Jurnal Ilmu Komputer Agri-Informatika

Tersedia secara daring di:
https://jurnal.ipb.ac.id/index.php/jika

Volume 12 Nomor 1 halaman 59–71
eISSN: 2654-9735, pISSN: 2089-6026

Pengembangan Model Prediksi Kelulusan Calon Mahasiswa Sarjana pada Sistem Seleksi SNMPTN IPB

Development of a Graduation Prediction Model for Undergraduate Applicants in the SNMPTN Admission System at IPB University

WADUDI MUTHAHARI[1]*, SONY HARTONO WIJAYA[1], UTAMI DYAH SYAFITRI[2]

Abstrak

Sejak tahun 2019, proses seleksi SNMPTN di IPB menggunakan media seleksi berbasis web dan algoritma tertentu. Namun, proses tersebut belum menerapkan pemodelan berbasis *machine learning* yang dapat memberikan rekomendasi peluang seorang siswa diterima sebagai mahasiswa IPB. Penelitian ini bertujuan mencari faktor-faktor yang mempengaruhi kelulusan calon mahasiswa jalur SNMPTN IPB dan mengembangkan pemodelan *machine learning* dengan menggunakan *Random Forest* dan Regresi Logistik Biner. Ada empat model yang dibangun dan semua model dilatih menggunakan *hyperparameter tuning*. Model pertama menggunakan semua fitur dan tanpa proses penyeimbangan data. Model kedua menggunakan semua fitur dan SMOTE. Model ketiga menggunakan seleksi fitur dan SMOTE serta model keempat dengan pemilihan fitur berdasarkan rekomendasi pakar dan SMOTE. Hasil pengujian menunjukkan bahwa model yang menggunakan data uji dengan penyeimbangan data SMOTE secara konsisten menunjukkan nilai *Recall* yang lebih tinggi dibandingkan dengan model tanpa penyeimbangan data. Model ketiga dengan Regresi Logistik Biner pada data Jawa Barat dan Model kedua dengan Regresi Logistik Biner pada data Non Jawa Barat menunjukkan nilai *Recall* terbaik sebesar 88.93% dan 86.91%. Hasil pemodelan juga menunjukkan bahwa urutan dalam pemilihan perguruan tinggi, kategori indeks sekolah, prestasi, dan pilihan program studi memiliki pengaruh signifikan terhadap prediksi kelulusan pelamar.

Kata Kunci**:** Model, *Random Forest*, Regresi Logistik Biner, SMOTE, SNMPTN IPB

Abstract

Since 2019, the SNMPTN selection process at IPB has used web-based selection media and specific algorithms. However, the process has not yet implemented machine learning-based modeling that can provide recommendations on a student's likelihood of being accepted as an IPB student. This study aims to find out what factors influence prospective students passing the IPB SNMPTN pathway and to develop machine learning modeling using Random Forest and Binary Logistic Regression. Four models were built and trained using hyperparameter tuning. The first model uses all features without balancing. The second model uses all features and SMOTE. The third model uses feature selection and SMOTE, and the fourth uses feature selection by Expert Adjustment (EA) and SMOTE. The results show that the models tested using test data with SMOTE data balancing consistently show higher recall values compared to models without data balancing. The third model with Binary Logistic Regression on West Java data and the second model with Binary Logistic Regression on Non-West Java data show the best recall values of 88.93% and 86.91%, respectively. The modeling results also show that the order of college selection, school index category, academic achievements, and program of study choice significantly impact the prediction of applicants' passing.

*Keywords**: Binary Logistic Regression, Model, Random Forest, SMOTE, SNMPTN IPB***

[1] Program Studi Ilmu Komputer, Sekolah Sains Data, Matematika dan Informatika, IPB University, Bogor 16680;
[2] Program Studi Statistika dan Sains Data, Sekolah Sains Data, Matematika dan Informatika, IPB University, Bogor 16680;
* Penulis Korespondensi: Tel/Faks: 085365045005; Surel: dudimuthahari03@apps.ipb.ac.id

PENDAHULUAN

Institut Pertanian Bogor (IPB) merupakan institusi pendidikan yang telah dan akan terus berperan penting dalam pembangunan sumber daya manusia (RJPIPB 2017). IPB membangun sumber daya manusia dengan melakukan seleksi kepada para calon mahasiswa baru melalui beberapa mekanisme, salah satunya melalui Seleksi Nasional Masuk Perguruan Tinggi Negeri (SNMPTN) (Permen 2018). SNMPTN merupakan pola seleksi berdasarkan potensi akademik siswa selama di sekolah pada masing-masing Perguruan Tinggi Negeri di bawah koordinasi Lembaga Tes Masuk Perguruan Tinggi (LTMPT). Penerimaan mahasiswa baru dilakukan dengan mengedepankan prinsip kredibel, adil, transparan, fleksibel, efisien, dan akuntabel serta tidak diskriminatif (LTMPT 2020). Sistem seleksi SNMPTN di IPB menggunakan tiga prinsip dasar yaitu pemerataan kesempatan, akuntabel, dan transparan (IPB 2021). Sejak tahun 2019, proses seleksi SNMPTN di IPB menggunakan media seleksi berbasis web (IPB 2021). Sistem seleksi yang berjalan masih belum menerapkan teknologi yang memudahkan dalam melakukan analisis data seperti pemanfaatan *machine learning* sehingga dibutuhkan sebuah model yang dapat memberikan rekomendasi apakah siswa yang mendaftar layak diterima atau tidak sebagai mahasiswa IPB.

Penelitian tentang studi akademik dan penerimaan mahasiswa baru sudah banyak dilakukan, antara lain AlGhamdi *et al.* (2020) melakukan penelitian terkait dengan pembelajaran *machine learning* untuk memprediksi penerimaan kelulusan mahasiswa yang ingin melanjutkan studi magister ke suatu universitas. Pada penelitian tersebut dibuat pemodelan dengan membandingkan beberapa metode. Model dengan metode Regresi Logistik memiliki hasil akurasi yang lebih baik dibandingkan dengan Regresi Linear dan *Decision Tree*. Penelitian lainnya yang dilakukan Devarapalli (2021) adalah memprediksi peluang masuk mahasiswa pada perguruan tinggi tertentu. Pada penelitian tersebut metode *Random Forest* memiliki nilai akurasi paling baik dibandingkan dengan *Decision Tree*, KNN, dan SVM.

