Penerapan Random Forest, SMOTEN, dan *Class Merging* untuk Klasifikasi Status Pekerjaan Lulusan pada Data *Tracer Study* UIN Raden Mas Said Surakarta

Febrianta Surya Nugraha[*1], Layyin Mahfiana[2]
[1,2]Universitas Islam Negeri Raden Mas Said Surakarta
Jl. Pandawa, Dusun IV, Pucangan, Kec. Kartasura, Kabupaten Sukoharjo, Jawa Tengah 57168
Email : [1]febrianta.surya@staff.uinsaid.ac.id, [2]layyin.mahfiana@staff.uinsaid.ac.id

Abstract

Tracer studies are an important instrument for evaluating graduate quality and the alignment of higher education curricula with labor market needs. However, the utilization of tracer study data remains limited, while its inherent class imbalance often reduces the performance of classification models. This study aims to develop a graduate employment status classification model using the Random Forest algorithm with the SMOTEN approach and to compare three class merging strategies. The dataset consisted of 2,258 graduate records with 30 variables collected from the tracer study of UIN Raden Mas Said Surakarta. The research followed the CRISP-DM framework, including data preparation, modeling, and evaluation using Stratified 5-Fold Cross-Validation. The results indicate that the aggressive class merging strategy (two classes) achieved the best performance, with a test accuracy of 76.99%, balanced accuracy of 69.54%, and a reduction in the overfitting gap from 35.83% to 15.06%. Feature importance analysis revealed that the study–job relevance (20.62%), graduation GPA (17.09%), the timing of job searching (11.95%), internship/practicum experience (9.63%), and fieldwork experience (6.68%) were the most influential factors affecting graduates' employment status. These findings demonstrate that appropriate class merging effectively improves classification performance on imbalanced tracer study data while highlighting the importance of curriculum relevance and practical learning experiences in enhancing graduate employability..

Keywords: *class imbalance, class merging, Random Forest, SMOTEN, tracer study*

Abstraksi

Tracer study merupakan instrumen penting untuk mengevaluasi kualitas lulusan dan relevansi kurikulum terhadap kebutuhan dunia kerja. Namun, pemanfaatan data tracer study masih terbatas, sementara karakteristik datanya umumnya menghadapi permasalahan ketidakseimbangan kelas (class imbalance) yang dapat menurunkan performa model klasifikasi. Penelitian ini bertujuan membangun model klasifikasi status pekerjaan lulusan menggunakan algoritma Random Forest dengan pendekatan SMOTEN serta membandingkan tiga strategi class merging pada data tracer study UIN Raden Mas Said Surakarta yang terdiri atas 2.258 responden dan 30 variabel. Tahapan penelitian mengikuti kerangka CRISP-DM yang meliputi persiapan data, pemodelan, dan evaluasi menggunakan Stratified 5-Fold Cross Validation. Hasil penelitian menunjukkan bahwa

strategi aggressive class merging (2 kelas) memberikan performa terbaik dengan test accuracy sebesar 76,99%, balanced accuracy sebesar 69,54%, serta menurunkan overfitting gap dari 35,83% menjadi 15,06%. Analisis feature importance menunjukkan bahwa tingkat hubungan studi–pekerjaan (20,62%), IPK kelulusan (17,09%), waktu mulai mencari pekerjaan (11,95%), praktikum (9,63%), dan kerja lapangan (6,68%) merupakan faktor yang paling berpengaruh terhadap status pekerjaan lulusan. Temuan ini menunjukkan bahwa penyederhanaan kelas efektif meningkatkan performa klasifikasi pada data tracer study yang tidak seimbang, sekaligus menegaskan pentingnya kesesuaian kurikulum dan pengalaman praktis dalam meningkatkan kesiapan kerja lulusan.

Kata Kunci: *class merging, ketidakseimbangan kelas, Random Forest, SMOTEN, tracer study*

1. PENDAHULUAN

Tracer study atau studi penelusuran lulusan merupakan instrumen strategis dalam sistem pendidikan tinggi yang berfungsi sebagai umpan balik untuk mengevaluasi relevansi kurikulum dengan kebutuhan dunia kerja [1]. Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi telah menetapkan tracer study sebagai bagian dari Indikator Kinerja Utama (IKU) perguruan tinggi, khususnya IKU 1 mengenai lulusan yang memperoleh pekerjaan layak dan IKU 2 mengenai pengalaman belajar di luar kampus [2]. Data tracer study mencakup informasi mengenai masa tunggu memperoleh pekerjaan, bidang pekerjaan, kesesuaian pekerjaan dengan bidang studi, kompetensi lulusan, serta berbagai indikator lain yang dapat dimanfaatkan untuk mengevaluasi kualitas lulusan dan proses pembelajaran [3].

