Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK) DOI: 10.25126/jtiik.202294311
Vol. 9, No. 4, Agustus 2022, hlm. 725-734 p-ISSN: 2355-7699
Akreditasi KEMENRISTEKDIKTI, No. 36/E/KPT/2019 e-ISSN: 2528-6579

MODEL KLASIFIKASI CALON MAHASISWA BARU UNTUK SISTEM REKOMENDASI PROGRAM STUDI SARJANA BERBASIS *MACHINE LEARNING*

Ahmad R. Pratama[*1], Rio Rizky Aryanto[2], Arif Taufiq M. Pratama[3]

[1,2,3]Universitas Islam Indonesia, Kabupaten Sleman
Email: [1]ahmad.rafie@uii.ac.id, [2]rio.aryanto@students.uii.ac.id, [3]arif.pratama@students.uii.ac.id
*Penulis Korespondensi

(Naskah masuk: 08 November 2020, diterima untuk diterbitkan: 15 Agustus 2022)

Abstrak

Proses pemilihan program studi bagi calon mahasiswa baru, khususnya bagi mereka yang masih duduk di bangku SMA atau sederajat, merupakan salah satu momen pengambilan keputusan penting. Tak jarang pilihan yang salah berujung pada kegagalan studi atau kesulitan lain selepas menamatkan studi. Meski sudah mulai marak dilakukan di berbagai negara maju, sistem rekomendasi program studi berbasis *machine learning* untuk calon mahasiswa baru masih belum banyak dikembangkan di Indonesia. Penelitian ini dilakukan sebagai upaya rintisan sistem rekomendasi tersebut dengan menggunakan data pribadi dan akademik dari semua mahasiswa dan alumni program sarjana di Universitas Islam Indonesia (UII), di mana data prestasi akademik di masing-masing program studi digunakan sebagai *ground truth label*. Dari hasil penelitian ini, didapatkan sebuah model berbasis *Random Forest* (RF) dengan tingkat akurasi 86%, presisi 84%, recall 86%, dan AUC 97%. Model ini memiliki kinerja yang jauh lebih baik jika dibandingkan dengan model berbasis *Multinomial Logistic Regression* (MLR) maupun *Support Vector Machine* (SVM). Sesuai peta jalan penelitian, model yang dihasilkan dari penelitian ini akan digunakan untuk pengembangan sistem rekomendasi yang dapat membantu calon mahasiswa baru dalam memilih program studi saat proses penerimaan mahasiswa baru (PMB), khususnya di lingkungan UII.

Kata kunci: *rekomendasi, program studi, machine learning, random forest*

A CLASSIFICATION MODEL OF PROSPECTIVE STUDENTS FOR A MACHINE LEARNING-BASED COLLEGE MAJOR RECOMMENDATION SYSTEM

Abstract

Choosing a major for the prospective undergraduate students is one of the most important moments in their life, especially for the high school graduates. Not infrequently, a wrong choice can lead to academic failure or even other difficulties after graduating from college. While a machine learning-based college major recommendation system is not strange in some developed countries, it is not the case in Indonesia. This study aims to serve as a pilot project for such a recommendation system by using personal and academic data of all students and alumni of the undergraduate programs in Universitas Islam Indonesia (UII) where academic achievement data is used as the ground truth label. Out of three models used and evaluated in this study, we found that Random Forest-based model to be the best option with an accuracy of 86%, precision on 84%, recall of 86%, and AUC of 97%. We also found that this model has a much better performance than other models with Multinomial Logistic Regression (MLR) or Support Vector Machine (SVM). The resulting model from this study will be deployed to develop a college major recommendation system that can help the prospective students choose their majors during college admission process, particularly in the context of UII as per research roadmap.

Keywords: *recommendation, undergraduate major, machine learning, random forest*

1. PENDAHULUAN

Proses pemilihan program studi bagi calon mahasiswa baru di jenjang sarjana, terlebih bagi mereka yang masih duduk di bangku Sekolah Menengah Atas (SMA) merupakan salah satu momen pengambilan keputusan penting dalam hidup seseorang. Pada proses perpindahan ke jenjang pendidikan tinggi ini lah untuk pertama kalinya siswa dihadapkan pada pilihan prodi bahkan sejak sebelum memulai proses pendidikan di jenjang barunya tersebut. Dalam banyak kasus, pilihan ini berperan besar dalam menentukan masa depan individu yang menjalaninya, baik di sisi kenyamanan dan kecocokan selama menjalani masa studi, peluang

keberhasilan menyelesaikan studi, hingga pilihan karir selepas berhasil menamatkan studi. Singkat kata, pilihan program studi ini menentukan banyak hal terkait dengan peluang dan risiko yang muncul dari pilihan yang telah diambil dan akan dijalani.

Faktanya, masih banyak hal yang belum diketahui terkait dengan proses pemilihan program studi oleh calon mahasiswa baru. Seperti yang ditunjukkan sebuah penelitian longitudinal, di sisi mahasiswa pun tak jarang terjadi pergeseran keyakinan akan program studi yang telah dipilih sebelumnya, terutama ketika mahasiswa yang bersangkutan menyadari bahwa kemampuannya untuk mengikuti pembelajaran di bidang tertentu, dalam hal ini matematika dan ilmu alam, lebih rendah daripada yang mereka bayangkan sebelumnya (Stinebrickner dan Stinebrickner, 2011).

Permasalahan terkait pemilihan program studi ini juga terlihat dari hasil survei yang dilakukan oleh Indonesia Career Center Network (ICNN) pada tahun 2017. Dilansir dari situs berita Jawa Pos National Network (JPNN) pada tanggal 7 Februari 2019 yang mengutip hasil survei ICNN, disebutkan bahwa terdapat 87% mahasiswa di Indonesia merasa telah mengambil program studi yang tidak tepat. Proporsi yang cukup besar tersebut menunjukkan bahwa kesalahan pemilihan program studi bagi mahasiswa Indonesia masih sering terjadi. Tantangan tersebut tentu perlu mendapatkan perhatian lebih mengingat urgensi dari pemilihan program studi terkait kesuksesan menamatkan studi di perguruan tinggi, bahkan juga dapat mempengaruhi jenjang karier selepas tamat studi.