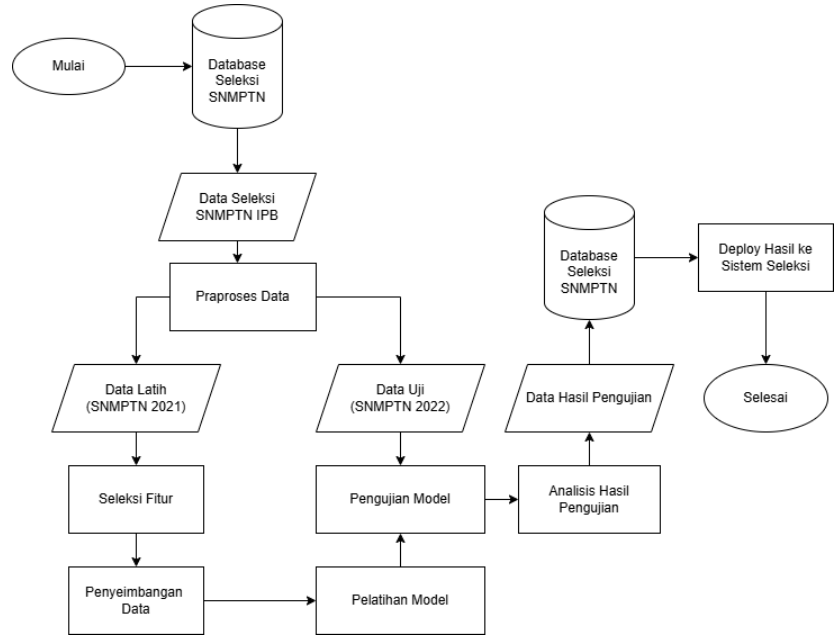
Pada penelitian ini dilakukan pemodelan untuk memprediksi peluang kelulusan calon mahasiswa jalur SNMPTN IPB. Peluang yang didapatkan dari hasil pemodelan akan menjadi rekomendasi atau bahan pertimbangan dalam penentuan kelulusan pelamar. Metode klasifikasi yang digunakan pada penelitian ini adalah Regresi Logistik dan *Random Forest*. Regresi Logistik adalah model regresi yang paling umum untuk memprediksi hasil biner. Hosmer dan Lemeshow (2000) menjelaskan bahwa Regresi Logistik merupakan analisis statistika yang mendeskripsikan hubungan antara variabel respon (kelas target) yang memiliki dua kategori dengan satu atau lebih variabel penjelas (fitur) berskala kategorik atau kontinu. *Random Forest* merupakan metode klasifikasi *ensemble* yang bekerja menerapkan metode *bootstrap aggregating* (*bagging*). Metode ini menumbuhkan banyak pohon sehingga membentuk hutan (*forest*). Analisis dilakukan pada kumpulan pohon yang terbentuk. Pendugaan gabungan (*aggregating*) berdasarkan *ntree* buah pohon menggunakan suara terbanyak (*majority vote*) untuk kasus klasifikasi (Breiman 2001; Kouchaki *et al*. 2020).

Setiap tahun data pelamar yang lulus dan tidak lulus memiliki ketidakseimbangan data karena pelamar yang lulus jauh lebih sedikit dibandingkan dengan jumlah pelamar yang tidak lulus. Sebagian besar algoritma *machine learning* akan menghasilkan model yang baik ketika jumlah datanya seimbang dan akan menjadi masalah ketika data yang tersedia tidak seimbang (James *et al.* 2013). Teknik penyeimbangan data yang dilakukan adalah *Synthetic Minority Oversampling Technique* (SMOTE). SMOTE pertama kali diperkenalkan oleh Nithes V. Chawla (Chawla *et al.* 2002). Pendekatan ini bekerja dengan membuat sintetik data, yaitu replikasi dari data minor. Metode SMOTE bekerja dengan mencari *K-Nearest Neighbors* (ketetanggaan data) untuk setiap data di kelas minor, setelah itu buat sintetik data sebanyak persentase duplikasi yang diinginkan antara data minor dan *K-Nearest Neighbors* yang dipilih secara acak (Baizal *et al.* 2009). Selain itu, dilakukan seleksi fitur yang bertujuan untuk mengurangi kompleksitas model, mempercepat proses pelatihan, menghindari *overfitting*, serta meningkatkan interpretabilitas model dengan hanya mempertimbangkan fitur-fitur yang relevan (Guyon dan Elisseeff 2003). Seleksi fitur yang digunakan adalah *Boruta Feature*

Selection (BFS) yang merupakan algoritma berbasis *wrapper* yang mampu mengidentifikasi semua fitur relevan terhadap target (Kursa dan Rudnicki 2010). BFS juga terbukti efektif dalam menangani data berdimensi tinggi serta data yang tidak seimbang (Kursa 2014). Penelitian ini bertujuan untuk mencari faktor-faktor apa saja yang mempengaruhi kelulusan calon mahasiswa jalur SNMPTN IPB dan mengembangkan pemodelan *machine learning* dengan menggunakan *Random Forest* (RF) dan Regresi Logistik Biner (RLB).

METODE

Penelitian ini dibagi ke dalam beberapa tahapan yang bertujuan untuk membantu peneliti memahami langkah kerja penelitian yang dilakukan. Tahapan penelitian dapat dilihat pada Gambar 1.



Gambar 1 Tahapan penelitian

Data Penelitian

Data penelitian yang digunakan merupakan data seleksi penerimaan mahasiswa baru SNMPTN IPB tahun 2021 dan 2022. Data diperoleh dari Direktorat Administrasi Pendidikan dan Penerimaan Mahasiswa Baru. Data berisikan informasi mengenai data diri pelamar, nilai akademik, prestasi yang diraih, indeks sekolah dan beberapa fitur pendukung lainnya yang akan menjadi pertimbangan dalam seleksi.

Data indeks sekolah atau gerombol sekolah yang menjadi salah satu fitur pada data penelitian ini didapatkan dari penelitian yang dilakukan Dewantari (2021). Pada penelitian tersebut dilakukan penggerombolan sekolah menggunakan metode *Two-Step Cluster*. Hasil dari penelitian tersebut mendapatkan penggerombolan sebanyak 4 gerombol sekolah. Gerombol 1 merupakan sekolah yang memiliki komitmen, konsistensi, dan kualitas yang rendah terhadap IPB. Gerombol 2 merupakan sekolah yang memiliki komitmen yang tinggi terhadap satu jalur namun rendah di jalur yang lain, kualitas sekolah yang tinggi, dan kekonsistenan yang rendah. Gerombol 3 merupakan sekolah yang memiliki komitmen yang cukup rendah, namun tidak lebih rendah dari Gerombol 1, kualitasnya cukup rendah dan konsistensinya juga sangat rendah. Gerombol 4 merupakan sekolah yang unggul pada komitmen, kualitas, dan dapat menjaga konsistensinya selama 5 tahun berturut-turut (Dewantari 2021).