Namun demikian, pemanfaatan data tracer study di banyak perguruan tinggi, termasuk perguruan tinggi keagamaan Islam, masih terbatas pada kebutuhan akreditasi dan pelaporan institusi [4]. Data yang telah terkumpul umumnya belum dimanfaatkan secara optimal untuk menggali pola-pola pengetahuan yang dapat mendukung pengambilan keputusan strategis dalam pengembangan kurikulum maupun peningkatan mutu lulusan [5]. Salah satu penelitian di Indonesia juga menunjukkan bahwa analisis data tracer study dapat dimanfaatkan untuk memprediksi employability lulusan sehingga mendukung pengambilan keputusan berbasis data di perguruan tinggi [6].

Perkembangan teknik data mining dan machine learning membuka peluang baru dalam pemanfaatan data tracer study. Miranda dan Lhaksamana menerapkan algoritma Decision Tree dan Support Vector Machine untuk mengklasifikasikan masa tunggu kerja alumni Telkom University [7]. Haikal dan Palupi mengembangkan model prediksi

employability lulusan menggunakan Support Vector Machine dan memperoleh performa klasifikasi yang baik [8]. Cholil *et al.* memanfaatkan algoritma K-Nearest Neighbor untuk memprediksi lama masa tunggu alumni Universitas Semarang dalam memperoleh pekerjaan [9]. Selain itu, Dewi *et al.* mengembangkan sistem tracer study berbasis Agile untuk mendukung pencapaian indikator kinerja perguruan tinggi [10], sedangkan Salabi *et al.* menunjukkan bahwa data tracer study dapat dimanfaatkan sebagai instrumen evaluasi efektivitas program studi [11].

Meskipun berbagai penelitian telah memanfaatkan algoritma *machine learning* untuk memprediksi masa tunggu kerja maupun *employability* lulusan [7]–[9], penelitian-penelitian tersebut umumnya berfokus pada perbandingan performa algoritma klasifikasi tanpa membahas permasalahan ketidakseimbangan distribusi kelas (*class imbalance*) pada data *tracer study*. Padahal, karakteristik data *tracer study* sering kali memiliki distribusi kelas yang tidak seimbang sehingga berpotensi menurunkan kinerja model klasifikasi.

Salah satu tantangan utama dalam analisis data tracer study adalah ketidakseimbangan distribusi kelas (*class imbalance*), yaitu kondisi ketika jumlah data pada setiap kategori target tidak sebanding. Teknik SMOTE menjadi metode yang paling banyak digunakan untuk mengatasi permasalahan tersebut [12]. Selain itu, He dan Garcia menjelaskan bahwa ketidakseimbangan data merupakan salah satu penyebab utama menurunnya performa model klasifikasi pada berbagai kasus machine learning [13]. Pengembangan metode seperti Borderline-SMOTE serta berbagai teknik penyeimbangan data lainnya juga terbukti mampu meningkatkan kemampuan model dalam mengenali kelas minoritas [14], [15].

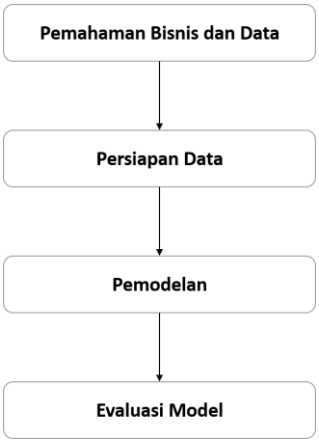
Selain penanganan ketidakseimbangan data, pemilihan algoritma klasifikasi juga memengaruhi kualitas prediksi. Random Forest merupakan salah satu algoritma ensemble yang memiliki kemampuan baik dalam menangani data berdimensi tinggi, relatif tahan terhadap *overfitting*, serta mampu menghasilkan informasi *feature importance* untuk mendukung interpretasi model [16]–[20]. Sebelum proses klasifikasi dilakukan, seleksi fitur juga diperlukan agar atribut yang digunakan benar-benar relevan sehingga mampu meningkatkan akurasi sekaligus mengurangi kompleksitas model [21].

Penelitian ini memanfaatkan data tracer study UIN Raden Mas Said Surakarta yang terdiri atas 2.258 responden dengan 30 variabel hasil tracer study tahun 2023–2024. Data

menunjukkan ketidakseimbangan yang cukup tinggi pada variabel status pekerjaan, di mana kategori bekerja *part time/full time* mendominasi sebanyak 1.365 responden (60,5%), sedangkan kategori wiraswasta sambil melanjutkan studi hanya berjumlah 36 responden (1,6%). Berdasarkan kondisi tersebut, penelitian ini bertujuan membangun model klasifikasi status pekerjaan lulusan menggunakan algoritma Random Forest dengan pendekatan SMOTEN, serta membandingkan tiga strategi *class merging* (tanpa *merging*, *moderate merging*, dan *aggressive merging*) untuk mengurangi dampak ketidakseimbangan data. Model dievaluasi menggunakan *Stratified 5-Fold Cross Validation* sehingga diharapkan mampu mengidentifikasi faktor-faktor yang paling berpengaruh terhadap status pekerjaan lulusan sekaligus memberikan rekomendasi berbasis data bagi pengembangan kurikulum dan peningkatan kualitas lulusan di perguruan tinggi.

