Masuk Perguruan Tinggi Negeri (SNMPTN) Menggunakan Metode Klasifikasi Naïve Bayes," *DoubleClick: Journal of Computer and Information Technology,* vol. 4, no. 1, pp. 55-60, 2020.

[10] R. A. I. Putri and T. H. Pudjiantoro, "Prediksi Perguruan Tinggi Negeri dengan Menggunakan Metode Naive Bayes," *Prosiding SISFOTEK,* vol. 4, no. 1, pp. 106-111, 2020.

[11] J. Tolles and W. J. Meurer, "Logistic Regression: Relating Patient Characteristics to Outcomes," *JAMA Guide to Statistics and Methods,* vol. 316, no. 5, pp. 533-534, 2016.

[12] M. El Zeweidy and E. Osman, "A comparative analysis of Techniques for Predicting Academic Performance," *Journal of the ACS,* vol. 7, 2013.

[13] J. G. Perez and E. S. Perez, "Predicting Student Program Completion Using Naive Bayes Classification Algorithm," *I.J. Modern Education and Computer Science,* vol. 3, pp. 57-67, 2021.

[14] H. Jiawei, M. Kamber and J. Pie, Data Mining: Concepts and Techniques, San Francisco: Morgan Kaufmann, 2012.

[15] N. A. M. R. Senaviratna and A. Cooray, "Diagnosing Multicollinearity of Logistic Regression Model," *Asian Journal of Probability and Statistics,* vol. 5, no. 2, pp. 1-9, 2019.