Praproses Data

Pada tahap praproses data, dilakukan beberapa metode untuk mempersiapkan data sebelum pemodelan. Proses yang dilakukan meliputi *integration*, imputasi, dan *encoding*.

Integration merupakan proses integrasi data dari sumber lain untuk memperkaya dataset yang digunakan. Dalam penelitian ini, fitur kategori indeks sekolah ditambahkan ke dalam dataset utama dengan melakukan *lookup* berdasarkan NPSN dari penelitian Dewantari (2021). Imputasi merupakan proses mengganti nilai *missing value* dengan nilai lain atau nilai *constant* (Rubin 1987). Beberapa fitur yang dilakukan proses *imputasi* adalah pelamar beasiswa, prestasi, tingkat prestasi, dan pilihan program studi. Proses *encoding* merupakan praproses data dengan mengubah fitur dengan format teks menjadi angka untuk memudahkan proses pemodelan. Setelah itu data dibagi menjadi data latih dan data uji. Data seleksi SNMPTN 2021 digunakan untuk data latih sedangkan data seleksi SNMPTN 2022 digunakan untuk data uji. Pada data tersebut, dilakukan pemisahan data berdasarkan daerah pelamar yang dibagi menjadi dua yaitu data pelamar Jawa Barat dan Non Jawa Barat. Pemisahan tersebut dilakukan karena distribusi pelamar Jawa Barat melebihi 50% dari keseluruhan distribusi pelamar serta pelamar dari Jawa Barat bisa memilih dua program studi pilihan di IPB (LTMPT 2021).

***Boruta Feature Selection* (BFS)**

BFS merupakan algoritma seleksi fitur dengan teknik *wrapper* yang mampu bekerja dengan metode klasifikasi untuk mengukur tingkat kepentingan fitur. BFS dibangun berdasarkan algoritma *random forest* dalam mengestimasi kepentingan fitur (Kursa *et al.* 2010; Kursa dan Rudnicki 2010). Setelah tahapan praproses data, dilakukan proses seleksi fitur. Seleksi fitur yang dilakukan memberikan hasil bahwa fitur-fitur yang berpengaruh akan ditandai 'Penting' (*Important*), 'Tidak dapat diputuskan' (*Tentative*), dan yang tidak berpengaruh akan ditandai 'Tidak Penting' (*Unimportant*) (Kursa dan Rudnicki 2010). Hasil fitur yang ditandai *Important* dan *Tentative* akan dimasukkan dalam pemodelan.

***Synthetic Minority Oversampling Technique* (SMOTE)**

SMOTE pertama kali diperkenalkan oleh Nithes V. Chawla (Chawla *et al.,* 2002). Pendekatan ini bekerja dengan membuat sintetik data, yaitu replikasi dari data minor. Metode SMOTE bekerja dengan mencari *K-Nearest neighbors* (ketetanggaan data) untuk setiap data di kelas minor, setelah itu buat sintetik data sebanyak persentase duplikasi yang diinginkan antara data minor dan *K-Nearest neighbors* yang dipilih secara acak (Baizal *et al.* 2009). SMOTE termasuk teknik pendekatan *oversampling*. Setelah melakukan praproses data, data latih yang memiliki distribusi kelas tidak seimbang akan dilakukan penyeimbang data menggunakan teknik *oversampling* SMOTE. Penyeimbangan data dilakukan agar distribusi kelas menjadi seimbang sehingga model yang dihasilkan bisa memprediksi data baru dengan baik.

Pemodelan

Ada empat model yang dibangun di tiap wilayah. Model pertama adalah model yang dibangun dari data pelamar menggunakan semua fitur dan tanpa penyeimbangan. Model kedua adalah model yang dibangun dari data pelamar menggunakan semua fitur dan dilakukan penyeimbangan. Model ketiga adalah model yang dibangun dari data pelamar menggunakan fitur seleksi dan dilakukan penyeimbangan data. Model keempat adalah model yang dibangun dari data pelamar dengan pemilihan fitur menggunakan *Adjustment* Pakar (AP) dan dilakukan penyeimbangan. Pemodelan dilakukan menggunakan metode Regresi Logistik Biner dan *Random Forest* dengan 10 *k-fold cross validation* dan *hyperparameter tuning*. Tabel 1 menunjukkan jumlah percobaan model yang dibangun, daftar *hyperparameter tuning*, fitur seleksi dan SMOTE.

Tabel 1 Jumlah percobaan model

Model	*Hyperparameter tuning*	Daftar Nilai	Model	Seleksi Fitur	SMOTE
RF	*ntree*	50, 100, 150, 200, 250, 300, 500, 1000, 1500, 2000	1	-	-
			2	-	Ya
	mtry	F, F 1/2, F 1/3, F 1/4 [b]	3	BFS	Ya
			4	AP [a]	Ya
RLB	*alpha*	seq(0, 1, *by* = 0.1)	1	-	-
			2	-	Ya
	lambda	10^seq(-4, 2, *length* = 50)	3	BFS	Ya
			4	AP [a]	Ya

[a] *Adjusment* Pakar, [b] F = Seluruh Fitur

.

Hyperparameter tuning merupakan proses pemilihan *hyperparameter* yang optimal untuk suatu model *machine learning* (Saini 2019). Proses pemilihan *hyperparameter tuning* pada penelitian ini menggunakan teknik *grid search* (Shekar dan Dagnew 2019). Pada Regresi Logistik Biner (RLB), parameter yang dipakai adalah *alpha* dan *lambda*, sedangkan parameter yang dipakai pada *Random Forest (RF)* adalah *ntree* dan *mtry*.

Pengujian dan Analisis Hasil Pengujian

Setelah beberapa model dibangun, model diuji menggunakan data uji. Data uji yang dipakai adalah data seleksi SNMPTN 2022 yang memiliki distribusi kelas tidak seimbang. Nilai evaluasi yang dicari adalah *Accuracy*, *Precision*, *Recall, dan F1-Score*. Perhitungan nilai tersebut menggunakan Persamaan 1-4 (Han dan Kamber 2006).