2. METODE PENELITIAN

Tahapan penelitian mengikuti kerangka CRISP-DM (Cross-Industry Standard Process for Data Mining) karena menyediakan pendekatan sistematis yang telah banyak digunakan dalam berbagai penelitian data mining, termasuk pada bidang pendidikan [22]–[24]. Tahapan tersebut meliputi enam fase, yaitu *business understanding, data understanding, data preparation, modeling, evaluation,* dan *deployment*. Pada penelitian ini, fokus utama berada pada empat tahap awal, yaitu pemahaman bisnis dan data, persiapan data, pemodelan, serta evaluasi model. Tahapan penelitian yang dilakukan digambarkan pada gambar 1 berikut.



Gambar 1. Tahapan Penelitian

2.1 Pemahaman Bisnis dan Data

Tahap pertama dimulai dengan identifikasi tujuan bisnis, yaitu menganalisis faktor-faktor yang mempengaruhi status pekerjaan lulusan UIN Raden Mas Said Surakarta menggunakan data tracer study. Data yang digunakan merupakan data sekunder dari tracer study lulusan yaitu Jumlah responden 2.258 lulusan, 30 variabel, periode data tahun 2023-2024, dan sumber data diperoleh melalui Pengembangan Karir UIN Raden Mas Said Surakarta.

Variabel yang digunakan dalam penelitian ini terdiri atas variabel prediktor dan satu variabel target yang diperoleh dari data *tracer study* lulusan. Variabel prediktor dikelompokkan ke dalam beberapa kategori, yaitu aspek akademik, metode pembelajaran, fasilitas, waktu, kesesuaian, dan kompetensi. Pada aspek akademik digunakan variabel IPK lulus yang bertipe numerik sebagai representasi capaian akademik mahasiswa, serta sertifikasi yang bertipe kategorial untuk menunjukkan kepemilikan sertifikat kompetensi. Aspek metode pembelajaran mencakup variabel perkuliahan, demonstrasi/peragaan, proyek riset, magang, praktikum, kerja lapangan, dan diskusi, yang seluruhnya bertipe kategorial untuk menggambarkan tingkat kontribusi masing-masing metode pembelajaran terhadap proses belajar mahasiswa.

Selanjutnya, aspek fasilitas terdiri atas variabel perpustakaan, teknologi informasi, modul belajar, ruang belajar, laboratorium, variasi mata kuliah, pusat kegiatan mahasiswa, fasilitas kesehatan, dan fasilitas ibadah, yang seluruhnya merupakan variabel kategorial untuk merepresentasikan persepsi lulusan terhadap kualitas sarana dan prasarana yang tersedia selama masa studi. Aspek waktu diwakili oleh variabel kapan mulai mencari pekerjaan, yang menunjukkan waktu lulusan memulai proses pencarian kerja. Sementara itu, aspek kesesuaian meliputi variabel tingkat hubungan studi dengan pekerjaan dan tingkat pendidikan yang sesuai dengan pekerjaan, yang digunakan untuk mengukur relevansi pendidikan tinggi terhadap pekerjaan yang diperoleh lulusan.

Pada aspek kompetensi digunakan sejumlah variabel kategorial yang menggambarkan kemampuan lulusan saat menyelesaikan studi, yaitu daya saing, etika, profesionalisme, kemampuan bahasa Inggris, penggunaan teknologi informasi, kemampuan komunikasi, kemampuan kerja tim, dan kemampuan pengembangan diri.

Adapun variabel target dalam penelitian ini adalah status pekerjaan, yang bertipe kategorial dan digunakan sebagai kelas yang akan diprediksi oleh model *machine learning*.

2.2 Persiapan Data

Tahap persiapan data meliputi penanganan *missing value*, transformasi data, dan penanganan ketidakseimbangan kelas sebelum proses pemodelan. Berdasarkan analisis awal, ditemukan 2.130 nilai hilang yang tersebar pada beberapa variabel, dengan proporsi terbesar pada Tingkat Pendidikan Sesuai Pekerjaan (26,0%) dan Tingkat Hubungan Studi–Pekerjaan (21,1%). Penanganan *missing value* dilakukan menggunakan median untuk variabel numerik dan modus untuk variabel kategorial, sedangkan variabel dengan proporsi data hilang lebih dari 25% diperlakukan secara hati-hati untuk meminimalkan bias imputasi [25], [26].

