 ISSN:2302-8149
e-ISSN:2540-9719

# Prediksi Ketereratan Karier Lulusan berdasarkan Aspek Akademik dan Non-Akademik berbasis Machine Learning

## *Prediction of Graduate Career Relevance Based on Academic and Non-Academic Aspects Using Machine Learning*

**[1]Muhammad Yusuf Luthfi Ijlal*, [2]Arif Setiawan, [3]Diana Laily Fithri**
[1,2,3]Program Studi Sistem Informasi, Fakultas Teknik, Universitas Muria Kudus
[1,2,3]Jl. Lkr. Utara, Gondangmanis, Bae, Kudus, Jawa Tengah, Indonesia 59327
*e-mail: yusufluthfi13@gmail.com

(***received*:** 27 May 2025, ***revised*:** 5 June 2025, ***accepted*:** 6 June 2025)

**Abstrak**

Salah satu indikator penting dalam menilai kualitas pendidikan tinggi adalah keberhasilan karier lulusan. Namun, tidak semua lulusan berhasil bekerja pada bidang yang relevan dengan disiplin ilmunya. Penelitian ini bertujuan untuk menganalisis pengaruh faktor akademik dan non-akademik terhadap ketereratan karier serta membangun model prediksi menggunakan algoritma *machine learning*. Data yang digunakan berasal dari hasil survei pelacakan alumni dan riwayat akademik mahasiswa Universitas Muria Kudus (UMK) dengan total 311 data setelah proses transformasi. Variabel prediktor yaitu Indeks Prestasi Kumulatif (IPK), lama studi, kemampuan kerja tim, dan keikutsertaan beasiswa, serta satu variabel target berupa ketereratan karier. Tiga algoritma digunakan dalam proses klasifikasi, yaitu *Decision Tree*, *Random Forest*, dan *Support Vector Machine* (*SVM*). Hasil menunjukkan bahwa variabel kemampuan kerja tim dan keikutsertaan beasiswa memiliki korelasi positif moderat terhadap ketereratan, lama studi berkorelasi negatif , dan IPK berkorelasi positif sangat lemah. *Random Forest* menunjukkan performa terbaik dengan nilai *AUC* sebesar 0.86 dan skor *Cohen's Kappa* sebesar 0.62, yang mengindikasikan klasifikasi berkualitas baik dan stabil. Hasil analisis pengaruh variabel dan model klasifikasi dapat digunakan pada sistem prediksi ketereratan karier serta dasar pengambilan keputusan dalam strategi peningkatan mutu lulusan yang selaras dengan kebutuhan dunia kerja.

**Kata kunci:** *machine learning*, *decision tree*, *random forest*, *support vector machine*, prediksi ketereratan karier

***Abstract***

*One of the key indicators for assessing the quality of higher education is the career success of its graduates. However, not all graduates succeed in securing employment in fields relevant to their area of study. This study aims to analyze the influence of academic and non-academic factors on career alignment and to develop a predictive model using machine learning algorithms. The data used in this study were obtained from an alumni tracer study and student academic records at Universitas Muria Kudus (UMK), comprising a total of 311 records after data transformation. The predictor variables include Grade Point Average (GPA), length of study, teamwork skills, and scholarship participation, with a single target variable representing career alignment. Three classification algorithms were applied: Decision Tree, Random Forest, and Support Vector Machine (SVM). The results show that teamwork skills and scholarship participation have a moderate positive correlation with career alignment, while length of study has a negative correlation, and GPA has a very weak positive correlation. Random Forest demonstrated the best performance with an AUC score of 0.86 and a Cohen's Kappa score of 0.62, indicating good and stable classification quality.*

 ISSN:2302-8149
e-ISSN:2540-9719

*The findings from the variable analysis and classification model can be utilized in developing a career alignment prediction system and as a basis for decision-making in strategies to improve graduate quality in line with labor market needs.*

***Keywords:*** *machine learning, decision tree, random forest, support vector machine, career alignment prediction*

## 1 Pendahuluan

Keberhasilan karier merupakan salah satu tujuan dari proses pendidikan tinggi. Lulusan perguruan tinggi diharapkan memiliki profesionalisme sesuai kebutuhan pasar kerja [1]. Didefinisikan sebagai profesional apabila dalam kariernya berposisi atau memiliki peran yang relevan dengan keterampilan yang dikuasai [2]. Namun fenomena di lapangan menunjukkan bahwa tidak semua lulusan berhasil bekerja di bidang yang relevan dengan disiplin ilmunya. Terdapat kasus di mana lulusan mendapatkan pekerjaan yang kurang linier dengan riwayat program studi. Kondisi demikian apabila tidak dilakukan evaluasi maka berpeluang menyebabkan inkompetensi dalam ruang lingkup pekerjaan. Ketidakmampuan menguasai bidang pekerjaan dapat berdampak buruk bagi kualitas kerja karena kompetensi menjadi salah satu aspek yang berpengaruh terhadap kinerja [3]. Kualitas kompetensi seseorang menurut [4] sangat dipengaruhi oleh kreativitas dan interpersonal. Kemampuan interpersonal dibentuk dari proses interaksi antar individu secara berkelanjutan secara selektif dan sistemis yang mampu membangun pengetahuan bersama [5], sementara itu kreativitas dilatih dan dikembangkan melalui proses akademis sebagai ajang menggali, mengamati, bertanya, hingga menguji [6]. Penguasaan aspek-aspek akademis tersebut proyeksikan sebagai nilai pembelajaran, sedangkan aspek non-akademis seperti kemampuan interpersonal dapat menjadi representasi dari penguasaan seseorang dalam keterlibatan, adaptasi, dan penempatan diri pada lingkungannya.

Menurut Teori Human Capital [7] bahwa pendidikan, keterampilan, dan produktivitas individu mempengaruhi kesesuaian dan keberhasilan dalam karier. Teori tersebut dikuatkan dengan Teori Person-Job Fit [8] yang menyatakan bahwa sesuai karakteristik keterampilan individu dengan tuntutan pekerjaan maka semakin besar peluang untuk sukses dalam karier. Berdasarkan dua teori tersebut dapat diasumsikan bahwa lulusan yang bekerja di luar bidang pendidikannya disebabkan tidak memenuhi kualifikasi keterampilan yang dibutuhkan. Lamanya masa studi dapat menyebabkan kurikulum tidak lagi relevan dengan kebutuhan pasar kerja dan terjadi kesenjangan kompetensi dan kebutuhan industri [9]. Akibatnya, lulusan cenderung memilih pekerjaan dengan syarat kualifikasi yang lebih umum dan mudah diakses. Keadaan tersebut menunjukkan perlunya analisis lebih lanjut untuk memahami faktor-faktor yang mempengaruhi ketereratan pekerjaan baik dari aspek akademik maupun non-akademik.

Analisis dilakukan melalui pendekatan *data mining* yang merupakan penggalian informasi dan pola dari sebuah dataset [10]. Pada penelitian ini proses tersebut menggunakan *pyhton* dan library *scikit-learn*. Dataset diproses menggunakan beberapa algoritma, diantaranya ID3 (*Iterative Dichotomiser* 3) yang menurut [11] adalah salah satu jenis algoritma yang bertipe pohon keputusan (*Decision Tree*), algoritma ini melakukan pencarian menyeluruh (*greedy*) pada semua pohon keputusan dari atas ke bawah untuk mencari atribut yang efektif mengklasifikasikan data. Kemudian *Random Forest* yang mengacak sampel (*sampling with replacement*) dan menghasilkan luaran berupa penggabungan beberapa pohon keputusan (*ensemble*) di mana menawarkan kemampuan pembobotan atau voting pada setiap *tree* untuk mendukung akurasi tingkat prediksi [12], serta algoritma *Support Vector Machine* (*SVM*) yang berfokus pada pencarian *hyperplane* terbaik untuk memisahkan dua kelas atau fitur secara optimal dengan mempertimbangkan titik-titik di sekitar batas yang ditentukan (*support vector*) guna meredam *overfitting* sehingga mendapatkan pola relevan tanpa gangguan dari data tidak penting [13].

