Jurnal Manajemen Informatika (JAMIKA)
Volume 14 Nomor 1 Edisi April 2024
E ISSN: 2655-6960 | P ISSN: 2088-4125
OJS: https://ojs.unikom.ac.id/index.php/jamika

Prediksi Kelulusan Tepat Waktu Siswa SMK Teknik Komputer Menggunakan Algoritma Random Forest

Prediction of On-Time Graduation for Computer Engineering Vocational School Students Using the Random Forest Algorithm

Arina Fatunnisa[1*], **Hendra Marcos**[2]
Program Studi Informatika, Universitas Amikom Purwokerto, Jawa Tengah, Indonesia
*E-mail: arinafatunnisa300803@gmail.com

Abstrak

Kinerja sekolah dapat diukur melalui tingkat kelulusan siswa, yang merupakan indikator kunci. Tingkat kelulusan yang rendah menunjukkan adanya masalah dalam sistem pendidikan dan pembelajaran, yang memerlukan intervensi tepat waktu untuk mencegah siswa tidak menyelesaikan pendidikannya. Oleh karena itu, memprediksi kelulusan sangat penting bagi sekolah dalam rangka menentukan siswa yang berpotensi tidak menyelesaikan pendidikan, sehingga dapat memberikan bantuan awal guna memperbaiki prestasi akademis mereka. Penelitian ini mendesak karena dengan memahami dan memprediksi kelulusan siswa, sekolah dapat mengalokasikan sumber daya secara lebih efektif untuk mendukung siswa yang berisiko, dengan tujuan akhir meningkatkan tingkat kelulusan dan kinerja sekolah secara keseluruhan. Penelitian ini menggunakan algoritma random forest dengan dataset kelulusan. Pendistribusian pemilihan data latih dan uji menggunakan metode stratified random sampling untuk memastikan representasi yang seimbang dari setiap kelas yang dihasilkan. Model Random Forest berhasil diperoleh melalui pelatihan dan evaluasi model menggunakan data uji, menunjukkan akurasi 1,0 atau setara dengan 100%. Penggunaan algoritma random forest pada dataset kelulusan siswa dapat menjadi pendekatan yang efektif dalam mendukung prediksi kelulusan tepat waktu karena memiliki akurasi yang tinggi dan kemampuan model untuk mengenali baik siswa yang lulus maupun tidak lulus.

Kata kunci: *Prediksi kelulusan, SMK Teknik Komputer MBM Rawalo, Algoritma Random Forest, Dataset kelulusan, Pembagian data latih dan uji*

Abstract

School performance can be measured through student completion rates, which is a key indicator. Low graduation rates indicate problems in the education and learning system, which require timely intervention to prevent students from not completing their education. Therefore, predicting graduation rates is crucial for schools in order to determine students who are likely to not complete their education, so as to provide early assistance to improve their academic performance. This research is urgent because by understanding and predicting student completion, schools can allocate resources more effectively to support at-risk students, with the ultimate goal of improving graduation rates and overall school performance. This research uses random forest algorithm with graduation dataset. The distribution of training and test data selection uses stratified random sampling method to ensure a balanced representation of each class generated. The Random Forest model was successfully obtained through training and model evaluation using test data, showing an accuracy of 1.0 or equivalent to 100%. The use of the random forest algorithm on student graduation datasets can be an effective approach in supporting timely graduation prediction due to its high accuracy and the model's ability to recognize both graduating and non-passing students.

Keywords: *Graduation prediction, MBM Rawalo Computer Engineering Vocational School, Random Forest Algorithm, Graduation dataset, Division of training and test data*

Naskah diterima 15 Jan. 2024; direvisi 25 Mar. 2024; dipublikasikan 05 Apr. 2024.

I. PENDAHULUAN

Pendidikan memiliki peran sentral dalam pembangunan masyarakat dan negara, menjadikannya elemen kunci dalam mencapai kemajuan. Salah satu parameter vital dalam menilai keberhasilan sistem pendidikan adalah tingkat kelulusan siswa tepat waktu [1]. Di berbagai negara, tingkat kelulusan tepat waktu di sekolah menengah kejuruan (SMK) menjadi fokus utama karena memiliki dampak langsung pada kesiapan siswa untuk memasuki dunia kerja atau melanjutkan pendidikan ke jenjang perguruan tinggi [2]. Namun, tantangan nyata muncul ketika menghadapi tingginya angka siswa yang tidak dapat menyelesaikan program pendidikan dalam