Terkait dengan hal ini, telah ada beberapa penelitian menggunakan berbagai macam pendekatan yang berbeda untuk memberikan solusi, semisal dengan model eksperimental (Wiswall dan Zafar, 2015) atau pemodelan yang didorong oleh data *(data-driven)* dalam mengidentifikasi faktor penentu pengambilan keputusan dalam pemilihan program studi. Singkat kata, pemanfaatan sains data dalam rangka proses pemilihan program studi di sisi calon mahasiswa baru atau proses PMB di sisi PT sudah mulai marak dilakukan di negara maju seperti Amerika Serikat (Picciano, 2012; Waters dan Miikkulainen, 2014; Liu dan Tan, 2020), ataupun negara lain seperti RRC (Wang dan Shi, 2016) dan Arab Saudi (Khanam dan Alkhaldi, 2019), namun masih cukup asing di Indonesia. Mengingat tiap negara memiliki ciri khas budayanya masing-masing, maka hasil dari satu negara tidak bisa otomatis diaplikasikan begitu saja ke negara lain.

Sementara itu, penelitian terkait pemilihan program studi di Indonesia banyak berkutat pada sistem pakar berbasis aturan *(rule-based)* baik yang berupa aturan asosiasi *(association rule)* yang relatif sederhana (Rumaisa, 2012), hingga yang lebih kompleks dengan menggunakan inferensi fuzzy (Sam'an, 2015; Rozi dan Purnomo, 2018), *naïve Bayes* (Suryadi, 2018) atau metode *Preference ranking organization method for enrichment evaluation (Promethee)* (Faizal, 2015; Kumala dkk., 2015). Meski tidak bisa dinafikan bahwa penelitian-penelitian tersebut juga memiliki kontribusinya tersendiri, namun masih ada ruang besar untuk perbaikan dan peningkatan dari sisi metode yang masih belum memanfaatkan teknologi *machine learning* dalam proses pemberian rekomendasi tersebut, dan terlebih lagi dari sisi kualitas hasil rekomendasi yang diberikan.

Kesenjangan inilah yang akan diisi oleh penelitian ini. Penelitian ini dilakukan dalam rangka inisiasi pemanfaatan sains data dalam proses penerimaan mahasiswa baru (PMB) di dunia pendidikan tinggi di Indonesia pada umumnya, dan di Universitas Islam Indonesia (UII) pada khususnya. Sumber data pada penelitian ini berasal dari data terkait calon mahasiswa baru, mahasiswa aktif, hingga lulusan di semua prodi di UII. Dalam rangka keamanan dan privasi, data yang digunakan berupa agregasi yang telah disinonimkan sebelumnya

2. METODE PENELITIAN

Penelitian menggunakan data internal UII. *Dataset* tersebut berisi informasi terkait mahasiswa baik selama duduk di bangku universitas maupun sekolah menengah atas (SMA). Beberapa informasi tersebut yaitu program studi mahasiswa, jumlah satuan kredit semester (SKS), indeks prestasi kumulatif (IPK) dan status mahasiswa pada saat data diambil. Selanjutnya terdapat juga informasi dari mahasiswa ketika duduk di bangku SMA seperti jenis sekolah dan jurusan yang diambil, nilai akademik pada mata pelajaran selama enam semester, total nilai rapor per semester dan total nilai ujian nasional. Informasi yang lebih umum juga terdapat pada *dataset*, seperti contohnya jenis kelamin, hobi, prestasi, tempat tanggal lahir dan informasi wali murid atau orang tua.

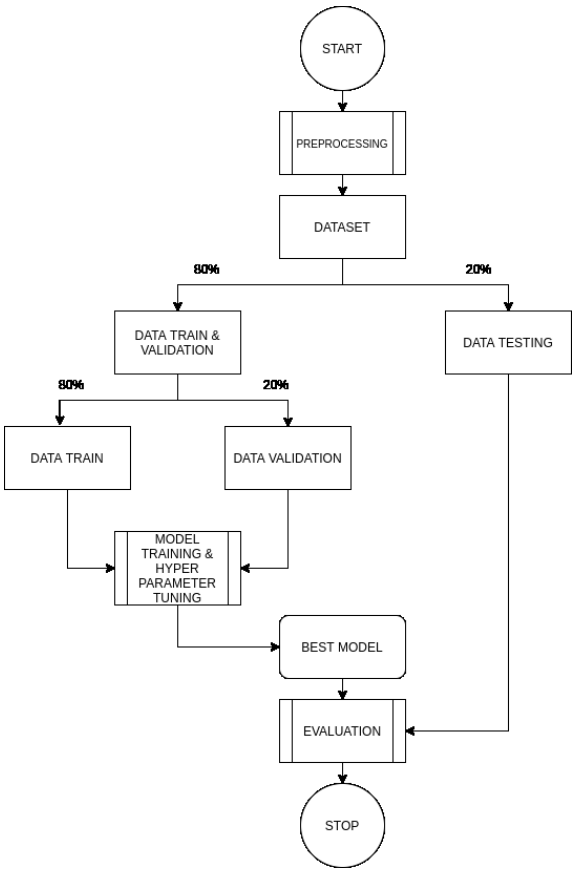
Sistem rekomendasi yang dikembangkan bertujuan untuk membantu calon mahasiswa mengetahui program studi yang cocok dilihat baik dari sisi akademik maupun non-akademik. Guna mendukung tujuan tersebut peneliti memilih beberapa atribut yang sekiranya mampu merepresentasikan lingkup permasalahan. Atribut tersebut terdiri dari variabel kategorikal maupun numerik seperti yang disajikan pada Tabel 1.

Tabel 1. Jenis Variabel

Jenis Variable	Contoh Variabel
Kategorikal	Jenis kelamin, hobi, jenis sekolah, jurusan SMA
Numerikal	nilai rata-rata dari setiap mata pelajaran seperti matematika, bahasa Indonesia, bahasa Inggris, biologi, fisika, kimia, geografi, sejarah, ekonomi, agama dan kompetensi keahlian/kejuruan

Secara garis besar terdapat empat tahapan utama dalam pengembangan model klasifikasi yang akan

digunakan sebagai algoritma untuk sistem rekomendasi program studi ini, yaitu *preprocessing*, *training & validation*, *testing,* dan *evaluation*. Gambar 1 menunjukkan diagram alir yang digunakan pada penelitian ini beserta beberapa detail informasi untuk masing-masing tahapan tersebut.