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (1)$$

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

$$F1 - Score = 2\,x\frac{Precision\ x\ Recall}{Precision + Recall} \quad (4)$$

Accuracy dalam klasifikasi adalah persentase ketepatan data yang diklasifikasikan secara benar setelah dilakukan pengujian pada hasil klasifikasi (Han dan Kamber 2006). *Precision* digunakan untuk mengukur tingkat akurasi hasil identifikasi kelas positif oleh model. *Recall* adalah matriks evaluasi yang digunakan untuk mengukur performa model dalam mendeteksi satu kelas, khususnya kelas positif (Powers 2011). Selain itu, *F1-Score* dihitung sebagai rata-rata harmonik antara *Precision* dan *Recall*, yang bertujuan untuk memberikan gambaran keseimbangan antara kedua metrik, terutama pada data dengan distribusi kelas yang tidak seimbang (Sokolova dan Lapalme 2009). Fokus utama dalam penelitian ini adalah pada evaluasi nilai *Recall.* Model yang mendapatkan nilai *Recall* tertinggi menjadi model terbaik dan dipilih untuk dilakukan analisis terhadap faktor-faktor yang mempengaruhi kelulusan pelamar (Fernandes *et al*. 2019). Hasil prediksi dari model terbaik akan diintegrasikan ke dalam *database* seleksi karena hasil pengujian tersebut berisi peluang prediksi kelulusan setiap peserta. Nilai peluang yang dihasilkan akan ditampilkan melalui sistem seleksi SNMPTN IPB sebagai bahan pertimbangan dalam proses seleksi.

HASIL DAN PEMBAHASAN

Data Penelitian

Data penelitian terdiri dari data seleksi SNMPTN 2021 dan SNMPTN 2022. Data seleksi SNMPTN 2021 digunakan untuk data latih sedangkan data SNMPTN 2022 digunakan untuk data uji. Data penelitian ini memiliki 39 fitur dan memiliki kelas target lulus dan tidak lulus. Tabel 2 menunjukkan distribusi jumlah data latih dan data uji.

Tabel 2 Distribusi jumlah data latih (2021) dan data uji (2022)

Data	Pembagian Data	Jumlah Data	Kelas Target	Jawa Barat		Non Jawa Barat	
				Jumlah	Persentase	Jumlah	Persentase
SNMPTN 2021	Data Latih	16588	Lulus	937	9.0%	996	16.1%
			Tidak Lulus	9473	91.0%	5182	83.9%
SNMPTN 2022	Data Uji	15876	Lulus	1048	10.3%	993	17.5%
			Tidak Lulus	9150	89.7%	4685	82.5%

Pada Tabel 2 terlihat data pelamar Jawa Barat dan Non Jawa Barat yang memiliki kelas tidak seimbang. Ketidakseimbangan kelas lulus dan tidak lulus terjadi karena kuota penerimaan SNMPTN yang sedikit, sedangkan pelamar SNMPTN yang mendaftar ke IPB mencapai belasan hingga puluhan ribu.

Praproses Data

Proses *integration* pada data latih menghasilkan fitur Indeks Sekolah dengan empat kategori. Selanjutnya dilakukan proses imputasi dan *encoding* secara bersamaan. Fitur seperti pelamar beasiswa, prestasi, tingkat prestasi, pilihan program studi, dan keputusan seleksi (kelas target) memiliki jumlah *missing value* yang signifikan. Pada fitur pelamar beasiswa, terdapat 14052 *missing value* pada data latih dan 13064 pada data uji. Untuk mengatasi hal ini, dilakukan imputasi dengan mengganti nilai-nilai yang hilang dengan nilai konstan. Konversi nilai yang *missing value* diisi dengan nilai yang sudah dilakukan *encoding* ke nilai numerik yaitu 0. Selain nilai *missing value* dilakukan label *encoding* dari nilai kategori menjadi nilai numerik. Fitur jenis kelamin yang awalnya bernilai kategori (L dan P) menjadi nilai *numeric* (0 dan 1) menggunakan teknik *encoding* biner yang umum dilakukan untuk fitur dengan dua kategori. Tabel 3 menunjukkan hasil dari proses imputasi dan *encoding*.

Tabel 3 Hasil praproses imputasi dan e*ncoding*

Fitur	*Missing Value* (MS)		*Result* MS	Variasi Nilai
	Data Latih	Data Uji		
Jenis Kelamin	0	0		0 – 1
Pelamar Beasiswa	14052	13064	0	0 – 1
Prestasi 1	10320	11699	0	0 – 1
Tingkat Prestasi 1	10320	11699	0	0 – 5
Prestasi 2	14187	14281	0	0 – 1
Tingkat Prestasi 2	14187	14281	0	0 – 5
Prestasi 3	16022	15485	0	0 – 1
Tingkat Prestasi 3	16022	15485	0	0 – 5
Pilihan Prodi 1	3514	2845	0	1 – 39
Pilihan Prodi 2	9312	9205	0	1 – 39
Keputusan Seleksi	14655	13835	0	0 – 1

Praproses data ini memastikan bahwa *dataset* yang digunakan sudah bersih dari *missing value* dan telah dikonversi ke format *numeric*, sehingga memudahkan penerapan algoritma *machine learning*. Setelah itu dilakukan pemisahan data berdasarkan daerah pelamar yang dibagi menjadi dua yaitu pelamar data Jawa Barat dan Non Jawa Barat untuk dilakukan pemodelan.

Seleksi Fitur menggunakan BFS

Seleksi fitur Boruta dipakai untuk percobaan model ketiga. Pada data Jawa Barat, didapatkan hasil bahwa sebanyak 34 fitur ditandai *Important*, dua fitur ditandai *Unimportant* dan 3 fitur ditandai *Tentative*. Pada pemrograman R, *library* Boruta menyediakan fungsi *TentativeRoughFix()* yang bertujuan memberikan keputusan untuk fitur-fitur yang ditandai *Tentative*, apakah masuk ke dalam fitur *Important* atau *Unimportant*. Keputusannya adalah terdapat 36 fitur yang ditandai *Important* dan tiga fitur yang ditandai *Unimportant*. Fitur yang *Unimportant* meliputi pindahan, prestasi 1 dan prestasi 2. Pada data Non Jawa Barat, terdapat 33 fitur yang ditandai *Important* dan enam fitur yang ditandai *Unimportant* yaitu pindahan, pelamar beasiswa, prestasi 1, prestasi 3, tingkat prestasi 1 dan tingkat prestasi 3. Secara umum, fitur-fitur yang ditandai *Important* meliputi perguruan tinggi pilihan, kategori indeks sekolah, pilihan prodi, jenis kelamin, mata pelajaran pendukung, rata-rata mata pelajaran pendukung dan fitur pendukung lainnya. Fitur tersebut digunakan untuk mengembangkan model yang optimal pada percobaan model ketiga.