waktu yang telah ditentukan. Kelulusan tepat waktu di SMK bukan hanya sebuah ukuran performa, tetapi juga mencerminkan efektivitas keseluruhan sistem pendidikan [3]. Ada banyak faktor yang dapat mempengaruhi kelulusan tepat waktu, termasuk faktor-faktor sosial, ekonomi, dan akademik. Dinamika kompleks ini dapat menciptakan kendala bagi sejumlah siswa, menghambat kemampuan mereka untuk menyelesaikan program pendidikan sesuai dengan jadwal yang telah ditetapkan. Akibatnya, perkembangan karir mereka dapat terganggu [4]. Pada tahun 2021, Kementerian Pendidikan dan Kebudayaan Republik Indonesia melaksanakan sebuah studi yang mengungkapkan bahwa hanya 65% siswa SMK yang berhasil lulus tepat waktu [5]. Angka ini menyoroti urgensi perlunya analisis dan prediksi lebih mendalam terkait faktor-faktor yang memengaruhi kelulusan tepat waktu siswa SMK. Dengan memahami secara lebih rinci dinamika ini, sistem pendidikan dapat mengambil langkah-langkah strategis untuk meningkatkan tingkat kelulusan tepat waktu dan dengan demikian memberikan kontribusi yang lebih efektif terhadap perkembangan individu dan kemajuan masyarakat secara keseluruhan.

Penelitian sebelumnya telah dengan tekun mencoba mengidentifikasi dan mengungkap faktor-faktor yang memberikan pengaruh terhadap kelulusan tepat waktu siswa SMK. Contoh penelitian yang signifikan adalah karya Wijayanto, yang menyoroti bahwa kurangnya dukungan sosial dan ekonomi dapat menjadi hambatan yang betul-betul signifikan bagi siswa SMK dalam mengejar kelulusan tepat waktu [6]. Penelitian lain oleh Saraswati menunjukkan bahwa faktor akademik, seperti prestasi belajar dan tingkat kehadiran, juga memiliki peran sentral dalam menentukan kelulusan tepat waktu [7]. Lebih jauh lagi, dunia penelitian telah mengintegrasikan teknologi pembelajaran mesin untuk meningkatkan pemahaman dan prediksi kelulusan tepat waktu siswa SMK. Salah satu pendekatan yang menonjol adalah penerapan algoritma random forest, yang terbukti mampu mengatasi kompleksitas prediksi berdasarkan berbagai atribut yang memengaruhi kelulusan tepat waktu [8]. Algoritma ini melibatkan penggunaan data yang komprehensif, mencakup faktor-faktor sosial, ekonomi, dan akademik siswa. Data tersebut kemudian diolah sebagai input untuk algoritma random forest, yang secara efektif menciptakan model prediktif. Model ini dirancang untuk mengidentifikasi siswa yang memiliki risiko tinggi untuk tidak lulus tepat waktu, memungkinkan lembaga pendidikan untuk mengambil langkah-langkah intervensi yang sesuai dan membantu siswa dalam mencapai kelulusan yang diinginkan [9]. Penerapan teknologi ini tidak hanya mewakili kemajuan dalam metodologi penelitian, tetapi juga membuka pintu untuk pendekatan yang lebih personal dan terfokus dalam membantu setiap siswa mencapai potensinya. Dengan terus mengembangkan dan meningkatkan teknologi prediksi kelulusan, kita dapat membangun sistem pendidikan yang lebih adaptif dan responsif terhadap kebutuhan unik setiap siswa, menjembatani kesenjangan kelulusan tepat waktu dan menciptakan lingkungan belajar yang inklusif.

Algoritma random forest, sebagai metode klasifikasi yang digunakan untuk mengidentifikasi siswa yang berisiko tinggi tidak lulus tepat waktu, membawa inovasi signifikan dalam analisis prediktif di bidang pendidikan [10]. Prosesnya melibatkan pembuatan pohon keputusan secara acak menggunakan subset acak dari data pelatihan, dan prediksi dari masing-masing pohon diintegrasikan sebagai suara mayoritas [11]. Dengan demikian, melalui langkah-langkah klasifikasi ini, model prediktif dapat secara akurat mengidentifikasi siswa yang memiliki risiko tinggi untuk tidak lulus tepat waktu. Penerapan algoritma random forest membuka pintu bagi pihak pendidikan untuk mengambil tindakan intervensi yang lebih spesifik dan terarah [12]. Identifikasi siswa yang berisiko tinggi melalui model ini memberikan kesempatan untuk memberikan bantuan ekstra kepada mereka yang memerlukan, memungkinkan lembaga pendidikan untuk merancang strategi pencegahan yang lebih efektif guna mengurangi tingkat ketidaklulusan tepat waktu di tingkat SMK. Inisiatif ini menjadi langkah proaktif dalam mendukung perkembangan akademis siswa dan memastikan bahwa setiap individu memiliki kesempatan yang sama untuk berhasil. Pentingnya kontribusi algoritma random forest dalam meningkatkan efisiensi dan keberhasilan sistem pendidikan di tingkat sekolah menengah kejuruan tidak dapat diabaikan [13]. Dengan memanfaatkan kekuatan analisis prediktif ini, lembaga pendidikan dapat lebih cermat dan responsif dalam mengatasi tantangan kelulusan tepat waktu. Melalui pemanfaatan teknologi ini, kita dapat membentuk lingkungan pembelajaran yang lebih adaptif dan mendukung, membantu siswa untuk mencapai potensi maksimal mereka, dan pada akhirnya, membangun fondasi masyarakat yang lebih terdidik dan produktif.