Gambar 1. Tahapan penelitian

2.1. *Preprocessing*

Tahapan ini dilakukan untuk menyiapkan *dataset* sebelum diimplementasikan pada model klasifikasi. Beberapa teknik seperti agregasi dan seleksi (*filtering*) dilakukan pada tahapan *preprocessing*. Teknik agregasi dilakukan untuk mendapatkan nilai rerata dari setiap mata pelajaran pada tiap individu. Peneliti juga membagi jurusan SMA ke dalam tiga kelompok yaitu Ilmu Pengetahuan Alam (IPA), Ilmu Pengetahuan Sosial (IPS), dan Non-IPA-IPS. Hal tersebut dilakukan karena pada *dataset* terdapat lebih dari sepuluh jenis jurusan SMA sehingga perlu pengelompokan ke dalam kategori yang lebih umum.

Selanjutkan melakukan seleksi atau *filtering dataset* yang dilakukan berdasarkan jumlah SKS dan nilai IPK mahasiswa. Nantinya, peneliti hanya menggunakan data mahasiswa yang mempunyai jumlah SKS minimal 80 dan IPK minimal 3.00. Peneliti berasumsi bahwa mahasiswa dengan karakteristik tersebut adalah mahasiswa yang dinilai cukup berhasil menjalankan studi pada jangka waktu yang relatif lama. Mahasiswa pada kelompok tersebut dinilai mempunyai kecocokan yang lebih tinggi terhadap program studi dibandingkan mahasiswa-mahasiswa lainnya. Tidak hanya berdasarkan jumlah SKS dan IPK, status mahasiswa juga diperhatikan pada saat seleksi *dataset*. Peneliti hanya menggunakan data mahasiswa dengan status yang sudah lulus maupun yang masih aktif. Sebaliknya, mahasiswa dengan status selain itu tidak akan dimasukkan ke dalam *dataset*, tidak peduli seberapa baik nilai IPK maupun jumlah SKS yang pernah diambil.

Seleksi *dataset* juga dilakukan berdasarkan jurusan SMA dan nilai rerata pada mata pelajaran tertentu. Bagi individu dengan jurusan SMA IPA, maka individu tersebut haruslah mempunyai nilai rerata pada beberapa mata pelajaran seperti matematika, bahasa Indonesia, bahasa Inggris, biologi, kimia, fisika dan agama agar dapat diikutsertakan pada data latih model. Untuk jurusan IPS, yang menjadi pertimbangan adalah mata pelajaran matematika, bahasa Indonesia, bahasa Inggris, geografi, sejarah, ekonomi dan agama sedangkan untuk jurusan Non-IPA-IPS yang menjadi pertimbangan hanyalah nilai rerata pada mata pelajaran kompetensi keahlian/kejuruan.

Terakhir, dilakukan seleksi berdasarkan jenis program studi. Pada penelitian ini, program studi yang digunakan hanya melibatkan program studi sarjana reguler yang merupakan mayoritas dari program sarjana yang ada di UII. Program studi dengan jenjang berbeda, seperti diploma dan pascasarjana, serta program internasional tidak disertakan dalam penelitian ini dikarenakan keterbatasan data yang akan sangat mempengaruhi hasil klasifikasi yang akan didapatkan. Setelah melakukan *preprocessing*, didapatkan total 24 program studi jenjang sarjana pada *dataset* (label 0 s.d label 23). Tabel 2 menunjukkan program studi dan jumlah kemunculannya pada *dataset* diurutkan dari kemunculan terbanyak sampai dengan yang paling sedikit, sementara contoh data yang telah melalui proses *preprocessing* dapat dilihat pada Gambar 2.

Tabel 2. Jumlah kemunculan masing-masing prodi pada *dataset*

Prodi	n	Prodi	n
Akuntansi	265	Ilmu Komunikasi	78
Manajemen	244	Informatika	76
Teknik Kimia	139	Teknik Sipil	65
Psikologi	123	Perbankan dan Keuangan	42
Ekonomi Pembangunan	116	Ekonomi Islam	40
Hukum	110	Pendidikan Agama Islam	30
Teknik Lingkungan	99	Teknik Elektro	29
Statistika	93	Ahwal Al-Syakhshiyah	23
Teknik Industri	83	Pendidikan Bahasa Inggris	17
Farmasi	83	Teknik Mesin	17
Kimia	82	Hubungan Internasional	16
Arsitektur	79	Kedokteran	7

JNS_KELAMIN_MHS	JENIS_SMTA	JURUSAN_SMTA	HOBI3	Matematika	Bahasa Indonesia	Bahasa Inggris	Biologi	Fisika	Kimia	Geografi	Sejarah	Ekonomi	Agama	Kompetensi Keahlian/ Kejuruan	NAMA_PRODI
L	SMU/SMA	IPA	SENI LUKIS	77.2	77.6	80.6	78.2	73.6	75.4	0.0	0.0	0.0	84.0	0.0	Arsitektur
P	SMU/SMA	IPA	TARI KREASI BARU	33.8	35.4	34.2	34.6	33.4	35.0	0.0	0.0	0.0	34.6	0.0	Farmasi
P	SMU/SMA	IPS	NYANYI TUNGGAL	14.32	14.4	13.72	0.0	0.0	0.0	14.720000000	14.080000000	13.919999999	14.0	0.0	Akuntansi
L	SMU/SMA	IPA	MTQ	39.0	38.0	37.6	38.2	37.8	35.6	0.0	0.0	0.0	38.0	0.0	Teknik Sipil
L	SMU/SMA	IPA	PADUAN SUARA	82.4	87.0	88.2	85.0	82.4	85.6	0.0	0.0	0.0	87.6	0.0	Teknik Mesin
P	SMU/SMA	IPA	TARI KREASI BARU	89.4	92.4	91.2	93.4	92.2	93.4	0.0	0.0	0.0	92.8	0.0	Akuntansi
P	Lainnya	IPS	FUTSAL	34.2	37.0	35.2	0.0	0.0	0.0	34.8	36.2	33.8	35.6	0.0	Hukum
P	SMU/SMA	IPS	MARCHING BAND	34.4	34.0	34.2	0.0	0.0	0.0	35.6	34.0	33.4	32.2	0.0	Ilmu Komunikasi
P	SMU/SMA	IPA	RENANG	36.6	36.6	35.2	35.4	36.4	36.2	0.0	0.0	0.0	36.0	0.0	Akuntansi
L	SMU/SMA	IPA	RENANG	88.0	88.0	89.5	88.0	83.5	84.0	0.0	0.0	0.0	90.0	0.0	Teknik Mesin
L	SMU/SMA	IPA	MENEMBAK	75.8	81.6	81.2	79.6	77.0	77.6	0.0	0.0	0.0	86.8	0.0	Ilmu Komunikasi
P	SMU/SMA	IPA	TARI TRADISIONAL	13.62	13.36	13.760000000000	13.36	13.439999999	12.919999999	0.0	0.0	0.0	6.92	0.0	Akuntansi
L	SMU/SMA	IPA	FUTSAL	34.67999999999999	32.739999999999995	32.2	33.04	35.32	34.760000000	0.0	0.0	0.0	33.92	0.0	Kedokteran
L	Lainnya	NON IPA-IPS	SENI LUKIS	90.4	85.4	90.6	0.0	0.0	89.8	0.0	0.0	0.0	89.0	93.8	Teknik Mesin
P	SMU/SMA	IPS	TEATER	34.2	34.8	34.2	0.0	0.0	0.0	36.4	37.8	33.6	33.0	0.0	Ahwal Al-Syakhshiyah