Variabel IPK didiskretisasi menjadi lima kategori, yaitu *Rendah* (3,00–3,19), *Cukup* (3,20–3,39), *Sedang* (3,40–3,59), *Tinggi* (3,60–3,79), dan *Sangat Tinggi* (3,80–4,00) sesuai rentang yang umum digunakan pada perguruan tinggi di Indonesia [27], [28]. Hasil distribusi menunjukkan bahwa mayoritas responden berada pada kategori Tinggi sebanyak 1.180 responden (52,3%), diikuti kategori Sedang sebanyak 754 responden (33,4%).

Selanjutnya, seluruh variabel kategorial dikonversi ke bentuk numerik menggunakan Ordinal Encoding karena sebagian besar variabel memiliki tingkatan yang bersifat ordinal sehingga informasi urutan antar kategori tetap dipertahankan [29], [30].
Untuk mengatasi ketidakseimbangan kelas pada variabel target, penelitian ini membandingkan tiga strategi class merging, yaitu tanpa penggabungan (7 kelas), moderate merging (6 kelas), dan aggressive merging (2 kelas). Pada strategi *aggressive merging*, seluruh kategori status pekerjaan dikelompokkan menjadi dua kelas utama, yaitu Bekerja sebanyak 1.682 data (74,5%) dan Tidak Bekerja sebanyak 576 data (25,5%), yang selanjutnya digunakan untuk mengevaluasi pengaruh penggabungan kelas terhadap performa model klasifikasi.

2.3 Pemodelan

Seleksi fitur dilakukan menggunakan Mutual Information untuk mengidentifikasi atribut yang paling berpengaruh terhadap variabel target sehingga dapat mengurangi dimensi data dan meningkatkan performa model [31], [32]. Hasil seleksi menghasilkan 20

fitur yang mewakili aspek akademik, metode pembelajaran, fasilitas, waktu, kesesuaian pendidikan dengan pekerjaan, serta kompetensi lulusan. Sebelum proses pemodelan, dataset dibagi menjadi 70% data latih dan 30% data uji untuk mencegah *data leakage* [33]. Distribusi data latih terdiri atas 1.177 data kelas Bekerja dan 403 data kelas Tidak Bekerja, sedangkan data uji terdiri atas 505 dan 173 data yang ditampilkan pada tabel 1 berikut.

Tabel 1. Distribusi data setelah pembagian data

Set	Bekerja	Tidak Bekerja	Total
Train (70%)	1.177	403	1.580
Test (30%)	505	173	678
Total	**1.682**	**576**	**2.258**

Ketidakseimbangan kelas pada data latih kemudian diatasi menggunakan SMOTEN, sehingga kedua kelas menjadi seimbang dengan masing-masing 1.177 sampel. SMOTEN *resampling* ditampilkan pada tabel 2 berikut.

Tabel 2. SMOTEN *resampling*

Kelas	Sebelum SMOTEN	Sesudah SMOTEN
Bekerja	1.177	1.177
Tidak Bekerja	403	**1.177**
Total	**1.580**	**2.354**

Model klasifikasi dibangun menggunakan algoritma Random Forest karena memiliki kemampuan yang baik dalam menangani data berdimensi tinggi, tahan terhadap *overfitting* dan *noise*, serta mampu memberikan informasi mengenai tingkat kepentingan setiap fitur [34]–[39]. Parameter yang digunakan meliputi max_depth = 10, min_samples_split = 5, min_samples_leaf = 2, max_features = 'sqrt', class_weight = 'balanced', dan random_state = 42, sedangkan jumlah pohon (*n_estimators*) ditentukan menggunakan pendekatan early stopping berdasarkan performa pada data validasi [40], [41].

2.4 Evaluasi

Evaluasi model dilakukan menggunakan metrik Accuracy, Balanced Accuracy, Macro F1-Score, dan Weighted F1-Score untuk memberikan gambaran performa model pada data yang tidak seimbang [42]. Selain itu, validasi model dilakukan menggunakan Stratified 5-Fold Cross Validation agar proporsi kelas tetap terjaga pada setiap fold [43].

Untuk mendeteksi *overfitting*, dihitung Overfitting Gap, yaitu selisih antara akurasi data latih dan data uji [44]. Model dikategorikan memiliki performa excellent apabila *gap*

kurang dari 5%, good pada rentang 5–10%, moderate pada rentang 10–15%, dan mengalami overfitting apabila *gap* melebihi 15%.