Tujuan dari penelitian ini adalah untuk menganalisis seberapa tinggi keterpengaruhan ketereratan karier dengan beberapa faktor di aspek akademik dan non-akademik, mengetahui performa masing-masing algoritma yang digunakan, serta memprediksi tingkat ketereratan pekerjaan lulusan menggunakan model yang telah dilatih dengan mekanisme pembelajaran mesin (*Machine Learning*). Hasil penelitian ini diharapkan menghasilkan model yang bisa diaplikasikan untuk

 ISSN:2302-8149
e-ISSN:2540-9719

memberikan rekomendasi maupun bimbingan karier bagi calon lulusan serta menjadi literatur untuk menyusun strategi peningkatan kualitas lulusan yang terarah selaras dengan pasar kerja.

## 2 Tinjauan Literatur

Sejumlah penelitian mengimplementasikan model machine learning di berbagai persoalan, khususnya dalam hal ini adalah penggunaan algoritma dengan basis klasifikasi. Penelitian oleh [14] menganalisis pengaruh kontribusi data keluarga terhadap prediksi mahasiswa lulus tepat waktu menggunakan algoritma *Support Vector Machine* (*SVM*) dengan kernel linear. Sebanyak 365 *record* data lulusan diuji menunjukkan nilai akurasi sebesar 0.86, *precision* sebesar 0.86, *recall* sebesar 0.99, dan *F1-score* sebesar 0.92. Hasil ini berguna untuk membantu institusi pendidikan dalam meramalkan dan mengantisipasi mahasiswa yang berisiko tidak lulus tepat waktu. Penelitian selaras juga dilakukan oleh [15] dengan fokus menganalisis dan mengklasifikasikan mahasiswa berdasarkan ketepatan lama studi menggunakan algoritma *Random Forest* dan *Gradient Boosting*. Dataset yang digunakan mencakup atribut program studi, jenis kelamin, IPK, IPS, dan status mahasiswa. Hasil akurasi tertinggi sebesar 82,64% untuk *Random Forest* (10-*fold cross-validation*) dan 79,66% untuk *Gradient Boosting* (3-*fold cross-validation*). Perbandingan algoritma model klasifikasi juga digunakan pada penelitian [16]. Penelitian ini membandingkan algoritma *Random Forest*, *Support Vector Machine*, *AdaBoost* untuk klasifikasi Indeks Pembangunan Manusia (IPM) menggunakan 505 data dengan 9 atribut. Hasil penelitian diketahui bahwa Random Forest mendapatkan akurasi tertinggi sebesar 0,85 disusul *SVM* 0,84 dan AdaBoost 0.80. Penelitian berikutnya oleh [17] yang membandingkan kinerja algoritma *Random Forest* dan *Support Vector Machine* dalam memprediksi capaian studi mahasiswa. Penelitian ini menggunakan 433 sampel mencakup atribut jenis kelamin, asal sekolah, IPS, IPK, total SKS, serta status capaian studi. Penelitian ini menggunakan teknik seleksi fitur *Forward Selection* untuk mengidentifikasi atribut yang paling relevan. Hasil penelitian menunjukkan bahwa sebelum seleksi fitur, *Random Forest* memiliki akurasi sebesar 97,67%, sedangkan *SVM* mencapai akurasi 91,47%. Setelah seleksi fitur, akurasi kedua algoritma menjadi sama, yaitu 97,67%. Seleksi fitur pada *Random Forest* tidak memengaruhi akurasi, sementara pada *SVM*, seleksi fitur meningkatkan akurasi sebesar 6,2%. Hasil ini menunjukkan bahwa pemilihan fitur dapat meningkatkan performa algoritma tertentu.

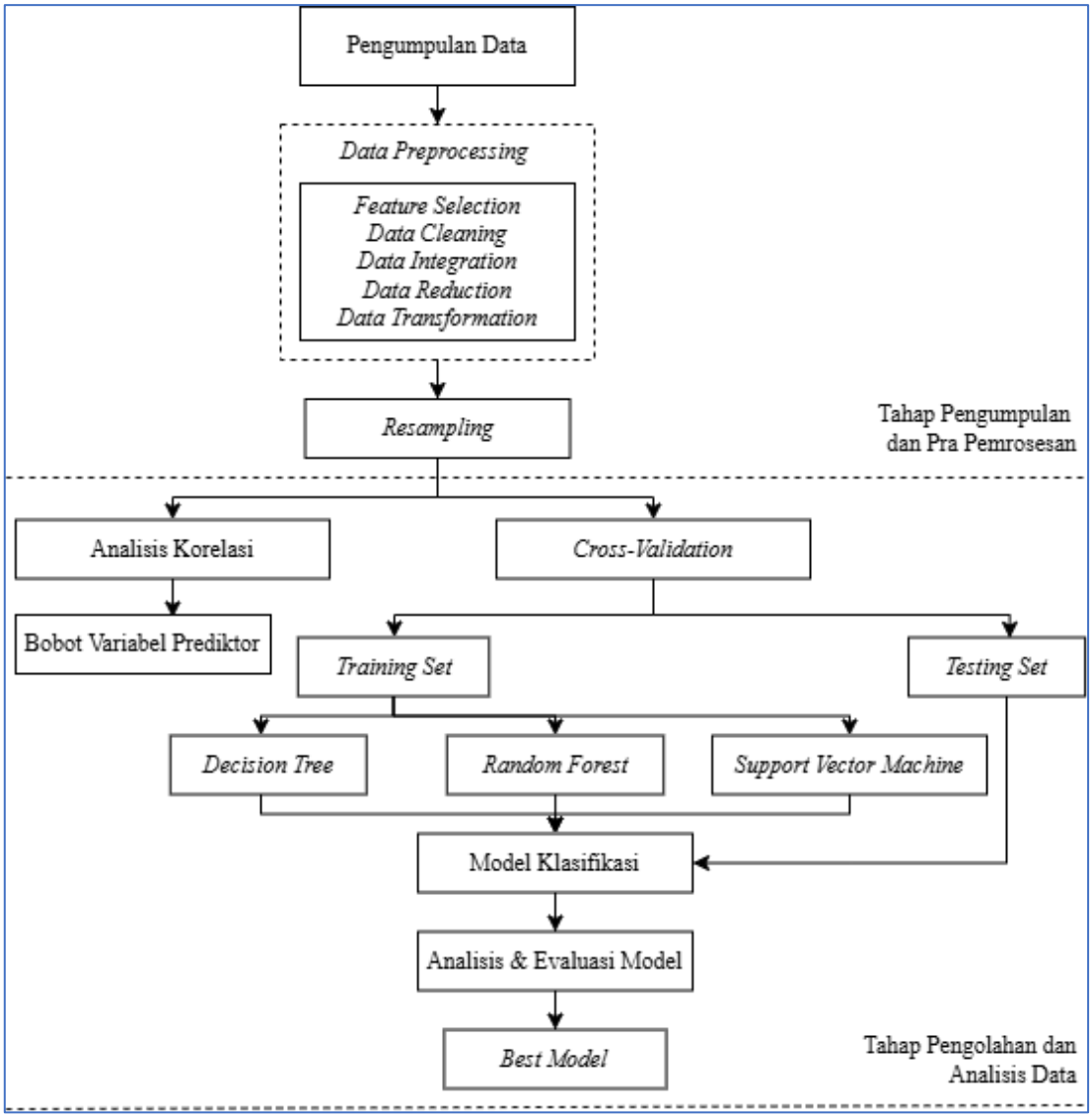
Sementara itu, komparasi algoritma guna merancang *machine learning* dalam ruang lingkup analisis faktor yang memengaruhi ketereratan bidang karier juga telah dilakukan di beberapa penelitian. Penelitian dari [18] mengimplementasikan algoritma *Decision Tree* untuk melihat keterkaitan antara mata kuliah yang diambil oleh mahasiswa dengan bidang karier yang dijalani setelah lulus. Sebanyak 91 data dari dataset lulusan Politeknik Gorontalo program studi Teknik Informatika lulusan tahun 2018 sampai 2021 dilakukan simulasi data mining berbasis pohon keputusan. Hasil penelitian menunjukkan bahwa algoritma *Decision Tree* mendapatkan skor *median accuracy*, 71,43%. Berdasarkan penelitian ini, *Decision Tree* dianggap efektif dalam memetakan pekerjaan lulusan yang ‘sesuai’ dan ‘sangat sesuai’ dengan kompetensi dari mata kuliah yang diambil. Penelitian lain juga dilakukan oleh [19] untuk memprediksi tingkat relevansi tingkat akademik terhadap bidang karier dengan menganalisis atribut proyek riset, praktikum, waktu tunggu kerja, bidang ilmu, dan hubungan studi dengan karier. Dataset diuji menggunakan algoritma C4.5, menggunakan standar *k-fold cross-validation* dengan ketentuan parameter 10 f*old* metode *confusion matrix*. Hasil dari penelitian mendapatkan skor akurasi 92% yang mengartikan C.45 optimal dalam pemilihan fitur atau atribut pengujian. Namun demikian hasil tersebut dinilai belum komprehensif karena hanya mengadopsi satu algoritma dalam memprediksi suatu probabilitas dari dataset. Berikutnya penelitian oleh [20], penelitian ini menganalisis ketereratan tingkat pendidikan dengan bidang karier alumni dengan mengolah 140 data alumni yang memuat atribut-atribut akademik seperti masa studi, indeks prestasi, bidang ilmu, waktu mencari pekerjaan, dan waktu memperoleh pekerjaan. Dataset diuji menggunakan algoritma *Iterative Dichitomiser* 3 (ID3) dengan metode *k-fold cross validation* serta *precentage split* guna mengukur tingkat keberhasilan algoritma ID3. Hasil penelitian menunjukkan algoritma ID3 mendapatkan skor rata-rata akurasi 76,87% yang dapat berarti bekerja cukup baik untuk memprediksi ketereratan tingkat pendidikan dengan pekerjaan. Metode perbandingan algoritma juga dilakukan [21] dengan menganalisis performa *Decision Tree* (J.48) dan Naïve Baiyes untuk memprediksi ketereratan bidang karier bagi lulusan program studi vokasi