Gambar 2. Contoh data penelitian setelah melalui tahapan *preprocessing*

2.2. *Training & Validation*

Berdasarkan hasil pada tahapan *preprocessing*, peneliti menemukan adanya ketidakseimbangan jumlah program studi pada *dataset* (*imbalanced dataset*). Ketimpangan jumlah data terlihat pada beberapa program studi seperti Teknik Elektro, Pendidikan Bahasa Inggris, Teknik Mesin, Hubungan Internasional dan Kedokteran. Hal tersebut dapat mempengaruhi performa model klasifikasi yang akan dikembangkan sehingga peneliti memutuskan untuk memperbaiki hal tersebut dengan teknik *oversampling*.

Teknik *oversampling* adalah suatu teknik yang bertujuan untuk menghasilkan *(generate)* data baru pada kelompok program studi dengan jumlah kemunculan yang lebih kecil dibandingkan kelompok lainnya. Proses *generate* data baru tersebut dilakukan dengan cara pengambilan acak atau *random sampling* dari *dataset* asli. Artinya, proses ini hanya akan menambah jumlah *records* data namun tidak mengubah karakteristik kelompok tersebut. Melalui *oversampling* dengan *Synthetic Minority Over-Sampling Method* (SMOTE) yang mensintesis data baru pada kelas minoritas dengan karakteristik menyerupai data aslinya, *dataset* yang sebelumnya berjumlah total 1,956 records data berubah menjadi total 6,360 *records* data.

Selain *oversampling*, pada proses pembelajaran model juga dilakukan standarisasi nilai pada variabel numerik. Hal ini untuk menyamakan satuan unit pada masing-masing variabel numerik tersebut. Langkah berikutnya adalah membagi *dataset* tersebut menjadi 2 bagian yaitu data latih *(train & validation)* dan data uji *(test)*. Proporsi yang digunakan adalah 80:20 di mana proporsi yang lebih besar diambil sebagai data latih model. Selanjutnya data latih tersebut akan dipecah lagi menjadi 2 bagian dengan proporsi yang sama yaitu 80:20. Proporsi yang lebih kecil digunakan sebagai data validasi pada proses *hyperparameter* tuning model klasifikasi. Setelah dipecah terdapat total 4,071 *records* untuk data latih atau training, 1,017 *records* data validasi dan 1,272 records data uji atau *testing*.

2.3. *Testing*

Tahapan berikutnya adalah menguji performa model menggunakan data *testing* yang belum pernah dikenali oleh model. Pada tahapan ini digunakan model dengan parameter hasil *hyperparameter tuning*. Artinya, model yang diuji adalah model terbaik dari semua kemungkinan model klasifikasi yang telah diujicoba pada proses validasi.

2.4. *Evaluation*

Tahapan yang terakhir adalah melakukan proses evaluasi atas ketiga model yang telah dikembangkan untuk didapatkan model terbaik yang layak untuk dikembangkan ke dalam sistem rekomendasi ini. Dalam hal ini, terdapat tujuh buah metrik yang digunakan, mulai dari *accuracy* yang mengukur tingkat akurasi atau rasio prediksi benar dari model klasifikasi, *precision* yang mengukur rasio prediksi positif benar jika dibandingkan dengan keseluruhan hasil yang diprediksi positif oleh model klasifikasi, *recall* yang mengukur sensitivitas model klasifikasi melalui rasio prediksi positif benar jika dibandingkan dengan keseluruhan data yang memang benar positif, *F1-Score* yang merupakan ukuran keseimbangan antara *precision* dan *recall, AUC (Area Under The Curve) ROC (Receiver Operating Characteristics)* dan *Gini Coefficient* yang dapat digunakan untuk evaluasi visual atas performa model untuk masing-masing *threshold* klasifikasi dan sangat bermanfaat dalam kondisi *class imbalance,* serta *Log Loss* yang membandingkan probabilitas hasil prediksi dengan *ground truth label* yang dikuantifikasi dengan *cross-entropy.* Adapun ketiga model yang akan diperbandingkan adalah model berbasis *Multinomial Logistic Regression* (MLR), *Random Forest* (RF), dan *Support Vector Model* (SVM).

Model pertama menggunakan *Multinomial Logistic Regression* (MLR) yang merupakan model klasifikasi berbasis regresi logistik sederhana. Model MLR sendiri dapat digunakan untuk kasus klasifikasi di mana hasil luaran yang akan diprediksi mempunyai lebih dari dua kelas. Hal ini lah yang membedakan MLR dengan model regresi logistik sederhana yang digunakan untuk memprediksi hasil luaran dengan kelas biner. Model klasifikasi ini mampu

mengidentifikasi bagaimana hubungan antara variabel dependen dengan variabel independen. Fungsi logistik berupa kurva S *(sigmoid curve)* digunakan pada model klasifikasi untuk memprediksi probabilitas semua hasil luaran yang memungkinkan. Dalam implementasinya, model MLR menggunakan pendekatan *one-versus-one* (OVO) atau *one-versus-rest* (OVS) sehingga dapat digunakan untuk kasus klasifikasi dengan banyak kelas. Pendekatan OVO akan membandingkan satu per satu antar kelas, sedangkan OVS akan membandingkan secara spesifik satu kelas terhadap semua kelas lainnya. Artinya, pada model MLR akan didapatkan banyak model regresi logistik sederhana. Untuk memprediksi kelas hasil luaran akhir, digunakan fungsi aktivasi *softmax* yang menormalisasi probabilitas kelas menggunakan semua probabilitas hasil luaran dari masing-masing model regresi logistik sederhana yang didapatkan Salah satu contoh sistem rekomendasi yang dibangun dengan MLR adalah untuk tim olahraga kriket (Jayanth dkk, 2018).