3. HASIL DAN PEMBAHASAN

3.1 Analisis Data Eksploratif

Analisis eksploratif terhadap 2.258 data lulusan menunjukkan bahwa distribusi variabel target bersifat tidak seimbang. Sebagian besar lulusan berada pada kategori bekerja part time/full time sebanyak 1.365 responden (60,5%), diikuti tidak bekerja sebanyak 297 responden (13,2%), pernah bekerja tetapi saat ini tidak bekerja sebanyak 187 responden (8,3%), wiraswasta sebanyak 152 responden (6,7%), bekerja sambil melanjutkan studi sebanyak 129 responden (5,7%), melanjutkan studi sebanyak 92 responden (4,1%), dan wiraswasta sambil melanjutkan studi sebanyak 36 responden (1,6%). Secara keseluruhan, 74,5% lulusan berada pada kategori bekerja, sedangkan 25,5% berada pada kategori tidak bekerja. Distribusi ini menunjukkan adanya ketidakseimbangan kelas yang dijumpai pada data tracer study.

Distribusi IPK menunjukkan bahwa mayoritas lulusan memiliki prestasi akademik yang baik, dengan 52,3% berada pada kategori Tinggi (3,60–3,79) dan 33,4% pada kategori Sedang (3,40–3,59). Secara keseluruhan, 85,7% lulusan memiliki IPK di atas 3,40 dengan rata-rata 3,62, yang mencerminkan kualitas akademik lulusan yang baik. Temuan ini mengindikasikan bahwa IPK berpotensi menjadi salah satu prediktor dalam klasifikasi status pekerjaan lulusan, meskipun hubungan tersebut perlu dibuktikan lebih lanjut melalui proses pemodelan.

3.2 Perbandingan Strategi *Class Merging*

Tiga strategi *class merging* dievaluasi menggunakan algoritma Random Forest dengan parameter yang sama (*n_estimators* = 80 dan *max_depth* = 10). Hasil evaluasi pada data uji yang merupakan 30% dari total data ditampilkan pada tabel 3 berikut.

Tabel 3. Hasil Evaluasi Tiga strategi *class merging*

Metrik	7 Classes (No Merging)	6 Classes (Moderate)	2 Classes (Aggressive)
Training Accuracy	92.61%	92.43%	92.06%
Test Accuracy	56.78%	57.67%	76.99%
Balanced Accuracy	26.09%	30.97%	69.54%
F1-Macro	0.2495	0.3080	0.6962

Metrik	7 Classes (No Merging)	6 Classes (Moderate)	2 Classes (Aggressive)
F1-Weighted	0.5444	0.5531	0.7695
CV Mean	87.43%	87.00%	83.81%
OOB Score	87.96%	86.37%	83.86%
Overfitting Gap	35.83%	34.76%	15.06%
Status	overfitting	overfitting	Moderate

Hasil pengujian menunjukkan bahwa strategi aggressive merging (2 kelas) memberikan performa terbaik dibandingkan strategi tanpa penggabungan (7 kelas) maupun *moderate merging* (6 kelas). Strategi ini menghasilkan akurasi sebesar 76,99%, balanced accuracy sebesar 69,54%, Macro F1-Score sebesar 0,6962, dan Weighted F1-Score sebesar 0,7695, lebih tinggi dibandingkan dua strategi lainnya. Selain itu, *overfitting gap* berhasil ditekan dari sekitar 35% menjadi 15,06%, sehingga model menjadi lebih stabil. Temuan ini menunjukkan bahwa penyederhanaan kelas mampu meningkatkan kemampuan klasifikasi pada data yang tidak seimbang.

Analisis per kelas menunjukkan bahwa pada strategi 7 kelas dan 6 kelas, model hanya mampu mengklasifikasikan kelas mayoritas dengan baik, sedangkan sebagian besar kelas minoritas memiliki F1-Score di bawah 0,20. Bahkan pada strategi tanpa penggabungan, kelas Wiraswasta + Studi tidak berhasil diprediksi sama sekali (F1-Score = 0,00), yang mengindikasikan adanya bias model terhadap kelas mayoritas.

Sebaliknya, strategi aggressive merging mampu menghasilkan performa yang lebih seimbang untuk kedua kelas. Kelas Bekerja memperoleh Precision, Recall, dan F1-Score sebesar 0,85, sedangkan kelas Tidak Bekerja memperoleh F1-Score sebesar 0,55. Peningkatan Macro F1-Score dari 0,25 menjadi 0,70 menunjukkan bahwa penyederhanaan menjadi klasifikasi biner menghasilkan model yang lebih andal dan memiliki kemampuan generalisasi yang lebih baik.

Hasil analisis *overfitting* juga menunjukkan bahwa strategi aggressive merging mampu menurunkan *overfitting gap* secara signifikan dibandingkan dua strategi lainnya. Meskipun selisih akurasi data latih dan data uji masih berada pada kategori moderate, hasil ini menunjukkan bahwa penggabungan kelas merupakan pendekatan yang efektif untuk mengurangi *overfitting* pada data tracer study yang tidak seimbang.