SMOTE

Penyeimbangan data dilakukan menggunakan teknik SMOTE pada skenario pemodelan kedua hingga keempat. Teknik ini digunakan untuk mengatasi ketidakseimbangan antara kelas "lulus" dan "tidak lulus" dalam data pelamar. Tabel 4 menunjukkan distribusi jumlah data latih setelah dilakukan penyeimbangan menggunakan SMOTE.

Tabel 4 Distribusi jumlah data latih (2021) setelah penyeimbangan

Data	Pembagian Data	Jumlah Data	Kelas Target	Jawa Barat	Non Jawa Barat
SNMPTN 2021	Data Latih	30129	Lulus	8433	5976
			Tidak Lulus	9744	5976

Pemodelan

Pemodelan pada penelitian ini menggunakan metode *Random Forest* (RF) dan *Regresi Logistik Biner* (RLB) dengan percobaan empat pemodelan di tiap wilayah dan tiap metode. Pemodelan dilakukan pada dua kelompok data yaitu Jawa Barat dan Non Jawa Barat dengan optimasi *hyperparameter* menggunakan *grid search*. Rentang nilai *hyperparameter* yang digunakan seperti yang terlihat pada Tabel 1.

Model 1 menggunakan semua fitur dengan data yang tidak seimbang. Model 2 menggunakan semua fitur dan dilakukan penyeimbangan data menggunakan SMOTE. Model 3 menggunakan SMOTE dan fitur seleksi BFS. Model 4 menggunakan SMOTE dan pengambilan fitur berdasarkan *Adjustment* Pakar atau pengetahuan para pimpinan seleksi yang berpengalaman selama beberapa tahun mengikuti langsung proses seleksi. Fitur yang digunakan saat percobaan model keempat sebanyak 19 dari 39 total keseluruhan fitur. Hasil pemodelan dapat dilihat pada Tabel 5 (Jawa Barat) dan Tabel 6 (Non Jawa Barat).

Tabel 5 Hasil pemodelan data Jawa Barat

Pemodelan (2021)	Jawa Barat					
	RF			RLB		
	mtry	*ntree*	Akurasi	*Alpha* (α)	*Lambda* (λ)	Akurasi
1	40	500	93.18%	0.9	$1.67x10^{-3}$	92.25%
2	10	500	98.07%	0.6	$1.32x10^{-4}$	86.55%
3	9.25	2000	98.06%	1.0	$1.00x10^{-4}$	86.35%
4	20	2000	97.19%	0.9	$1.68x10^{-3}$	86.04%

Tabel 6 Hasil pemodelan data Non Jawa Barat

Pemodelan (2021)	Non Jawa Barat					
	RF			RLB		
	mtry	*ntree*	Akurasi	*Alpha* (α)	*Lambda* (λ)	Akurasi
1	20	150	86.40%	0.1	$3.91x10^{-3}$	86.85%
2	6.6	1500	96.43%	0.8	$6.87x10^{-3}$	80.41%
3	8.5	2000	96.63%	0.4	$2.22x10^{-3}$	81.00%
4	20	500	95.45%	0.8	$5.18x10^{-3}$	80.10%

Pada data Jawa Barat, hasil pemodelan menunjukkan akurasi terbaik diperoleh pada model 2 menggunakan RF dengan parameter terbaik *mtry* = 10 dan *ntree* = 500, menghasilkan akurasi 98.07%. Pada data Non Jawa Barat, akurasi tertinggi diperoleh oleh model 3 metode RF dengan *mtry* = 8.5 dan *ntree* = 2000 menghasilkan akurasi 96.63%.

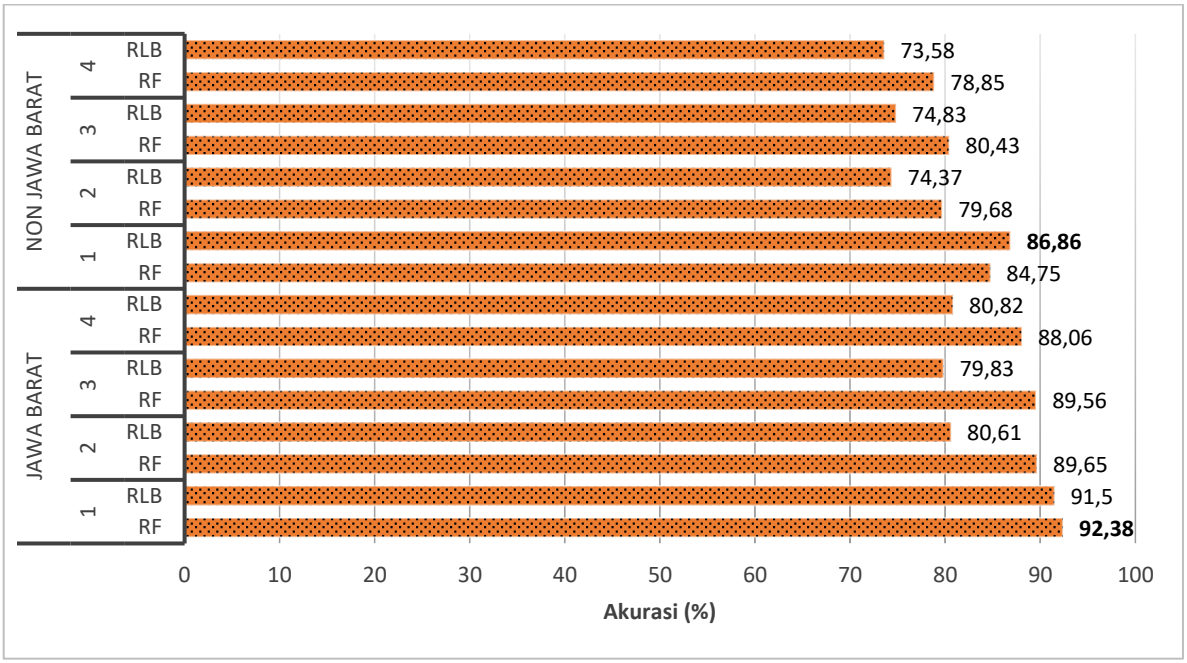
Pengujian dan Analisis Hasil Pengujian

Pengujian dilakukan terhadap empat model yang telah dibangun dengan data uji SNMPTN 2022, baik untuk wilayah Jawa Barat maupun Non Jawa Barat. Setiap model diuji menggunakan matriks evaluasi *Accuracy*, *Precision*, *Recall* dan *F1-Score*. Hasil pengujian dapat dilihat pada Tabel 7.