Model kedua adalah *Random Forest* (RF) yang merupakan model klasifikasi berbasis konsep pohon keputusan *(decision tree)*. Model ini mengagregasikan banyak pohon keputusan yang independen guna memprediksi kelas hasil luaran. Tidak seperti pohon keputusan biasa yang menggunakan semua prediktor dan semua data sampel, RF hanya menggunakan beberapa prediktor saja. Sedangkan untuk data sampelnya, RF dapat memilih untuk menggunakan semua data sampel atau melakukan *sampling*, tergantung konfigurasi yang digunakan oleh peneliti. Pemilihan prediktor dan data sampel tersebut dilakukan secara acak *(random)*. Hal ini membuat setiap pohon keputusan yang terdapat pada model RF akan memiliki atribut yang unik dan berbeda dari pohon keputusan lainnya. Konsep yang diusung tersebut membuat model RF mampu mengenali dan mempelajari banyak opsi dari satu jenis *dataset* saja. Konsep tersebut juga yang membuat model RF semakin kaya akan informasi meskipun mengusung konsep yang cukup sederhana. Untuk memprediksi kelas hasil luaran digunakan skema voting. Kelas yang paling banyak muncul sebagai hasil prediksi dari semua pohon keputusan yang tersedia akan dipilih sebagai kelas hasil luaran akhir. Beberapa sistem rekomendasi yang dibangun menggunakan RF di antaranya adalah yang terkait produk (Khanvilkar dan Vora, 2018) dan kredit bank (Putra, 2019)

Model ketiga adalah *Support Vector Machine* (SVM) yang secara konsep digunakan untuk menemukan suatu batas atau dikenal dengan *hyperplane* yang mampu memaksimalkan jarak antar kelas. Pada dasarnya, model SVM digunakan untuk kasus linear, artinya hyperplane yang dicari adalah suatu garis linear yang mampu menjadi pemisah antar kelas. Namun bukan berarti model SVM tidak dapat digunakan pada kasus non-linear. Konsep kernel atau kernel tricks dapat digunakan pada kasus klasifikasi non-linear. Konsep tersebut bekerja dengan cara membawa data ke dalam dimensi yang lebih tinggi sehingga lebih mudah untuk dipetakan hyperplane-nya. Dengan konsep kernel, hyperplane pada model SVM dapat berbentuk non-linear. Beberapa contoh kernel yang dapat digunakan antara lain polynomial, radial basis function (RBF) dan sigmoid. Sama seperti MLR, model SVM sendiri sebenarnya digunakan untuk kasus klasifikasi dengan kelas biner. Untuk kasus dengan banyak kelas maka pendekatan OVO atau OVS dapat digunakan. Fungsi aktivasi softmax juga dapat digunakan untuk memprediksi kelas hasil luaran akhir model SVM. Beberapa sistem rekomendasi yang dibangun menggunakan SVM di antaranya adalah yang terkait berita (Fortuna, Fortuna, & Mladenić, 2010) dan penentuan dosen pembimbing (Pradana, 2020).

3. HASIL DAN PEMBAHASAN

3.1. Performa Model Tahap *Validation*

Model dikembangkan menggunakan bahasa pemrograman Python dan mengimplementasikan paket Scikit-Learn (Sklearn). Tabel 3 menunjukkan performa 3 model klasifikasi *machine learning* yang dilatih menggunakan nilai *default* atau parameter bawaan paket Sklearn. Dilihat dari beberapa metrik evaluasi, terlihat bahwa model RF memiliki performa yang paling baik dibandingkan MLR maupun SVM dilihat dari sisi akurasi maupun metrik evaluasi lainnya.

Metrik evaluasi yang digunakan pada penelitian ini antara lain *accuracy*, *precision*, *recall*, F1-score, *Area Under Receiver Operating Characteristic Curve* (AUC-ROC), Gini coefficient, dan Log loss. Formula untuk menghitung skor dari beberapa metrik seperti *accuracy*, *precision*, *recall*, *F1-score*, dan AUC-ROC dapat dilihat pada penelitian (Hossin dan Sulaiman, 2015), sedangkan formula untuk metrik Gini coefficient dan Log loss masing-masing didapatkan dari referensi (Srivastava, 2019) dan (Bishop, 2006).

Walaupun sudah mendapatkan performa yang cukup bagus pada model RF, akan tetapi proses *hyperparameter tuning* akan tetap dilakukan. Hal ini untuk melihat apakah proses tersebut mampu mendongkrak performa model terutama pada MLR dan SVM yang dinilai masih belum maksimal. Hyperparameter tuning dilakukan dengan metode random search setelah mengatur beberapa kemungkinan parameter model *(grid parameter*). Pada proses optimisasi tersebut, performa model akan diukur menggunakan data validasi yang proporsinya diambil 20% dari data latih.