 ISSN:2302-8149
e-ISSN:2540-9719

Sekretaris Universitas Budi Luhur menggunakan data lulusan tahun 2017. Penelitian tersebut bertujuan untuk menemukan algoritma yang cocok guna pengembangan Sistem Prediksi Ketereratan Kerja Alumni. Hasil penelitian memperlihatkan bahwa algoritma *Decision Tree* (J.48) mendapatkan akurasi tertinggi sebesar 95,8% sementara Naïve Baiyes mencapai 81,9%.

## 3 Metode Penelitian

Implementasi metode machine learning menggunakan model yang dirancang dari algoritma *Decision Tree*, *Random Forest*, dan *SVM* untuk menganalisis dan memperkirakan ketereratan karier lulusan dengan bidang studi mereka. Tahapan penelitian secara keseluruhan divisualisasikan pada Gambar 1 berikut.




**Gambar 1. Diagram tahapan penelitian**

### *3.1 Pengumpulan Data*

Data yang digunakan dalam penelitian ini merupakan hasil survei Unit Pusat Karir dan Pelacakan Alumni (PKPA) Universitas Muria Kudus (UMK) dari tahun 2020 hingga 2024 yang berjumlah 2.646 data. Selain itu, data juga diambil dari Biro Administrasi Akademik dan Kemahasiswaan (BAAK) UMK yang memuat riwayat akademik mahasiswa dari tahun akademik 2020/2021 hingga 2022/2023 sejumlah 3.985 data. Variabel penelitian mencakup 4 atribut prediktor yakni Indeks Prestasi Kumulatif (IPK), lama studi, kemampuan kerjasama, keikutsertaan beasiswa serta 1 variabel kriterium yakni ketereratan bidang karier. Dua sumber data dari objek berbeda dilakukan dekomposisi dan

 ISSN:2302-8149
e-ISSN:2540-9719

transformasi menggunakan program untuk memperoleh tabel data sesuai kebutuhan pengujian dengan akumulasi data terpakai 311 data. Daftar atribut data PKPA dan BAAK tertera pada Tabel 1 dan 2.

**Tabel 1. Atribut data PKPA**

| Deskripsi | Satuan | Tipe Data |
|---|---|---|
| 1. NIM | Tidak Berlaku (ID Unik) | Numerik |
| 2. Nama | Tidak Berlaku | Karakter |
| 3. Skill Kerjasama Tim | Skala 1 - 5 | Numerik |
| 4. Status Beasiswa | Kategori | Numerik |
| 5. Ketereratan Karier | Skala 1 – 5 | Numerik |

**Tabel 2. Atribut data BAAK**

| Deskripsi | Satuan | Tipe Data |
|---|---|---|
| 1. NIM | Tidak Berlaku (ID Unik) | Numerik |
| 2. Nama | Tidak Berlaku | Karakter |
| 3. Lama Studi | Tahun | Numerik |
| 4. IPK | Skala 0 - 4 | Numerik |

### *3.2 Data Preprocessing*

Data yang didapatkan dalam pengumpulan data terdapat beberapa permasalahan yang dapat menimbulkan gangguan saat pemrosesan atau *mining*. Oleh sebab itu, diperlukan pra pemrosesan data, yaitu tahapan yang dilakukan sebagai upaya dalam mendukung kualitas data melalui pembersihan data yang rumpang atau *missing*, pengintegrasian data, transformasi, serta pengurangan dimensi atau *Data Reduction* [22].

### *3.3 Resampling*

Dataset hasil *preprocessing* dilakukan pembagian menjadi dua bagian yakni data pelatihan (*training set*) untuk membangun dan melatih model serta data pengujian (*testing set*) yang berfungsi untuk menguji kemampuan model mengklasifikasikan data [23]. Dataset yang dibagi harus terdistribusi dengan baik (seimbang antar kelas), ketidakseimbangan kelas (class imbalance) dapat menyebabkan kesalahan klasifikasi (*missclassification*) pada kelas minoritas yang berdampak penurunan kinerja model [24]. Oleh karena itu pada penelitian ini dataset akan diseimbangkan terlebih dahulu dengan cara *resampling* [25] melalui metode *Synthetic Minority Oversampling Technique* (SMOTE) jika terindikasi tidak seimbang. SMOTE bekerja dengan meningkatkan bobot bagi kelas minoritas hingga setara dengan kelas mayoritas melalui penumbuhan data sintetik berdasarkan data terdekat [26].

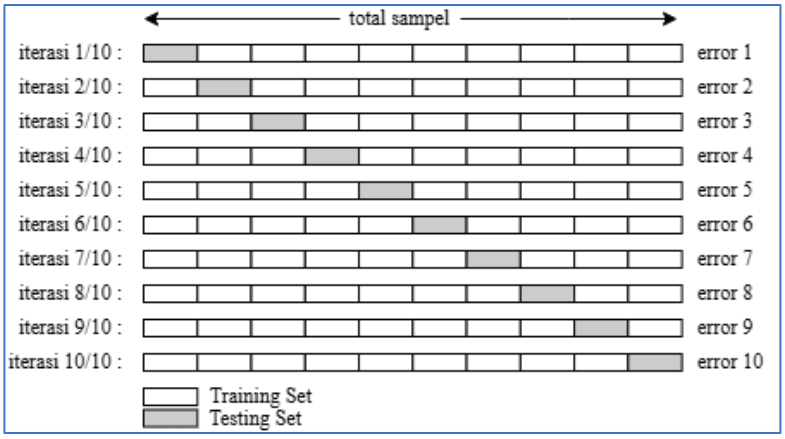
### *3.4 Analisis Korelasi*

Pengambilan nilai korelasi dapat dilakukan persis setelah pengumpulan data mentah (*raw data*) namun berisiko bias korelasi karena cenderung melihat kelas mayoritas. Penelitian [27] menyarankan korelasi pada data yang seimbang, dalam konteks ini setelah *resampling* dapat menggunakan *Pearson Correlation* karena mampu mengukur kekuatan dan arah korelasi variabel. Analisis korelasi diperlukan untuk memastikan fitur memiliki kadar keterkaitan yang cukup sebelum digunakan dalam pemodelan.

### *3.5 Cross-Validation*

Apabila kondisi kelas dataset seimbang dan variabel telah ditentukan, dilakukan *cross-validation* di mana merupakan metode pembagian data secara acak menjadi beberapa k bagian untuk *training* dan *testing*. Pada penelitian ini akan mengaplikasikan k-*fold cross-validation*, sub metode dari *cross-validation* yang umum digunakan karena mampu menghasilkan model dengan minim bias [28]. Ilustrasi k-*fold cross-validation* divisualisasikan pada Gambar 2.