Tabel 7 Hasil pengujian performa model data Jawa Barat dan Non Jawa Barat

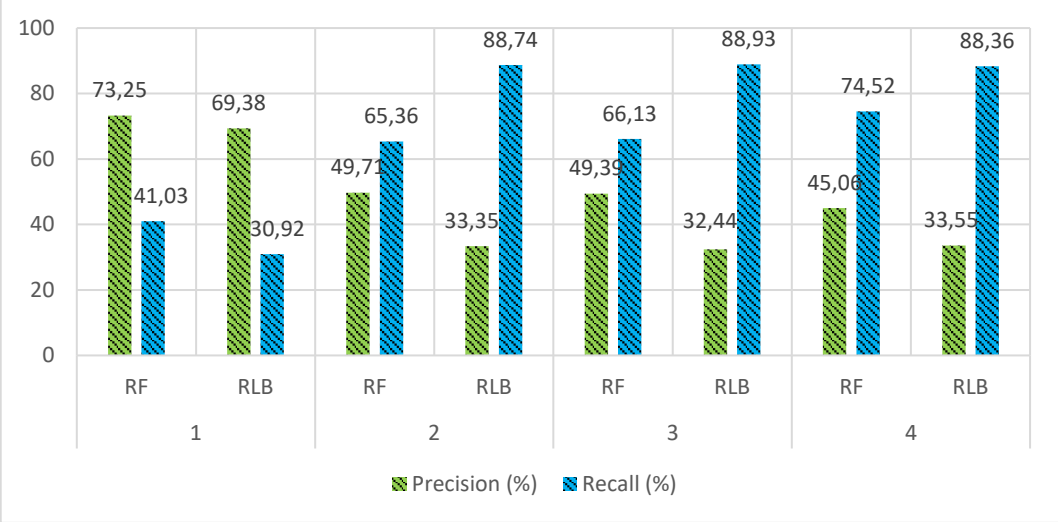
Data	Metode	Model	*Accuracy* (%)	*Precision* (%)	*Recall* (%)	*F1-Score*
Jabar	RF	1	92.38	73.25	41.03	52.69
		2	89.65	49.71	65.36	56.69
		3	89.56	49.39	66.13	56.76
		4	88.06	45.06	74.52	55.89
	RLB	1	91.50	69.38	30.92	42.50
		2	80.61	33.35	88.74	48.41
		3	79.83	32.44	**88.93**	47.62
		4	80.82	33.55	88.36	48.64
Non Jabar	RF	1	84.75	63.71	29.71	40.53
		2	79.68	43.99	59.41	50.87
		3	80.43	45.28	57.10	50.35
		4	78.85	43.3	67.77	53.85
	RLB	1	86.86	72.41	40.18	51.77
		2	74.37	39.44	**86.91**	54.43
		3	74.83	39.74	85.09	53.41
		4	73.58	38.59	86.40	53.17

Berdasarkan nilai evaluasi akurasi, pada data Jawa Barat menunjukkan bahwa Model 1 (RF) menghasilkan nilai akurasi tertinggi yaitu 92.38%. Sedangkan pada data Non Jawa Barat hasil yang diperoleh berbeda. Model 1 (RLB) menunjukkan performa terbaik dengan akurasi 86.86%. Akurasi terendah terdapat pada Model 3 (RLB) yaitu 79.83% pada data Jawa Barat dan Model 4 (RLB) yaitu 73.58% pada data Non Jawa Barat. Berdasarkan hal tersebut, Model 1 menggunakan metode RF atau RLB memiliki nilai akurasi yang tinggi dalam memprediksi data uji dibanding model lainnya yang sudah dilakukan penyeimbangan data. Namun evaluasi akurasi tidak bisa dijadikan dasar bahwa model tersebut bagus dalam memprediksi data uji karena data latih (Model 1) dan data uji memiliki data yang tidak seimbang. Model cenderung bisa mencapai akurasi tinggi hanya dengan memprediksi mayoritas kelas (tidak lulus), namun performanya mungkin tidak sebaik itu saat memprediksi kelas minoritas (lulus). Detail terkait grafik nilai akurasi dapat dilihat pada Gambar 2.



Gambar 2 Grafik perbandingan akurasi tiap model data Jawa Barat dan Non Jawa Barat

Pada matriks evaluasi *Precision*, nilai tertinggi terdapat pada Model 1 (RF) yaitu 73.25% pada data Jawa Barat, sedangkan pada data Non Jawa Barat nilai tertinggi terdapat pada Model 1 (RLB) sebesar 72.41%. Nilai *Precision* terendah terdapat pada Model 3 (RLB) yaitu 32.44% pada data Jawa Barat dan Model 2 (RLB) yaitu 39.44% pada data Non Jawa Barat. *Precision* yang tinggi berarti model berhasil meminimalkan kesalahan prediksi lulus (*minimizing false positive*). Detail grafik perbandingan *Precision* dan *Recall* pada data Jawa Barat dan Non Jawa Barat dapat dilihat pada Gambar 3 dan Gambar 4.