Tabel 3. Performa model klasifikasi dengan *default parameter* pada Sklearn

Model	Akrs	Pres.	Recall	F1-Score	Avg AUCROC	Gini Coef	Log Loss
MLR	0.17	0.17	0.17	0.14	0.74	0.48	2.73
RF	0.84	0.83	0.84	0.83	0.97	0.95	0.87
SVM	0.11	0.07	0.11	0.07	0.69	0.41	2.90

Tabel 4. Performa model klasifikasi menggunakan Parameter hasil *hyperparameter tuning*

Model	Akurasi	Presisi	Recall	F1-Score	Avg AUCROC	Gini Coef	Log Loss
MLR	0.21	0.23	0.20	0.19	0.76	0.53	2.62
RF	0.84	0.82	0.84	0.83	0.97	0.94	0.74
SVM	0.20	0.23	0.20	0.17	0.78	0.57	2.61

Cross validation dengan jumlah validasi sebanyak 3 dilakukan untuk mengukur kemampuan model, sedangkan untuk kombinasi parameter model akan dicoba sebanyak 100 kombinasi parameter. Artinya, proses *hyperparameter tuning* akan membandingkan performa dari 300 model yang berbeda untuk dicari parameter mana yang mampu memberikan performa terbaik. Tabel 4 menunjukkan performa dari masing-masing model klasifikasi menggunakan parameter hasil *hyperparameter tuning*. Pada model RF, didapatkan performa yang tidak jauh berubah walaupun terlihat nilai *log loss*-nya yang menurun dibandingkan model sebelumnya, sedangkan untuk metrik evaluasi lain terlihat tidak ada perubahan yang signifikan. Berbeda dari model RF, pada model MLR maupun SVM terlihat bahwa terdapat perbaikan terutama terkait akurasi model. Bahkan pada model SVM, nilai *precision* dan F1 score-nya naik cukup banyak. Akan tetapi, secara keseluruhan performa terbaik masih diberikan oleh model klasifikasi RF.

Tabel 5 menunjukkan parameter pada setiap model klasifikasi setelah melalui hyperparameter tuning. Parameter-parameter tersebut menjadi inisiasi parameter pada model klasifikasi yang dilatih menggunakan paket Sklearn pada bahasa pemrograman Python. Peneliti menggunakan nilai *default* bawaan paket Sklearn Python untuk parameter-parameter model yang tidak disebutkan pada Tabel 5. Model klasifikasi yang dilatih menggunakan parameter tersebut selanjutnya akan disimpan untuk kemudian diujikan pada data *testing*.

Tabel 5. Parameter pada model klasifikasi

Model	Parameter model paket Sklearn Python
MLR	solver = 'newton-sg'; penalty = 'l2'; multi_class = 'multinomial'; max_iter = 640; C = 1438.450
RF	n_estimator = 400; min_samples_split = 2; min_samples_leaf = 2; max_features = 'sqrt'; max_depth = 64; bootstrap = false
SVM	probability = true; kernel = 'linear'; gamma = 'scale'; decision _function_shape = 'ovo'; class_weight = 'balanced'; C = 1.0

3.2. Performa Model Tahap *Testing*

Proses selanjutnya adalah menguji performa dari ketiga model menggunakan data *testing*. Hal ini dilakukan untuk mengetahui bagaimana performa model klasifikasi menggunakan *dataset* yang belum pernah dikenali. Model yang diimplementasikan adalah model yang sudah menggunakan parameter hasil *hyperparameter tuning*. Tabel 6 menunjukkan performa dari masing-masing model klasifikasi pada tahap ini. Model RF masih mengungguli kedua model lainnya dengan nilai akurasinya di angka 86%. Nilai akurasi tersebut tidak jauh berbeda dibandingkan performa model pada proses sebelumnya. Namun model RF pada data testing memiliki nilai skor AUC dan koefisien Gini terbaik dibandingkan pengujian sebelumnya sekaligus terbaik dibandingkan model MLR dan SVM. Selain itu model RF juga memiliki nilai Log loss yang paling rendah. Sedangkan pada model MLR dan SVM terlihat performa yang sedikit lebih baik walaupun tidak terlalu signifikan jika dibandingkan dengan performa pada proses sebelumnya.

Selain menggunakan metrik evaluasi di atas, performa model juga dapat dilihat melalui kurva ROC yang dapat dilihat pada Gambar 3, Gambar 4, dan Gambar 5. Baik model MLR, RF dan SVM ketiganya memiliki kurva ROC yang mana masing-masing garis kurvanya berada di atas garis diagonal. Artinya, setiap model klasifikasi tersebut sebenarnya memiliki performa prediksi yang cukup bagus, bahkan lebih baik dibandingkan model klasifikasi acak atau *random* yang divisualisasikan menggunakan garis diagonal pada kurva ROC. Meskipun performa yang bagus tersebut mungkin hanya terjadi pada beberapa kelas program studi tertentu saja. Hal ini ditunjukkan pada beberapa garis kurva yang letaknya tidak jauh di atas garis diagonal. Kurva sejenis tersebut dapat ditemukan pada kurva ROC model MLR dan SVM.

Di antara ketiga model klasifikasi yang dikembangkan, model RF kembali menunjukkan performa yang paling baik jika dinilai hanya menggunakan kurva ROC. Hal ini ditunjukkan dari garis kurva yang letaknya cukup jauh di atas garis diagonal dan lebih mendekati ke sudut kiri atas.

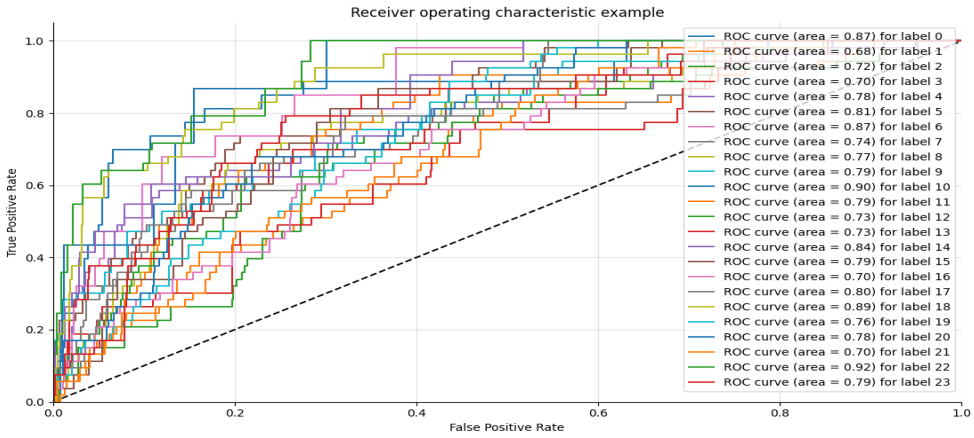
Tabel 6. Performa Model Klasifikasi pada Tahap *Data Testing*