 ISSN:2302-8149
e-ISSN:2540-9719




**Gambar 2. Visualisasi 10-*fold cross-validation***

Pengujian pada penelitian ini akan menerapkan 12 pembagian atau 10-*fold cross-validation* ($S_1$, $S_2$, ..., $S_{10}$) dengan skema bagian $S_1$ sampai $S_9$ sebagai data training untuk pembelajaran model, kemudian hasil model diuji menggunakan $S_{10}$ . Performa cross-validation dapat diukur berdasarkan rata-rata error yang didapatkan melalui rumus 1 berikut.

$$Error = \frac{1}{10}\sum_{i=1}^{10} Error_i \quad (1)$$

### *3.6 Modeling*

Pemodelan dilakukan setelah dataset terbagi menjadi *training set* dan *testing set* menggunakan algoritma. Penelitian ini akan menggunakan beberapa algoritma berjenis klasifikasi dan sub jenis *supervised learning* yakni *Decision Tree*, *Random Forest*, dan *Support Vector Machine*. *Supervised learning* yang mempunyai mekanisme kerja dengan membangkitkan suatu fungsi yang memetakan masukan dan luaran yang diinginkan diharapkan mampu menghasilkan model klasifikasi yang akurat [29]. Selain itu, adanya Random Forest bertujuan untuk menguji apakah algoritma berbasis *bagging ensemble learning* tersebut mampu menghasilkan performa klasifikasi lebih baik dengan mengurangi generalisasi [30].

### *3.7 Evaluasi Model*

*1) Confusion Matrix*

Sesuai dengan tujuan penelitian yaitu mengidentifikasi algoritma yang paling optimal dalam pengujian dataset lulusan, oleh sebab itu setiap luaran pengujian (*testing*) algoritma akan divisualisasikan menggunakan *confusion matrix* yaitu metode untuk evaluasi dengan tabel indikator atau matrix [31]. *Confusion matrix* berisi 4 indikator yaitu *True Positive* (TP), *True Negative* (TN), *False Positive* (FP), *False Negative* (FN). Struktur *confussion matrix* digambarkan pada Tabel 3 berikut.

**Tabel 3. Confusion matrix**

| *Confusion Matrix* | | ***Predicted Label*** | |
|---|---|---|---|
| | | *TRUE* | *FALSE* |
| ***Actual Label*** | *TRUE* | TP | FP |
| | *FALSE* | FN | TN |

Apabila hasil klasifikasi kelas prediksi bernilai *true* atau sama dengan kelas sebenarnya (TP dan TN) maka dinilai *correct classification* dan jika klasifikasi kelas prediksi bernilai *false* atau tidak sama dengan kelas sebenarnya (FP dan FN) maka dikategorikan sebagai *misclasification*. Kualitas kinerja model algoritma dinilai baik dari semakin banyaknya nilai pada TP dan TN [32]. Sebagai tolok ukur kemampuan masing-masing algoritma, pada penelitian ini model diukur berdasarkan *accuracy*, *precision*, *recall*, *f1-score*, *specificity* dan *Negative Predicted Value* (*NPV*), di mana nilai-nilai tersebut berkorelasi linear dengan peningkatan nilai TP dan TN [33].

 ISSN:2302-8149
e-ISSN:2540-9719

a. *Accuracy*
Indikator ini merepresentasikan presentase kelas yang diklasifikasikan secara tepat. Nilai *accuracy* dihitung menggunakan rumus 2.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad (2)$$

b. *Recall*/Sensitivitas
*Recall* mengukur kemampuan model dalam menangkap semua sampel positif yang ada di data. Nilai *recall* tinggi artinya sedikit *false negative*. Indikator ini dihitung dengan rumus 3.

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

c. *Precision*
Indikator ini mengukur ketepatan prediksi kelas positif pada model. Nilai *precision* tinggi artinya sedikit *false positive*. *Precision* diukur menggunakan rumus 4 berikut:

$$Precision = \frac{TP}{TP + FP} \quad (4)$$

d. *F1-Score*
*F1 Score* berperan dalam memberikan nilai rata-rata harmonis dari *precision* dan *recall* untuk menghindari *false positive* dan *false negative*. F1-Score dihitung dengan rumus 5 berikut ini:

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (5)$$

e. *Specificity*/Spesifisitas
*Specificity* mengukur kemampuan model dalam menangkap semua sampel negatif yang sebenarnya ada di data. Nilai *specificity* tinggi artinya sedikit *false positive*. Indikator ini dihitung dengan rumus 6 berikut:

$$Specificity = \frac{TN}{TN + FP} \quad (6)$$

f. *Negative Predicted Value* (*NPV*)
Indikator ini mengukur ketepatan prediksi kelas negatif pada model. Nilai NPV tinggi artinya sedikit *false negative*. NPV diukur menggunakan rumus 7 berikut:

$$NPV = \frac{TN}{TN + FN} \quad (7)$$

*2) ROC-AUC Curve*

Selain itu, digunakan parameter ROC-AUC (*Receiver Operating Characteristic - Area Under Curve*) untuk melihat kemampuan model membedakan kelas secara probabilistik [34]. Mekanisme pengukuran kinerja model dengan ROC-AUC didasarkan pada nilai AUC (*Area Under Curve*) . AUC diperoleh dari rumus 8 berikut.

$$AUC = \int_0^1 TPR(FPR)dFPR \quad (8)$$

$$TPR = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{FP + TN}$$

Skor AUC divisualisasikan dalam kurva ROC (*Receiver Operating Characteristic*) dan diinterpretasikan dalam beberapa kelompok [35] dapat dilihat pada Tabel 4.

**Tabel 4. Interpretasi nilai AUC**

| Interval | Kategori |
|---|---|
| 0.90 - 1.00 | *Excellent Classification* |
| 0.80 - 0.90 | *Good Classification* |
| 0.70 - 0.80 | *Fair Classification* |
| 0.60 - 0.70 | *Poor Classification* |
| 0.50 - 0.60 | *Failur* |

*3) Cohen's Kappa*

Penelitian ini juga menggunakan indikator *Cohen's Kappa* yang berperan untuk menilai konsistensi kesepakatan saat membedakan kelas. Penerapan *Cohen's Kappa* untuk evaluasi model dinilai penting karena dapat memberikan estimasi jumlah kesepakatan yang dihasilkan dari ketidaksengajaan dan murni kesepakatan [36]. Nilai *Cohen's Kappa* didapatkan dari perhitungan pada rumus 9 berikut [37].

$$\kappa = \frac{p_o - p_e}{1 - p_e} \quad (9)$$

 ISSN:2302-8149
e-ISSN:2540-9719

Nilai *observed accuracy* dan *expected accuracy* dapat diperoleh dari evaluasi nilai pada *confussion matrix* melalui rumus 10 dan 11 berikut.

$$p_o = \frac{TP + TN}{TP + TN + FP + FN} \quad (10)$$

$$p_e = \left(\frac{(TP + FP)(TP + FN)(FN + TN)(FP + TN)}{(TP + TN + FP + FN)^2}\right) \quad (11)$$

Interpretasi kualitas keputusan model berdasarkan nilai cohen's kappa diklasifikasikan dalam beberapa kelompok [38], diuraikan pada Tabel 5 berikut.

**Tabel 5. Interpretasi *cohen's kappa***

| Interval | Kategori |
|---|---|
| 0,81 – 1,00 | Sangat Tinggi |
| 0,61 – 0,80 | Tinggi |
| 0,41 – 0,60 | Sedang |
| 0,21 – 0,40 | Rendah |
| 0,01 – 0,20 | Sangat Rendah |

## 4 Hasil dan Pembahasan

### *4.1 Data Preprocessing*

Tahap *data preprocessing* pada penelitian ini mencakup *data cleaning*, *feature selection*, *data integration*, *data reduction*, dan *data transformation*. Alur pra proses diawali dengan mengunggah data PKPA (sampel dari tahun 2020 sampai 2023) dan data BAAK (utuh) dengan menerapkan *data cleaning* untuk menghindari *missing data*. Kedua data selanjutnya dilakukan *feature selection* untuk memilah atribut yang dibutuhkan, dari data PKPA dipilih atribut nim, skill kerjasama tim, status beasiswa, ketereratan karier. Sementara itu data BAAK dipilih atribut nim, ipk, dan lama studi. Tahap berikutnya dilakukan *data reduction* dengan mengonversi rentang data numerik menjadi data kategorikal. Setelah data menjadi kategorikal dilanjutkan proses *data integration* atau penggabungan data menggunakan mekanisme *inner join* berdasarkan atribut nim yang berada di kedua data. Struktur kedua data yang telah tergabung menjadi satu dataset dapat dilihat pada Tabel 6.