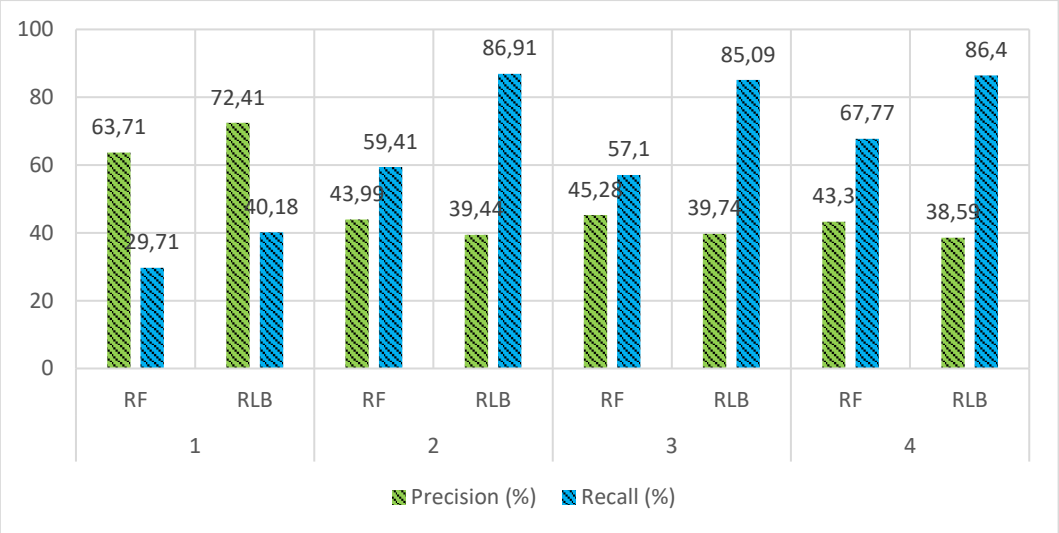


Gambar 3 Grafik perbandingan *precision* dan *recall* data Jawa Barat

Nilai *Recall* tertinggi pada data Jawa Barat terdapat pada Model 3 (RLB) sebesar 88.93%, diikuti Model 2 (RLB) 88.74% dan Model 4 (RLB) 88.36%. Pada data Non Jawa Barat, Model dua (RLB) mendapatkan nilai *Recall* tertinggi yaitu 86.91%, diikuti Model 4 (RLB) 86.40% dan Model 3 (RLB) 85.09%. *Recall* yang tinggi memastikan bahwa mayoritas pelamar yang benar-benar lulus dapat dideteksi oleh model. Nilai *Recall* terendah terdapat pada Model 1 (RLB) yaitu 30.92% pada data Jawa Barat dan Model 1 (RF) yaitu 29.71% pada data Non Jawa Barat.

Pada data Jawa Barat, nilai *F1-score* tertinggi diperoleh oleh Model 3 (RF) sebesar 56.76%, diikuti oleh Model 2 (RF) sebesar 56.69%, dan Model 4 (RF) sebesar 55.89%. Keempat model berbasis RF menghasilkan performa *F1-score* yang lebih tinggi dibanding model berbasis RLB, yang tertinggi hanya mencapai 48.64% pada Model 4 (RLB). Sementara itu, pada data Non Jawa Barat, model dengan *F1-score* tertinggi adalah Model dua (RLB) sebesar 54.43%, diikuti oleh Model 4 (RF) sebesar 53.85%, dan Model 3 (RLB) sebesar 53.41%. Berbeda dengan data Jawa Barat, model berbasis RLB justru menunjukkan performa yang relatif lebih baik dibandingkan RF dalam konteks *F1-score* pada data Non Jawa Barat. Perbedaan model terbaik antara data Jawa Barat dan Non Jawa Barat memberikan gambaran bahwa karakteristik dan distribusi data antar wilayah memiliki pengaruh terhadap performa model.

Data uji yang tidak seimbang dipengaruhi oleh tingkat kompetitif yang berbeda setiap tahunnya, variasi data yang besar antar sekolah, serta distribusi pelamar yang tidak seimbang antara Jawa Barat dan Non Jawa Barat. Dalam penelitian ini, fokus utama dalam melakukan pemodelan dan pengujian adalah memaksimalkan deteksi pelamar yang lulus (kelas minoritas), sehingga nilai *Recall* yang tinggi menjadi acuan dalam pengambilan model terbaik. Terlihat pada grafik *Precision* dan *Recall* pada Gambar 3 dan Gambar 4, model yang memiliki nilai *Recall* yang tinggi dan *Precision* yang rendah adalah model yang sudah dilakukan penyeimbangan data menggunakan SMOTE. Penyeimbangan data terlihat sangat mempengaruhi hasil pemodelan dan pengujian di mana model mampu lebih banyak memprediksi data yang benar-benar lulus. Pada data Jabar, Model 3 dengan metode RLB yang menggunakan fitur seleksi BFS dan penyeimbangan data (SMOTE) mampu memperoleh nilai *Recall* yang lebih baik dibandingkan dengan model lainnya. Pada data Non Jawa Barat, Model 2 dengan metode RLB yang menggunakan semua fitur memperoleh nilai *Recall* yang lebih baik dibandingkan dengan model lainnya. Berdasarkan informasi tersebut penyeimbangan data menggunakan SMOTE dengan metode RLB dapat meningkatkan performa model ketika memprediksi data kelas lulus (minoritas).



Gambar 4 Grafik perbandingan *precision* dan *recall* data Non Jawa Barat

Analisis Faktor-Faktor yang Mempengaruhi Kelulusan

Analisis terhadap faktor-faktor atau fitur yang berpengaruh pada model prediksi kelulusan pelamar dilakukan dengan menggunakan fungsi *variable importance* (varImp) dalam pemrograman R. Setelah pemodelan dan pengujian selesai, model dengan nilai *recall* tertinggi yaitu Model 3 (RLB) untuk data Jawa Barat dan Model 2 (RLB) untuk data Non Jawa Barat digunakan untuk mengidentifikasi faktor-faktor yang paling signifikan. Tabel 8 merangkum nilai *Importance* dari fitur-fitur yang mempengaruhi kelulusan pelamar Jabar dan NonJabar.

Dari hasil analisis, Fitur perguruan tinggi pilihan memiliki nilai *Importance* tertinggi baik untuk pelamar dari Jawa Barat maupun Non Jawa Barat, menandakan fitur ini sebagai faktor yang paling berpengaruh dalam menentukan kelulusan. Hasil menunjukkan bahwa IPB lebih cenderung meluluskan pelamar yang menjadikan IPB sebagai pilihan pertama, sehingga fitur ini memberikan dampak besar dalam prediksi kelulusan.

Tabel 8 Fitur-fitur yang mempengaruhi kelulusan

Jabar		NonJabar	
Fitur	*Importance*	Fitur	*Importance*
Perguruan Tinggi Pilihan	100.00	Perguruan Tinggi Pilihan	100.00
Kategori Indeks Sekolah	43.41	Pilihan Prodi 1	80.93
Pilihan Prodi 2	43.36	Kategori Indeks Sekolah	62.22
Pilihan Prodi 1	25.65	Jenis Kelamin	11.75
Tingkat Prestasi 3	22.54	Prestasi 2	6.91
Jenis Kelamin	15.14	Pilihan Prodi 2	4.40
Tingkat Prestasi 2	11.52	Rata-rata mata pelajaran pendukung	4.08
Tingkat Prestasi 1	7.90	Pelamar Beasiswa	3.62
Rata-rata mata pelajaran pendukung	4.88	Tingkat Prestasi 2	2.13
Pelamar Beasiswa	3.24	Tingkat Prestasi 1	2.02

Fitur kategori indeks sekolah berada pada urutan kedua untuk pelamar Jawa Barat dengan nilai *Importance* 43.41 dan urutan ketiga untuk Non Jawa Barat dengan nilai 62.22. Kategori indeks sekolah memiliki peran penting karena pelamar dari sekolah yang memiliki indeks sekolah yang baik umumnya lebih disukai, karena performa akademik mereka dianggap berpotensi tinggi selama menempuh pendidikan di IPB. Selanjutnya fitur pilihan program studi juga menunjukkan pengaruh yang signifikan terhadap kelulusan. Ketika pelamar memilih program studi favorit, tingkat persaingan menjadi lebih ketat, sehingga peluang kelulusan lebih rendah, sementara program studi dengan peminat yang lebih rendah memberikan peluang kelulusan yang lebih besar. Fitur lain seperti jenis kelamin, tingkat prestasi, rata-rata mata pelajaran pendukung dan prestasi juga memiliki kontribusi terhadap kelulusan.