Model	Akurasi	Presisi	Recall	F1-Score	Avg AUCROC	Gini Coef	Log Loss
MLR	0.21	0.21	0.21	0.19	0.78	0.57	2.58
RF	0.86	0.84	0.86	0.84	0.97	0.95	0.66
SVM	0.22	0.24	0.22	0.17	0.79	0.58	2.58

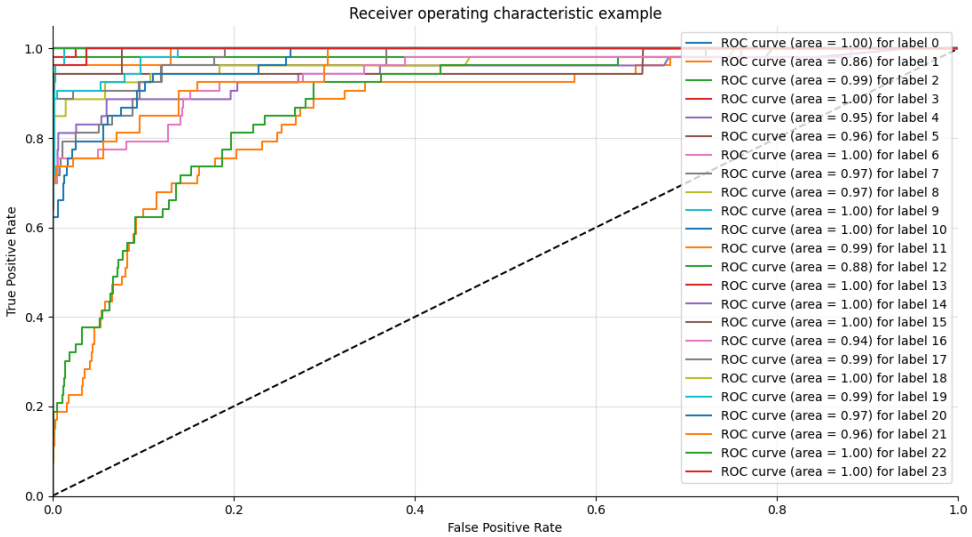
Artinya, tidak seperti model MLR dan SVM yang hanya cukup bagus untuk memprediksi program studi tertentu, model RF mempunyai kemampuan prediksi yang cukup bagus hampir untuk semua program studi yang terdapat pada *dataset*.

Berdasarkan evaluasi yang telah dilakukan menggunakan beberapa metrik evaluasi sekaligus kurva ROC, dapat diambil kesimpulan bahwa model RF memberikan performa yang paling baik dibandingkan model MLR dan SVM. Tidak hanya untuk beberapa program studi tertentu, akan tetapi model RF mempunyai kekuatan prediksi yang cukup bagus untuk hampir semua jenis program studi.

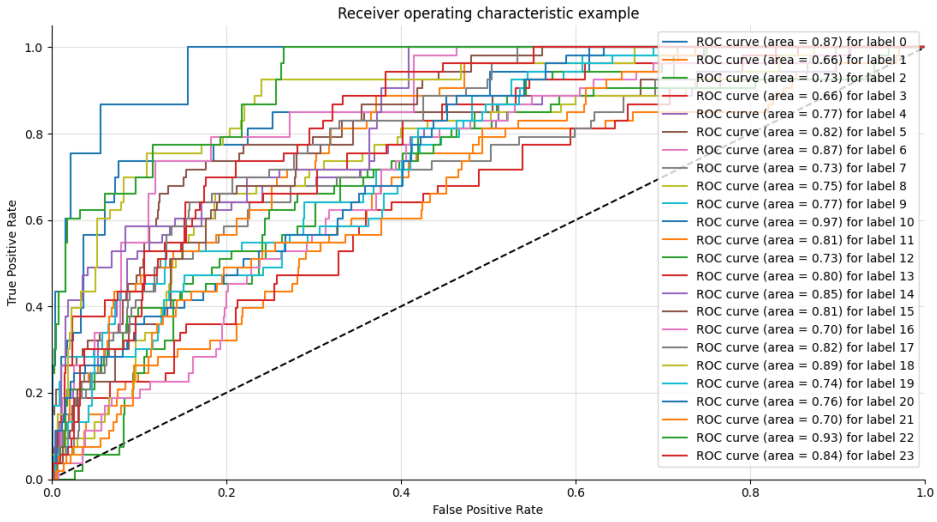
Secara detail, model RF tersebut mempunyai total 400 pohon keputusan untuk menentukan hasil klasifikasi. Masing-masing pohon keputusan memiliki kedalaman maksimal sebanyak 64 dan menggunakan semua *records* data dalam pembentukannya. Sebaliknya, setiap pohon keputusan tidak menggunakan semua atribut yang tersedia pada *dataset*, melainkan hanya menggunakan tiga sampai dengan empat atribut saja. Artinya, pohon keputusan yang terdapat pada model RF memiliki karakteristik yang berbeda-beda.



Gambar 3. Kurva ROC untuk model MLR



Gambar 4. Kurva ROC untuk model RF



Gambar 5. Kurva ROC untuk model SVM

4. KESIMPULAN

Dalam penelitian ini, tim penulis telah berhasil mengembangkan sebuah sistem rekomendasi pemilihan program studi sarjana berbasis *machine learning* dengan menggunakan data mahasiswa dan lulusan program sarjana di Universitas Islam Indonesia. Di antara ketiga model yang dikembangkan (MLR, RF, dan SVM), algoritma RF memberikan kinerja terbaik berdasarkan semua metrik yang ada, mulai dari akurasi 86%, presisi 84%, recall 86%, dan F1-Score 84%, hingga AUC-ROC 97%, Koefisien Gini 95%, dan Log-Loss 0,66.

Sesuai dengan peta jalan penelitian yang menaungi penelitian ini, tahapan penelitian berikutnya bertujuan untuk meningkatkan kualitas hasil rekomendasi yang diberikan, salah satunya dengan eksplorasi beberapa teknik tingkat lanjut seperti *multi-stage machine learning* (Mardani dkk., 2020) atau *collaborative filtering* (Nguyen dkk, 2020; Wei dkk., 2017), serta dengan mengintegrasikan data baru berbasis psikometri, khususnya dalam rangka memfasilitasi minat dan bakat calon mahasiswa baru dalam proses rekomendasi pemilihan prodi yang akan diberikan. Selain itu, penelitian selanjutnya juga akan menambahkan purwarupa sistem rekomendasi berbasis web dan aplikasi perangkat bergerak agar hasil dari penelitian ini dapat langsung dimanfaatkan oleh calon mahasiswa baru dalam proses PMB sebelum memilih program studi yang akan mereka jalani nantinya.