3.3 Analisis *Feature Importance*

Untuk mengetahui faktor-faktor yang paling berpengaruh dalam prediksi status pekerjaan lulusan, dilakukan analisis *feature importance* menggunakan model Random Forest terbaik yang dihasilkan pada skenario 2 classes. Nilai *feature importance* menunjukkan besarnya kontribusi masing-masing variabel terhadap proses pengambilan keputusan model. Tabel 4 menampilkan 10 fitur dengan nilai *feature importance* tertinggi, yang secara kumulatif memberikan kontribusi terbesar dalam membedakan lulusan yang bekerja dan tidak bekerja.

Tabel 4. 10 fitur dengan nilai *feature importance* tertinggi

Rank	Fitur	Importance	Kategori
1	Tingkat Hubungan Studi-Pekerjaan	20.62%	Kesesuaian
2	IPK Lulus	17.09%	Akademik
3	Kapan Mulai Mencari Pekerjaan	11.95%	Waktu
4	Praktikum	9.63%	Pengalaman
5	Kerja Lapangan	6.68%	Pengalaman
6	Variasi Mata Kuliah	5.90%	Kurikulum
7	Ruang Belajar	5.35%	Fasilitas
8	Teknologi Informasi	3.45%	Fasilitas
9	Tingkat Pendidikan Sesuai Pekerjaan	3.38%	Kesesuaian
10	Modul Belajar	2.91%	Fasilitas

Analisis *feature importance* pada tabel 4 menunjukkan bahwa lima fitur utama berkontribusi sebesar 65,8% terhadap keseluruhan model. Fitur dengan kontribusi terbesar adalah Tingkat Hubungan Studi–Pekerjaan (20,62%), diikuti IPK Lulus (17,09%), Kapan Mulai Mencari Pekerjaan (11,95%), Praktikum (9,63%), dan Kerja Lapangan (6,68%). Hasil ini menunjukkan bahwa faktor kesesuaian pendidikan, capaian akademik, waktu persiapan karier, serta pengalaman praktik merupakan penentu utama dalam prediksi status pekerjaan lulusan.

Berdasarkan pengelompokan fitur, dimensi kesesuaian menjadi faktor yang paling berpengaruh dengan kontribusi sebesar 24,00%. Temuan ini mendukung penelitian Schomburg dan Teichler yang menyatakan bahwa kesesuaian antara bidang studi dan pekerjaan merupakan salah satu prediktor utama keberhasilan karier lulusan. Selanjutnya, dimensi akademik, yang diwakili oleh IPK Lulus (17,09%), juga memberikan pengaruh yang besar terhadap status pekerjaan.

Dimensi waktu, khususnya variabel Kapan Mulai Mencari Pekerjaan (11,95%), mengindikasikan bahwa lulusan yang mempersiapkan karier sejak sebelum lulus memiliki peluang lebih besar untuk memperoleh pekerjaan. Selain itu, dimensi pengalaman, yang terdiri atas Praktikum dan Kerja Lapangan dengan total kontribusi 16,31%, menunjukkan bahwa pengalaman belajar berbasis praktik berperan penting dalam meningkatkan kesiapan kerja lulusan.

Sementara itu, dimensi kurikulum dan fasilitas memberikan kontribusi sebesar 20,48%, yang terdiri atas Variasi Mata Kuliah, Ruang Belajar, Teknologi Informasi, Modul Belajar, dan Pusat Kegiatan Mahasiswa. Hasil ini mengindikasikan bahwa kualitas lingkungan pembelajaran turut memengaruhi kesiapan lulusan dalam memasuki dunia kerja, sejalan dengan panduan pengembangan kurikulum pendidikan tinggi.

4. KESIMPULAN

Berdasarkan hasil analisis data tracer study UIN Raden Mas Said Surakarta dengan pendekatan machine learning, dapat disimpulkan bahwa Model Random Forest dengan SMOTEN berhasil mengidentifikasi pola-pola penting dalam data tracer study yang sebelumnya belum tergali, membuktikan bahwa data tracer study memiliki potensi besar untuk diekstraksi menjadi pengetahuan berharga bagi pengembangan pendidikan tinggi. Strategi aggressive class merging dari tujuh kelas menjadi dua kelas (Bekerja dan Tidak Bekerja) terbukti paling efektif dalam mengatasi ketidakseimbangan data tracer study. Pendekatan ini meningkatkan test accuracy dari 56,78% menjadi 76,99% dan balanced accuracy dari 26,09% menjadi 69,54%, sekaligus menurunkan overfitting gap sebesar 58% (dari 35,83% menjadi 15,06%). Hasil tersebut menunjukkan bahwa penggabungan kelas yang tepat tidak hanya meningkatkan performa klasifikasi, tetapi juga memperbaiki kemampuan generalisasi model.