**Tabel 6. Struktur dataset hasil *preprocessing***

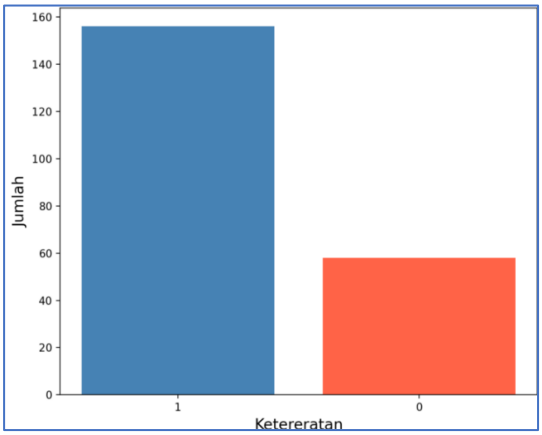
| Deskripsi | Variabel | Satuan | Tipe Data |
|---|---|---|---|
| 1. IPK | Independen | Kategori | Numerik |
| 2. Lama Studi | Independen | Kategori | Numerik |
| 3. Skill Kerjasama Tim | Independen | Kategori | Numerik |
| 4. Status Beasiswa | Independen | Kategori | Numerik |
| 5. Ketereratan Karier | Dependen | Kategori | Numerik |

### *4.2 Resampling*

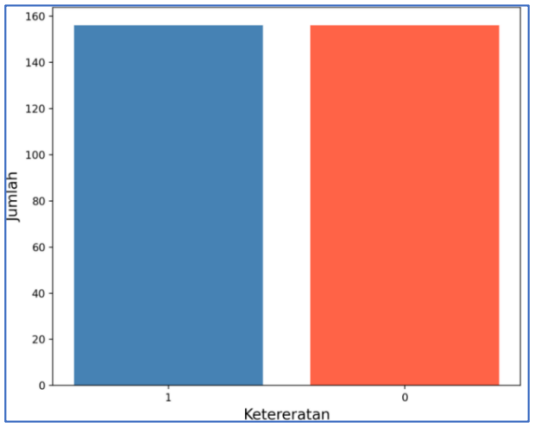
Setelah didapatkan dataset dengan variabel dependen dan independen, berikutnya dilakukan pemeriksaan komposisi data berdasarkan variabel dependen untuk memastikan bahwa dataset terdistribusi seimbang pada masing-masing fitur. Pada Gambar 3 terlihat bahwa sebelum dilakukan resampling, distribusi data menurut fitur ketereratan karier tersebar tidak seimbang (*imbalance*). Kelas 1 mewakili kelas positif sebagai 'erat' lebih mendominasi dibandingkan kelas 0 yang mewakili kelas negatif sebagai 'tidak erat'. Sebaran data yang imbalance dapat mengakibatkan misclasification atau kesalahan klasifikasi oleh model saat proses training dan berpengaruh buruk pada tingkat akurasi. Oleh karena itu diperlukan proses penyeimbangan data melalui resampling dengan metode *Synthetic Minority Oversampling Technique* (SMOTE). SMOTE berperan untuk meminimalisasi redundansi data tanpa menghilangkan informasi yang dibutuhkan. Pada kasus ini penggunaan SMOTE untuk resampling berhasil menyeimbangkan distribusi data berdasarkan kelas 'erat' dan 'tidak erat' seperti

 ISSN:2302-8149
e-ISSN:2540-9719

pada Gambar 4. Dataset hasil SMOTE tergolong sudah seimbang karena dari total 312 data, jumlah sebaran kedua fitur sama bernilai 156 data.



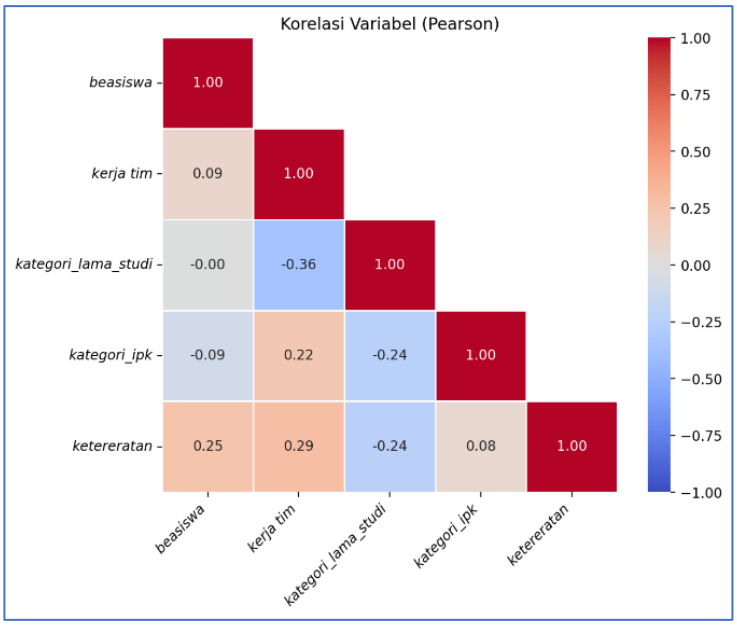

**Gambar 3. Distribusi data sebelum *resampling***




**Gambar 4. Distribusi data setelah *resampling***

### *4.3 Analisis Korelasi*

Dataset yang telah diseimbangkan, selanjutnya mengukur tingkat korelasi menggunakan *pearson* antara variabel prediktor dan kriteria untuk mengetahui kekuatan dan arah pengaruh. Pada Gambar 5 plot korelasi menjukkan adanya hubungan positif dari variabel kerja tim (0.29), beasiswa (0.25), dan kategori IPK (0.08), serta hubungan negatif kategori lama studi (-0.24) terhadapat ketereratan. Hasil tersebut mengindikasikan peningkatan kerja tim dan beasiswa berbanding lurus dengan skor ketereratan dengan kekuatan sedang, sementara peningkatan IPK berbanding lurus dengan skor ketereratan dengan kekuatan sangat lemah, dan penurunan angka lama studi berbandi terbalik dengan ketereratan.




**Gambar 5. Plot korelasi atribut**

### *4.4 Cross Validation*

Setelah *data resampling* dan variabel ditentukan, proses berikutnya adalah membagi dataset menjadi *training set* dan *testing set* menggunakan metode k-*fold cross-validation* dengan ketentuan 10 fold. Menurut penelitian [39] 10 *fold* dinilai memberikan estimasi akurasi minim bias dibandingkan dengan metode *cross-validation* lain seperti *leave-one-out* atau *boostrap*.

 ISSN:2302-8149
e-ISSN:2540-9719

```
skf = StratifiedKFold(n_splits=10, shuffle=True, random_state=42)
for train_idx, test_idx in skf.split(X, y):
    X_train, X_test = X[train_idx], X[test_idx]
    y_train, y_test = y[train_idx], y[test_idx]
```

**Gambar 6. *StratifiedKFold* untuk *cross-validation***

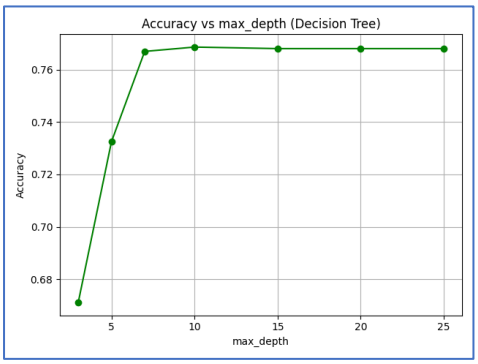
Pada Gambar 6 diatas, penerapan k=10 (*n-splits*) akan mendistribusikan dataset ke dalam 10 bagian yang terdiri atas 9 bagian sebagai data latih (*training set*) dan 1 bagian sebagai data uji (*testing set*) secara bergiliran sehingga seluruh data berkesempatan menjadi data uji 1 kali. Seperti pada Gambar 6, proses tersebut diulang 10 kali menggunakan *StratifiedKFold* dari *library Scikit-learn* logika *looping* sehingga setiap bagian dataset (*fold*) digunakan 9 kali untuk *training* dan 1 kali untuk *testing.*