UCAPAN TERIMA KASIH

Tim penulis mengucapkan terima kasih kepada Jurusan Informatika, Fakultas Teknologi Industri, Universitas Islam Indonesia yang telah mendanai penelitian ini dan kepada Badan Sistem Informasi Universitas Islam Indonesia yang telah menyediakan data yang dibutuhkan dalam penelitian ini

DAFTAR PUSTAKA

BISHOP, C.M., 2006. Pattern Recognition and Machine Learning. Switzerland: Springer New York.

FAIZAL, E., 2015. Analisis Pemilihan Jurusan Favorit Menggunakan Metode Promethee (Studi Kasus Pada STMIK El Rahma Yogyakarta). Jurnal Fahma, 13.

FORTUNA, B., FORTUNA, C. dan MLADENIĆ, D., 2010, September. Real-time news recommender system. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases (pp. 583-586). Springer, Berlin, Heidelberg.

HOSSIN, M. dan SULAIMAN, M.N., 2015. A Review on Evaluation Metrics for Data Classification Evaluations. International Journal of Data Mining & Knowledge Management Process (IJDKP), 4-5

JAYANTH, S.B., ANTHONY, A., ABHILASHA, G., SHAIK, N. dan SRINIVASA, G., 2018. A team recommendation system and outcome prediction for the game of cricket. Journal of Sports Analytics, 4(4), pp.263-273.

KHANVILKAR, G. dan VORA, D., 2018. Sentiment analysis for product recommendation using random forest. International Journal of Engineering & Technology, 7(3), pp.87-89.

KHANAM, Z., dan ALKHALDI, S., 2019. An Intelligent Recommendation Engine for Selecting the University for Graduate Courses in KSA: SARS Student Admission Recommender System. In International Conference on Inventive Computation

Technologies (pp. 711-722). Springer, Cham.

KUMALA, A.T., BENARKAH, N. dan TJANDRA, E., 2015. Pembuatan Sistem Pendukung Keputusan Pemilihan Jurusan Kuliah Bagi Siswa SMA Berbasis Web dengan Metode Promethee. Calyptra, 4(2), pp.1-10

LIU, R., dan TAN, A., 2020. Towards interpretable automated machine learning for STEM career prediction. Journal of Educational Data Mining, 12(2), pp.19-32

MARDANI, A., LIAO, H., NILASHI, M., ALRASHEEDI, M., dan CAVALLARO, F. 2020. A multi-stage method to predict carbon dioxide emissions using dimensionality reduction, clustering, and machine learning techniques. Journal of Cleaner Production, 275.

MESYA, 2019. JPNN [online]. Tersedia di: https://www.jpnn.com/news/hasil-survei-87-persen-mahasiswa-pilih-jurusan-tidak-sesuai-minat [Diakses 11 Mei 2021]

NGUYEN, L. V., HONG, M. S., JUNG, J. J., dan SOHN, B. S. 2020. Cognitive Similarity-Based Collaborative Filtering Recommendation System. Applied Sciences, 10(12).

PICCIANO, A. G., 2012. The evolution of big data and learning analytics in American higher education. Journal of asynchronous learning networks, 16(3), pp.9-20.

PRADANA, Y.R., 2020. Sistem Rekomendasi Dosen Pembimbing Berdasarkan Latar Belakang Menggunakan Metode Multi-Class Support Vector Machine Dan Weighted Product (Doctoral dissertation, Universitas Brawijaya).

PUTRA, M.I., 2019. Sistem rekomendasi kelayakan kredit menggunakan metode Random Forest pada BRI Kantor Cabang Pelaihari (Doctoral dissertation, UIN Sunan Ampel Surabaya).

ROZI, A.F., and PURNOMO, A.S., 2018. Rekomendasi Pemilihan Minat Studi Menggunakan Metode Mamdani Studi Kasus: Program Studi Sistem Informasi FTI UMBY. INFORMAL: Informatics Journal, 2(3), pp.138-147.

RUMAISA, F., 2012. Penentuan Association Rule Pada Pemilihan Program Studi Calon Mahasiswa Baru Menggunakan Algoritma Apriori Studi Kasus pada Universitas Widyatama Bandung. In Seminar Nasional Aplikasi Teknologi Informasi 2012, Jurusan Teknik Informatika, Universitas Islam Indonesia, Yogyakarta.

SAM'AN, M., 2015. Implementasi Fuzzy Inference System sebagai Sistem Pengambilan Keputusan Pemilihan Program Studi di Perguruan Tinggi. UNNES Journal of Mathematics, 4(1).

SRIVASTAVA, T., 2019. Analytics Vidhya [online]. Tersedia di: https://www.analyticsvidhya.com/blog/2019/08/11-important-model-evaluation-error-metrics/ [Diakses 11 Mei 2021]

STINEBRICKNER, T.R. dan STINEBRICKNER, R., 2011. Math or science? Using longitudinal expectations data to examine the process of choosing a college major (No. w16869). National Bureau of Economic Research

SURYADI, A., 2018. Sistem Rekomendasi Penerimaan Mahasiswa Baru Menggunakan Naive Bayes Classifier Di Institut Pendidikan Indonesia. Joutica, 3(2), pp.171-182.

WANG, Z., dan SHI, Y., 2016. Prediction of the admission lines of college entrance examination based on machine learning. In 2016 2nd IEEE International Conference on Computer and Communications (ICCC) (pp. 332-335). IEEE.

WATERS, A. dan MIIKKULAINEN, R., 2014. Grade: Machine learning support for graduate admissions. AI Magazine, 35(1), pp.64-64.

WEI, J., HE, J., CHEN, K., ZHOU, Y., dan TANG, Z., 2017. Collaborative filtering and deep learning-based recommendation system for cold start items. Expert Systems with Applications, 69, pp.29-39.

WISWALL, M. dan ZAFAR, B., 2015. Determinants of college major choice: Identification using an information experiment. The Review of Economic Studies, 82(2), pp.791-824.

Halaman ini sengaja dikosongkan