Faktor yang paling memengaruhi status pekerjaan lulusan adalah tingkat hubungan studi–pekerjaan (20,62%), IPK kelulusan (17,09%), waktu mulai mencari pekerjaan (11,95%), pengalaman praktikum (9,63%), dan kerja lapangan (6,68%). Secara keseluruhan, dimensi kesesuaian (24,00%) dan pengalaman praktis (16,31%) memiliki pengaruh yang lebih besar dibandingkan dimensi akademik (17,09%), yang menunjukkan bahwa relevansi

kurikulum dan pengalaman kerja lebih menentukan keberhasilan lulusan memasuki dunia kerja dibandingkan hanya prestasi akademik.

5. SARAN

Penelitian selanjutnya disarankan untuk menggunakan dataset yang lebih besar dan beragam dengan melibatkan beberapa angkatan serta perguruan tinggi agar hasil penelitian memiliki tingkat generalisasi yang lebih baik. Selain itu, penambahan variabel seperti program studi, lokasi kerja, tingkat gaji, dan metode pencarian kerja diharapkan dapat meningkatkan kemampuan prediksi model. Penelitian berikutnya juga dapat mengeksplorasi algoritma yang lebih canggih, seperti XGBoost, LightGBM, atau *stacking ensemble*, serta menerapkan analisis SHAP (*SHapley Additive exPlanations*) untuk memperoleh interpretasi yang lebih komprehensif terhadap kontribusi setiap fitur.

DAFTAR PUSTAKA

[1] H. Schomburg and U. Teichler, Higher Education and Graduate Employment in Europe: Results from Graduate Surveys from Twelve Countries. Dordrecht, The Netherlands: Springer, 2006.

[2] Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi Republik Indonesia, Indikator Kinerja Utama Perguruan Tinggi Negeri. Jakarta, Indonesia: Kemendikbudristek, 2021.

[3] H. Schomburg, Handbook for Graduate Tracer Studies. Turin, Italy: European Training Foundation, 2016.

[4] U. Teichler, "Does Higher Education Matter? Lessons from a Comparative Graduate Survey," European Journal of Education, vol. 42, no. 1, pp. 11–34, 2007.

[5] R. Chandra, S. Ruhama, and M. W. Sarjono, "Exploring Tracer Study Service in Career Center Web Site of Indonesia Higher Education," arXiv, arXiv:1304.5869, 2013.

[6] F. F. Abdulloh, M. Rahardi, A. Aminuddin, S. D. Anggita, and A. Y. A. Nugraha, "Observation of Imbalance Tracer Study Data for Graduates Employability Prediction in Indonesia," International Journal of Advanced Computer Science and Applications, vol. 13, no. 8, pp. 169–174, 2022, doi:10.14569/IJACSA.2022.0130820.

[7] A. Miranda and K. M. Lhaksamana, "Classification Analysis of Waiting Period for Telkom University Alumni to Get Jobs Using Decision Tree and Support Vector Machine," Building of Informatics, Technology and Science (BITS), vol. 4, no. 2, pp. 866–875, 2022, doi:10.47065/bits.v4i2.1963.

[8] M. F. Haikal and I. Palupi, "Predicting University Graduates Employability Using Support Vector Machine Classification," Building of Informatics, Technology and Science (BITS), vol. 6, no. 2, pp. 911–920, 2024, doi:10.47065/bits.v6i2.5655.

[9] S. R. Cholil, V. Vydia, Susanto, and Y. Cahyono, "Prediksi Lama Masa Tunggu Alumni Universitas Semarang dalam Mendapatkan Pekerjaan Menggunakan Algoritma K-Nearest Neighbor," Indonesian Journal on Computer and Information Technology, vol. 9, no. 2, pp. 246–255, 2024.

[10] D. A. Indah Cahya Dewi, N. K. Bagiastuti, and P. W. Sunu, "Development of a Tracer Study System Using Agile Approach for Meeting Higher Education Key Performance Indicators," in Proc. International Conference on Sustainable Applied Science (ICOSTAS-SAS), Atlantis Press, 2024.

[11] A. Salabi, M. A. M. Prasetyo, H. Halil, and M. Maulina, "Effectiveness of the Islamic Education Management Study Program Using Alumni Tracer Study Data," Jurnal Akuntabilitas Manajemen Pendidikan, vol. 12, no. 2, pp. 164–177, 2024.

[12] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, 2002.

[13] H. He and E. A. Garcia, "Learning from Imbalanced Data," IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 9, pp. 1263–1284, 2009.

[14] H. Han, W. Y. Wang, and B. H. Mao, "Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning," in Advances in Intelligent Computing. Berlin, Germany: Springer, 2005, pp. 878–887.

[15] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard, "A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data," ACM SIGKDD Explorations Newsletter, vol. 6, no. 1, pp. 20–29, 2004.

[16] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[17] A. Liaw and M. Wiener, "Classification and Regression by randomForest," R News, vol. 2, no. 3, pp. 18–22, 2002.