### *4.5 Modeling*

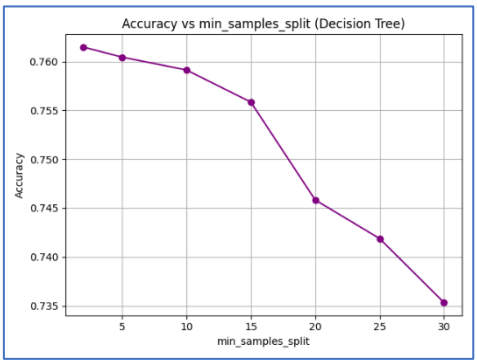
Setelah *training set* dan *testing set* terbentuk melalui *Stratified K-Fold Cross-Validation* , tahapan selanjutnya melakukan pemodelan. Proses ini meliputi pelatihan model menggunakan data latih pada setiap *fold*. Parameter algoritma disesuaikan dengan *GridSearchCV* untuk mendapat kombinasi dengan akurasi terbaik. Pada penelitian ini alur modeling diimplementasikan terhadap tiga algoritma yaitu:

1) *Decision Tree*

Pelatihan menggunakan *Decision Tree* dilakukan dengan kelas *DecisionTreeClassifier* untuk membangun model klasifikasi berbasis pohon (*tree*) melalui pembagian data berdasarkan atribut paling informatif. *GridSearchCV* diterapkan untuk mencari kombinasi *max_depth* dan *min_samples_split*. Hasil pengujian untuk menentukan model terbaik berdasarkan pemilihan kedalaman pohon dan minimal pembagian data menunjukkan akurasi tertinggi yang diperoleh saat *max_depth* 10 dan *min_samples_split* kurang dari 5. Stabilitas akurasi terjaga walaupun kedalaman pohon ditambah namun menurun saat jumlah sampel *split* dinaikkan. Pencarian model dengan penyesuaian parameter divisualisasikan pada Gambar 7 dan Gambar 8.




**Gambar 7. Kedalaman pohon dan skor akurasi**

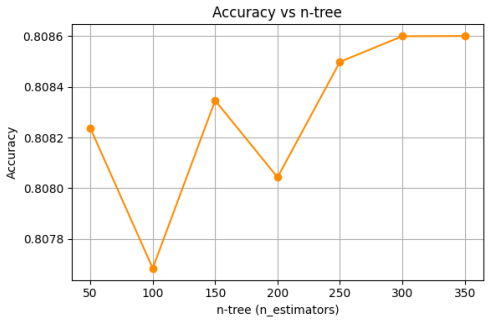



**Gambar 8. Jumlah sampel saat *split* dan skor akurasi**

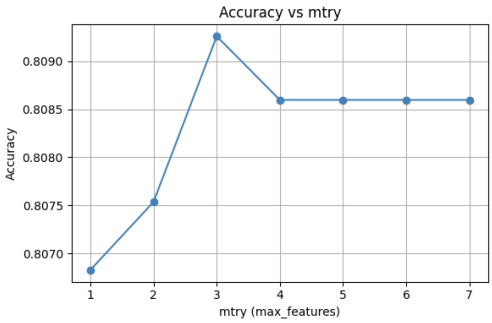
2) *Random Forest*

Pemodelan *Random Forest* dilakukan dengan kelas *RandomForestClassifier* melalui pemilihan kombinasi *n_estimators* dan *max_features*. Hasil percobaan menunjukkan model mendapatkan performa terbaik pada kombinasi jumlah pohon sebanyak 300 dan jumlah atribut sebanyak 3. Percobaan model dengan penyesuaian parameter divisualisasikan pada Gambar 9 dan Gambar 10.

 ISSN:2302-8149
e-ISSN:2540-9719




**Gambar 9. Jumlah pohon dan skor akurasi**




**Gambar 10. Jumlah atribut dan skor akurasi**

*3) Support Vector Machine*

Percobaan pada model *Support Vector Machine* (*SVM*) dilakukan dengan beberapa pengujian menggunakan tiga kernel yakni *linear*, *polynomial*, dan *Radial Basis Function* (*SVM*). Hasil percobaan menunjukkan bahwa model mendapatkan akurasi terbaik pada percobaan ke-119 saat menerapkan kernel *rbf*, *C*=10 , *gamma* = 0.2519 seperti yang ditampilkan pada Gambar 11. Nilai *C*=10 yang tergolong besar menunjukkan bahwa model lebih fokus pada mengklasifikasikan semua titik data dengan benar dan sangat tidak mentoleransi kesalahan.

| | kernel | C | gamma | best_accuracy |
|---|---|---|---|---|
| 114 | linear | 1 | scale (0.2519) | 0.7296 |
| 58 | poly | 10 | scale (0.2478) | 0.7687 |
| 119 | rbf | 10 | scale (0.2519) | 0.7971 |

**Gambar 11. Hasil percobaan model *SVM***

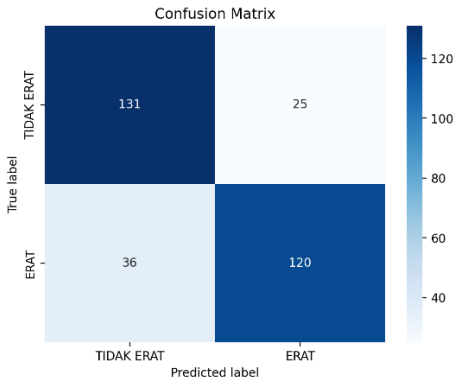
### *4.6 Evaluasi Model*

Setelah model klasifikasi terbangun dari tiga algoritma berbeda, langkah berikutnya yakni membandingkan performa setiap model berdasarkan beberapa hasil evaluasi yang diperoleh. Berikut beberapa hasil evaluasi berdasarkan indikator *confusion matrix*, *cohen's kappa,* dan *ROC-AUC* yang menjadi tolok ukur model.

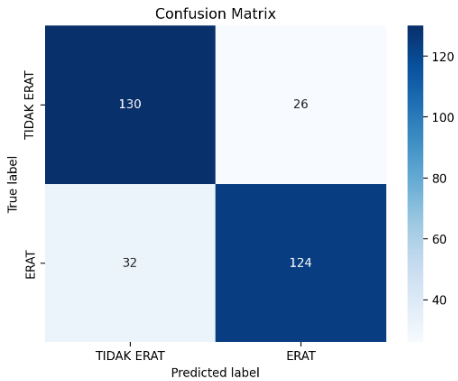
*1) Confusion Matrix*

Berdasarkan hasil akurasi yang didapatkan, selanjutnya mengukur kemampuan model dalam mengklasifikasikan label menggunakan visualisasi confusion matrix untuk membandingkan ketepatan prediksi label pada kelas prediksi terhadap kelas sebenarnya.

*a)*

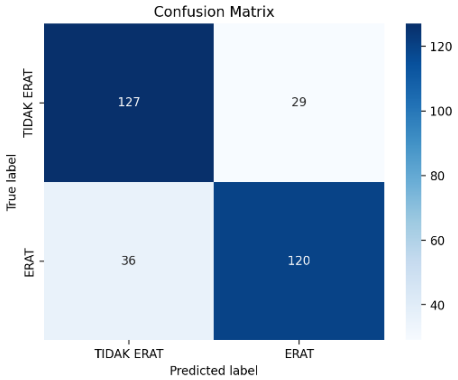



*b)*



 ISSN:2302-8149
e-ISSN:2540-9719

*c*)




**Gambar 12 *Confusion matrix* (a) *DT*, (b) *RF*, (c) *SVM***

Pada Gambar 11, menunjukkan bahwa ketiga model mendapatkan performa berbeda. Kemampuan klasifikasi terbaik diperoleh model *Random Forest* dengan *False Positive* 32 *False Negative 26* dari 156 data di masing-masing fitur. Hasil *confusion matrix* selanjutnya dipertimbangkan melalui perbandingan indikator *accuracy*, *precision*, *recall*, *f1-score*, *specificity* dan *Negative Predicted Value* (*NPV*) yang berkorelasi liniear dengan *confusion matrix.*