SIMPULAN

Penelitian ini telah berhasil mengimplementasikan dua metode *machine learning*, yaitu *Random Forest* (RF) dan Regresi Logistik Biner (RLB), untuk memprediksi kelulusan pelamar SNMPTN di IPB. Model 3 (RLB) pada data Jabar dan Model 2 (RLB) pada data NonJabar menunjukkan performa terbaik. Penyeimbangan data menggunakan SMOTE terbukti efektif dalam meningkatkan kemampuan model untuk mendeteksi pelamar yang benar-benar lulus, yang terlihat dari peningkatan nilai *Recall*. Metrik ini sangat penting dalam konteks penelitian karena kemampuan mendeteksi pelamar yang layak lulus menjadi fokus utama analisis. Hasil pemodelan juga menunjukkan bahwa beberapa fitur seperti urutan pemilihan perguruan tinggi, peringkat akademik sekolah, prestasi, dan pilihan program studi memiliki pengaruh signifikan terhadap prediksi kelulusan pelamar. Secara keseluruhan, penelitian ini menunjukkan bahwa model *machine learning* dapat diimplementasikan dalam sistem seleksi SNMPTN di IPB untuk mendukung pengambilan keputusan.

DAFTAR PUSTAKA

AlGhamdi A, Barsheed A, AlGhamdi H, AlMshjary H. 2020. A machine learning approach for graduate admission prediction. Proceedings of the 2020 2nd International Conference on Image, Video and Signal Processing. Hlm 155-158; [diakses 2022 Januari 02].

Baizal ZA, Bijaksana MA, Sastrawan AS. 2009. Analisis Pengaruh Metode Over Sampling Dalam Churn Prediction untuk Perusahaan Telekomunikasi. Seminar Nasional Aplikasi Teknologi Informasi 2009 (SNATI 2009). ISSN: 1907-5022.

Breiman L. 2001. Random forests. *Machine Learning*. 45:5–32.

Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. 2002. SMOTE : Synthetic minority oversampling technique. *Journal of Artificial Intelligence Research* 16.

Devarapalli DJ. 2021. Classification Method to Predict Chances of Students' Admission in a Particular College. Proceedings of International Conference on Recent Trends in Machine Learning, IoT, Smart Cities and Applications. Hlm 225-238; [diakses 2024 Maret 22].

Dewantari NKM. 2021. Penggerombolan Sekolah pada Penerimaan Mahasiswa Baru Jalur SNMPTN di IPB menggunakan Metode Two-Step Cluster [skripsi]. Bogor (ID): Institut Pertanian Bogor.

Fernandes E, Holanda M, Victorino M, Borges V, Carvalho R, Van Erven G. 2019. Educational data mining: Predictive analysis of academic performance of public school students in the capital of Brazil. *Journal of Business Research*. 94:335–343. DOI: 10.1016/j.jbusres.2018.02.012.

Guyon I, Elisseeff A. 2003. An introduction to variable and feature selection. *Journal of Machine Learning Research*. 3(Mar):1157–1182.

Han J, Kamber M. 2006. *Data Mining: Concepts and Techniques, 2nd ed*. San Francisco (US): Morgan Kaufmann.

Hosmer DW, Lemeshow S. 2000. *Applied Logistic Regression*. *Ed ke-2*. New York(US): John Wiley and Sons.

[IPB] Institut Pertanian Bogor. 2021. Laporan SNMPTN IPB. Bogor: IPB.

James G, Witten D, Hastie T, Tibshirani R. 2013. *An Introduction to Statistical Learning with Applications in R. Ed. ke-7*. New York(US): Springer. DOI:10.1007/978-1-4614-7138-7.

Kouchaki S, Yang Y, Lachapelle A, Walker TM, Walker AS, Peto TEA, Crook DW, Clifton DA. 2020. Multi-label random forest model for tuberculosis drug resistance classification and mutation ranking. *Front. Microbiol*. DOI:10.3389/fmicb.2020.00667.

Kursa MB, Jankowski A, Rudnicki WR. 2010. Boruta – a system for feature selection. *Fundamenta Informaticae*. 101(4):271-285. DOI:10.3233/FI-2010-288.

Kursa MB, Rudnicki WR. 2010. Feature selection with the boruta package. *Journal of Statistical Software.* 36(11):1-13. DOI:10.18637/jss.v036.i11.

Kursa MB. 2014. Robustness of random forest-based gene selection methods. *BMC Bioinformatics*. 15(1):8. DOI:10.1186/s12859-014-0155-9.

[LTMPT] Lembaga Tes Masuk Perguruan Tinggi. 2020. Informasi SNMPTN 2020. Jakarta (ID): LTMPT.

[LTMPT] Lembaga Tes Masuk Perguruan Tinggi. 2021. Informasi Umum LTMPT. [diakses 2022 Agustus 1]. https://ltmpt.ac.id/?mid=7.

[Permen] Peraturan Menteri Riset, Teknologi, dan Pendidikan Tinggi Republik Indonesia Nomor 60 Tahun 2018 Tentang Penerimaan Mahasiswa Baru Program Sarjana pada Perguruan Tinggi Negeri. 2018.

Powers DMW 2011. Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation. Journal of Machine Learning Technologies. 2(1):37-63.

[RJPIPB] Rencana Jangka Panjang Institut Pertanian Bogor Periode 2019-2045. 2017. Bogor (ID): IPB.

Rubin DB. 1987. *Multiple Imputasi for Nonresponse in Surveys*. New York(US): Wiley.

Saini N. 2019. Multi-label Image Classification to Detect Air Traffic Controllers' Drowsiness Using Facial Features. MSc Res. Proj. Data Anal.

Shekar BHS, Dagnew G. 2019. Grid search-based hyperparameter tuning and classification of microarray cancer data. *International Journal of Data Science and Analysis*. 7(3):145-155. DOI:10.1016/j.ijdsa.2023.06.009.

Sokolova M, Lapalme G. 2009. A systematic analysis of performance measures for classification tasks. *Information Processing & Management*. 45(4):427–437. DOI:10.1016/j.ipm.2009.03.002.