[18] T. K. Ho, "Random Decision Forests," in Proc. 3rd International Conference on Document Analysis and Recognition, 1995, pp. 278–282.

[19] G. Biau and E. Scornet, "A Random Forest Guided Tour," TEST, vol. 25, no. 2, pp. 197–227, 2016.

[20] C. Strobl, A.-L. Boulesteix, A. Zeileis, and T. Hothorn, "Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution," BMC Bioinformatics, vol. 8, Art. no. 25, 2007.

[21] G. Chandrashekar and F. Sahin, "A Survey on Feature Selection Methods," Computers & Electrical Engineering, vol. 40, no. 1, pp. 16–28, 2014.

[22] P. Chapman et al., CRISP-DM 1.0: Step-by-Step Data Mining Guide. Chicago, IL, USA: SPSS Inc., 2000.

[23] C. Shearer, "The CRISP-DM Model: The New Blueprint for Data Mining," Journal of Data Warehousing, vol. 5, no. 4, pp. 13–22, 2000.

[24] R. Wirth and J. Hipp, "CRISP-DM: Towards a Standard Process Model for Data Mining," in Proc. Practical Applications of Knowledge Discovery and Data Mining, 2000, pp. 29–39.

[25] R. J. A. Little and D. B. Rubin, Statistical Analysis with Missing Data, 3rd ed. Hoboken, NJ: Wiley, 2019.

[26] S. Van Buuren, Flexible Imputation of Missing Data, 2nd ed. Boca Raton, FL: CRC Press, 2018.

[27] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning, 2nd ed. New York: Springer, 2009.

[28] M. Kuhn and K. Johnson, Applied Predictive Modeling. New York: Springer, 2013.

[29] J. R. Vergara and P. A. Estévez, "A Review of Feature Selection Methods Based on Mutual Information," Neural Computing and Applications, vol. 24, pp. 175–186, 2014.

[30] R. Kohavi, "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection," in Proc. IJCAI, 1995, pp. 1137–1145.

[31] T. Fawcett, "An Introduction to ROC Analysis," Pattern Recognition Letters, vol. 27, no. 8, pp. 861–874, 2006.

[32] D. M. W. Powers, "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness and Correlation," Journal of Machine Learning Technologies, vol. 2, no. 1, pp. 37–63, 2011.

[33] G. C. Cawley and N. L. C. Talbot, "On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation," Journal of Machine Learning Research, vol. 11, pp. 2079–2107, 2010.

[34] C. Romero and S. Ventura, "Educational Data Mining and Learning Analytics: An Updated Survey," WIREs Data Mining and Knowledge Discovery, vol. 14, no. 2, 2024.

[35] Z. Ersozlu, S. Taheri, and I. Koch, "A Review of Machine Learning Methods Used for Educational Data," Education and Information Technologies, vol. 29, 2024.

[36] M. H. Alkhudhayr et al., "A Review of Educational Data Mining Trends," Procedia Computer Science, vol. 237, pp. 88–95, 2024.

[37] H. N. Mpia, L. W. Mburu, and S. N. Mwendia, "Applying Data Mining in Graduates' Employability: A Systematic Literature Review," International Journal of Engineering Pedagogy, vol. 13, no. 2, pp. 4–24, 2023.

[38] N. M. Almutairi et al., "Employability Prediction: A Survey of Current Approaches, Research Challenges and Applications," Journal of Ambient Intelligence and Humanized Computing, vol. 14, pp. 3479–3503, 2023.

[39] R. Scandurra et al., "Do Employability Programmes in Higher Education Improve Skills and Labour Market Outcomes?," Studies in Higher Education, vol. 49, no. 8, pp. 1381–1396, 2024.

[40] J. H. Guanin-Fajardo, J. Guaña-Moya, and J. Casillas, "Predicting Academic Success of College Students Using Machine Learning Techniques," Data, vol. 9, no. 4, Art. no. 60, 2024.

[41] A. H. Alqahtani et al., "Predictive Models for Educational Purposes: A Systematic Review," Big Data and Cognitive Computing, vol. 8, no. 12, Art. no. 187, 2024.

[42] D. Sartika et al., "The Impact of Internship Experience on the Employability of Vocational Students: A Bibliometric and Systematic Review," Cogent Education, vol. 11, 2024.

[43] M. S. N. Al-Din and H. A. Al Abdulqader, "Students' Academic Performance Prediction Using Educational Data Mining and Machine Learning: A Systematic Review," International Journal of Research and Innovation in Social Science, vol. 8, no. 8, 2024.

[44] V. A., S. B. E., and K. Sathish, "A Comprehensive Study of Employable and Technical Skill Development Programme in Higher Education Scenario: A Systematic Review," Educational Administration: Theory and Practice, vol. 30, no. 4, 2024.