**Tabel 7. Hasil evaluasi model**

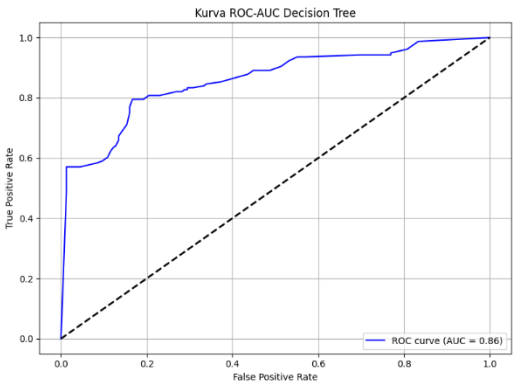
| Model | *accuracy* | *precision* | *recall* | *f1-score* | *specificity* | *NPV* |
|---|---|---|---|---|---|---|
| DT | 0.8045 | 0.806 | 0.8045 | 0.8042 | **0.8397** | 0.7844 |
| RF | **0.8141** | **0.8146** | **0.8141** | **0.8104** | 0.8333 | **0,8024** |
| SVM | 0.7971 | 0.7923 | 0.7971 | 0.7916 | 0.8141 | 0.7791 |

Berdasarkan Tabel 7 dapat disimpulkan bahwa model *Random Forest* menunjukkan performa terbaik dalam klasifikasi dibandingkan *Decision Tree* dan *SVM.* Kondisi tersebut dibuktikan dari nilai akurasi tertinggi (81.41%) serta perolehan skor indikator penunjang lainnya yang unggul di hampir semua metrik. Meskipun *Decision Tree* lebih unggul dalam *specificity*, *Random Forest* memberikan keseimbangan terbaik sehingga dapat dinilai sebagai model paling optimal untuk klasifikasi data ini.

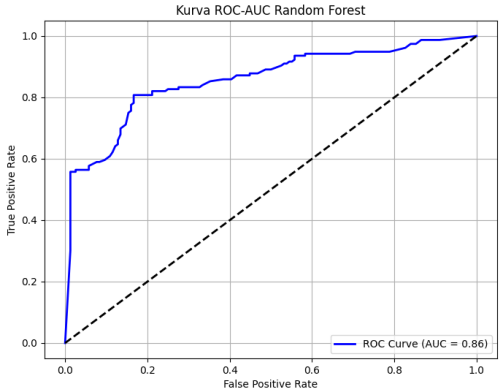
*2) ROC-AUC*

Setelah mengetahui bahwa *Random Forest* memiliki kemampuan yang lebih optimal dibanding model lain, perlu dilakukan pengecekan seberapa baik model benar-benar memisahkan kelas ‘erat dan ‘tidak erat’ melalui pengukuran *Area Under Curve* (AUC).

*a*)

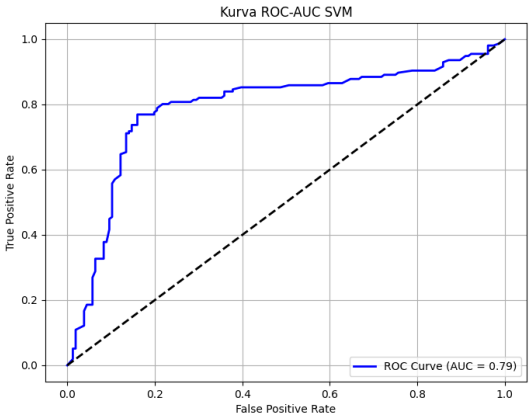



*b*)



 ISSN:2302-8149
e-ISSN:2540-9719

*c*)




**Gambar 13. Grafik *AUC* model**

Hasil perbandingan pada Gambar 13 menunjukkan perolehan *AUC* dari *Random Forest* sebesar 0.86 sama dengan *Decision Tree* dan lebih besar dari *SVM* (0.79). Skor 0.86 menegaskan bahwa model tidak hanya akurat, tetapi juga mampu membedakan kelas dengan baik. Sesuai dengan ketentuan interpretasi maka klasifikasi *Random Forest* tergolong sebagai *Good Classification*.

*3) Cohen's Kappa*

Tahapan berikutnya adalah pengecekan *cohen's kappa*, nilai dari parameter ini sebagai indikator tingkat validitas kesepakatan model. Kemampuan prediksi setiap model perlu dipastikan apakah benar-benar sesuai dengan label sebenarnya atau terjadi secara tidak sengaja (acak).

**Tabel 8. Perolehan *cohen's kappa* model**

| Model | *kappa* |
|---|---|
| DT | 0.60 |
| RF | **0.62** |
| SVM | 0.58 |

Berdasarkan nilai Cohen's kappa pada Tabel 8, model *Random Forest* mencatatkan skor 0.62 lebih besar dari *Decision Tree* dan *SVM*. Hasil tersebut menunjukkan bahwa performa model memiliki validitas kesepakatan kategori tinggi sehingga tidak hanya besar secara numerik tetapi cukup valid dan stabil secara statistik.

## 5 Kesimpulan

Berdasarkan pembahasan serta hasil penelitian yang telah dipaparkan, maka dapat diperoleh beberapa kesimpulan. Pertama, korelasi lama studi bernilai negatif mengindikasikan bahwa semakin kecil lama studi atau lulus lebih cepat menambah peluang karier lebih erat. Di sisi lain, aspek IPK, beasiswa, dan skill kerjasama tim berkorelasi positif menunjukkan bahwa semakin tinggi nilai dari ketiga aspek tersebut maka menambah peluang karier lebih erat. Aspek kerja tim berpengaruh paling signifikan dalam menentukan ketereratan karier dibandingkan dengan ketiga aspek lain, sementara itu aspek IPK memiliki korelasi paling kecil dan sangat lemah yang berarti tidak terlalu memengaruhi ketereratan karier. Kesimpulan berikutnya, kemampuan klasifikasi ketiga model memiliki selisih perbedaan yang kecil. Model *Random Forest* terbukti lebih unggul dibandingkan *Decision Tree* serta *SVM*, dibuktikan dengan perolehan skor *accuracy* dan metrik evaluasi tertinggi. Hasil pengujian tersebut menunjukkan implementasi *Random Forest* tergolong baik dalam membangun model klasifikasi ketereratan karier lulusan. Model yang telah terbangun dengan kemampuan mumpuni tersebut dapat diaplikasikan dalam sistem prediksi dini ketereratan karier calon lulusan, selain itu angka korelasi aspek akademik dan non-akademik dapat dimanfaatkan sebagai indikasi awal pengambilan keputusan guna peningkatan kualitas calon lulusan.

## Referensi

[1] K. D. Pramesti, N. I. Meisya, And R. Amrillah, "Relevansi Lulusan Perguruan Tinggi dengan Dunia Kerja," *J. Pendidik. Islam Dan Sos. Agama*, Vol. 03, No. 04, Pp. 236234–243, 2024.

 ISSN:2302-8149
e-ISSN:2540-9719

[2] I. Ilyas, "Strategi Peningkatan Kompetensi Profesional Guru," *J. Inovasi, Eval. Dan Pengemb. Pembelajaran*, Vol. 2, No. 1, Pp. 34–40, 2022.
[3] A. R. Safiih And A. Waliyudin, "Pengaruh Kompetensi dan Disiplin Kerja Terhadap Kinerja Karyawan Pt Kerta Mulya Sejahtera Di Rumpin Bogor," *Jorapi J. Res. Publ. Innov.*, Vol. 2, No. 1, Pp. 1384–1394, 2024, [Online]. Available: Https://Jurnal.Portalpublikasi.Id/Index.Php/Jorapi/Index
[4] Anis Dela Desela, Rianti Setiadi, And Yekti Widyaningsih, "Analisis Faktor-Faktor yang memengaruhi Kompetensi Karyawan Sos Children's Villages Indonesia," *J. Stat. Dan Apl.*, Vol. 4, No. 2, Pp. 95–108, 2020.
[5] Nur Maghfirah Aesthetika, "Komunikasi Interpersonal," *Komun. Interpers.*, 2018, [Online]. Available: Https://Doi.Org/10.21070/2018/978-602-5914-13-3
[6] Widiyaningrum and Harnanik, "Faktor-Faktor yang mempengaruhi Kreativitas Belajar Siswa Kelas XII Pemasaran pada Pembelajaran Produktif Pemasaran di SMK Negeri 1 Purbalingga," *Econ. Educ. Anal. J.*, Vol. 5, No. 3, Pp. 729–735, 2016.
[7] D. I. K. Sukasari And E. R. Tem-, "Human Capital: A Theoretical and Empirical Analysis With Special Reference to Education (3rd Edition)," No. Suppan, Pp. 1–16, 1994.
[8] A. L. Kristof, "Person-Organization Fit: An Integrative Review of Its Conceptualizations, Measurement,and Implications," *Pers. Psychol.*, Vol. 49, No. 1, Pp. 1–49, 1996.
[9] A. Isma, A. Isma, A. Isma, And A. Isma, "Peta Permasalahan Pendidikan Abad 21 di Indonesia," *J. Pendidik. Terap.*, Vol. 01, No. September, Pp. 11–28, 2023.
[10] Y. Mardi, "Data Mining : Klasifikasi menggunakan Algoritma C4.5," *Edik Inform.*, Vol. 2, No. 2, Pp. 213–219, 2017, Doi: 10.22202/Ei.2016.V2i2.1465.
[11] A. Rahmawati, D. Wintana, and D. A. Ambarsari, "Penerapan Algoritma Iterative Dichotomiser Three ( ID3 ) dalam Mendiagnosa Kesehatan Kehamilan," Vol. 06, No. 02, Pp. 116–127, 2019.
[12] D. Truong, *Data Science And Machine Learning For Non-Programmers: Using Sas Enterprise Miner*. In Chapman \& Hall/Crc Data Mining And Knowledge Discovery Series. Crc Press, 2024. [Online]. Available: Https://Books.Google.Co.Id/Books?Id=Jghzeaaaqbaj
[13] A. S. Nugroho, A. B. Witarto, and D. Handoko, "Support Vector Machine," 2003.
[14] W. Wijiyanto And S. Sopingi, "Kontribusi Keluarga dalam Prediksi Mahasiswa Lulus Tepat Waktu menggunakan Model Support Vector Machine," *Dutacom*, Vol. 17, No. 1, Pp. 25–36, 2023.
[15] M. L. Mu'tashim And A. Zaidiah, "Klasifikasi Ketepatan Lama Studi Mahasiswa dengan Algoritma Random Forest dan Gradient Boosting (Studi Kasus Fakultas Ilmu Komputer Universitas Pembangunan Nasional Veteran Jakarta)," *Semin. Nas. Mhs. Bid. Ilmu Komput. Dan Apl.*, Vol. 4, No. 1, Pp. 155–166, 2023.
[16] R. I. Arumnisaa And A. W. Wijayanto, "Comparison Of Ensemble Learning Method: Random Forest, Support Vector Machine, Adaboost For Classification Human Development Index (Hdi)," *Sistemasi*, Vol. 12, No. 1, P. 206, 2023.
[17] E. Novianto, S. Suhirman, And D. Prasetyo, "Perbandingan Metode Klasifikasi Random Forest dan Support Vector Machine dalam Memprediksi Capaian Studi Mahasiswa," Vol. 9, No. 4, Pp. 1821–1833, 2024.
[18] N. Wati And S. Mamase, "Simulasi Analisis Hubungan Antara Kompetensi Belajar dan Kesesuaian Pekerjaan Alumni Lulusan Kampus Vokasi Menggunakan Metode Decision Tree," *J. Teknol. Inf. Indones.*, Vol. 7, No. 2, Pp. 106–111, 2023, Doi: 10.30869/Jtii.V7i2.1094.
[19] B. Santoso, "Penerapan Teknik Data Mining Algoritma C4. 5 untuk Prediksi Kesesuaian Tingkat Pendidikan Dengan Bidang Pekerjaan Pada Alumni Universitas Sriwijaya," 2021.
[20] R. P. Oktarian, A. Desiani, and I. Irmeilyana, "Penerapan Algoritma Iterative Dichotomiser 3 (Id3) untuk memprediksi Kesesuaian Tingkat Pendidikan dengan Bidang Pekerjaan …," 2018, [Online]. Available: Https://Repository.Unsri.Ac.Id/4889/%0ahttps://Repository.Unsri.Ac.Id/4889/1/Rama_44201_08011381419053_0011127702_0017057403_01_Front_Ref.Pdf
[21] D. R. Utari, "Prediksi Bidang Kerja bagi Lulusan Program Studi Vokasi Sekretaris menggunakan Teknik Klasifikasi Data Mining," *J. Sekr. Adm.*, Vol. 16, No. 2, Pp. 115–123, 2018.

 ISSN:2302-8149
e-ISSN:2540-9719

[22] I. M. Hamdani[1] *Et Al.*, "Intisari Jurnal Inovasi Pengabdian Masyarakat Edukasi dan Pelatihan Data Science dan Data Preprocessing," *Juni*, Vol. 2, No. 1, Pp. 19–26, 2024.
[23] T. Utami, D. A. Shofiana, And Y. Heningtyas, "Penerapan Algoritma C4 . 5 untuk Prediksi Churn Rate Pengguna Jasa Telekomunikasi," Vol. 8, No. 2, Pp. 69–76, 2020.
[24] S. Diantika, F. Teknik, U. Bina, And S. Informatika, "Penerapan Teknik Random Oversampling untuk mengatasi Imbalance Class dalam Klasifikasi Website Phishing," Vol. 7, No. 1, Pp. 19–25, 2023.
[25] Ciciana, "Metode Resampling Dan Random Forest Untuk Klasifikasi Data Program Studi Matematika Universitas Sulawesi Barat Tahun 2024," 2024.
[26] N. V Chawla, K. W. Bowyer, L. O. Hall, And W. P. Kegelmeyer, "Smote : Synthetic Minority Over-Sampling Technique," Vol. 16, Pp. 321–357, 2002.
[27] I. M. Nasir *Et Al.*, "Pearson Correlation-Based Feature Selection For Document Classification using Balanced Training," *Sensors (Switzerland)*, Vol. 20, No. 23, Pp. 1–18, 2020.
[28] And K. P. Y. Widyaningsih, G. Puspita Arum, "Aplikasi K-Fold Cross Validation dalam Penentuan Model Regresi Binomial Negatif Terbaik Application Of K-Fold Cross Validation In Determining The Best Negative Binomial Regression Model," Vol. 15, No. 2, Pp. 315–322, 2021.
[29] F. Sodik, B. Dwi, And I. Kharisudin, "Perbandingan Metode Klasifikasi Supervised Learning pada Data Bank Customers menggunakan Python," Vol. 3, Pp. 689–694, 2020.
[30] G. Valentini And F. Masulli, "Ensembles Of Learning Machines," Pp. 3–4, 2002.
[31] R. Indransyah *Et Al.*, "Klasifikasi Sentimen Pergelaran Motogp di Indonesia menggunakan Algoritma Correlated Naïve Bayes Clasifier," Pp. 60–66, 2022.
[32] H. A. J. I. Pratama *Et Al.*, "Implementasi Algoritma C4.5 dan Naïve Bayes Classification untuk Mengklasifikasi Mahasiswa Berpotensi Drop Out," 2024.
[33] O. Arifin And T. B. Sasongko, "Analisa Perbandingan Tingkat Performansi Metode Support Vector Machine dan Naïve Bayes Classifier," No. 2016, Pp. 67–72, 2018.
[34] A. Setiawan, Z. H. Nst, Z. Khairi, And L. Efrizoni, "Klasifikasi Tingkat Risiko Diabetes menggunakan Algoritma," Vol. 7, No. 2, Pp. 263–271, 2024.
[35] Fatmawati, "Perbandingan Algoritma Klasifikasi Data Mining Model C4 . 5 dan Naive Bayes untuk Prediksi Penyakit Diabetes," Pp. 50–59, 2016.
[36] F. Bahiyyah And I. Indiati, "Analisis Kesalahan Siswa SMP dalam Menyelesaikan Soal Literasi Matematika berdasarkan Metode Newman ditinjau dari Kemandirian Belajar Pendahuluan Kemampuan untuk Memahami dan menggunakan Matematika Dasar dalam Kegiatan Sehari-Hari Dikenal Sebagai Literasi," Vol. 12, No. 3, Pp. 436–446, 2021.
[37] W. M. C. And Hardeli, "Development Of E-Module For Material Classification and Its Changes Based On A Scientific Approach In Grade Seven Of Junior High Schools".
[38] S. Boslaugh, *Statistics In A Nutshell: A Desktop Quick Reference*. " O'reilly Media, Inc.," 2012.
[39] R. Kohavi, "A Study Of Cross-Validation and Bootstrapfor Accuracy Estimation and Model Selection," *Ijcai Int. Jt. Conf. Artif. Intell.*, Vol. 2, No. June, Pp. 1137–1143, 1995.