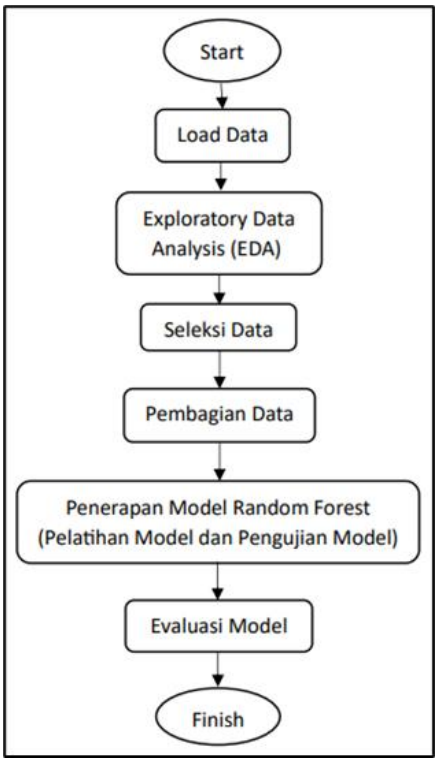
Pemilihan metode algoritma Random Forest sebagai pendekatan prediktif untuk kelulusan tepat waktu siswa SMK berdasarkan rekam jejak keberhasilannya dalam berbagai konteks prediksi, terutama di bidang pendidikan, merupakan inti dari penelitian ini [6]. Keunggulan utama dari algoritma ini adalah kemampuannya untuk mengatasi data yang kompleks dan menghasilkan prediksi yang akurat, memungkinkan kita untuk mengidentifikasi pola dan variabel signifikan yang mempengaruhi kelulusan [11]. Penelitian ini menghadirkan inovasi dengan fokus khusus pada penggunaan algoritma Random Forest dalam konteks pendidikan kejuruan,

Jurnal Manajemen Informatika (JAMIKA)
Volume 14 Nomor 1 Edisi April 2024
E ISSN: 2655-6960 | P ISSN: 2088-4125
OJS: https://ojs.unikom.ac.id/index.php/jamika

sebuah area yang relatif kurang dieksplorasi dibandingkan dengan pendidikan umum atau pendidikan tinggi, menawarkan pendekatan baru dalam mengidentifikasi siswa yang berisiko dan merancang strategi intervensi yang disesuaikan. Melalui pemahaman mendalam tentang variabel-variabel yang memengaruhi kelulusan tepat waktu siswa SMK dan penggunaan data untuk merancang strategi intervensi yang lebih efektif, penelitian ini bertujuan memberikan dukungan yang lebih tepat sasaran kepada siswa, membantu mereka mengatasi hambatan yang mungkin menghambat proses kelulusan tepat waktu. Dengan demikian, penelitian ini tidak hanya berkontribusi pada literatur akademis tetapi juga berdampak nyata dalam meningkatkan efektivitas sistem pendidikan di tingkat sekolah menengah kejuruan, membentuk generasi pelajar yang lebih terampil dan siap bersaing di dunia kerja.

II. METODE PENELITIAN

Penelitian ini melibatkan sejumlah fase penelitian yang dirancang dengan cermat untuk memperoleh hasil yang akurat dan relevan. Urutan tahap-tahap penelitian tersebut mencerminkan alur sistematis yang diikuti selama proses penelitian.



Gambar 1. Alur Penelitian

Gambar 1 memvisualisasikan dengan jelas dan sistematis tahapan-tahapan dalam penelitian ini, membantu dalam memahami alur kerja secara terinci. Tahap pertama dalam gambar tersebut adalah "load data," yang menunjukkan proses pengumpulan dan impor data ke dalam sistem. Ini merupakan langkah awal penting, di mana data yang relevan dengan penelitian ini diambil dan disiapkan untuk analisis selanjutnya. Setelah data ter-load, langkah berikutnya adalah "eda" atau Exploratory Data Analysis. Tahap ini mencakup eksplorasi mendalam terhadap karakteristik dan distribusi data, membantu peneliti memahami pola yang mungkin muncul sebelum penerapan model. Selanjutnya, terdapat fase "seleksi data," yang menunjukkan proses penentuan variabel-variabel yang akan diikutsertakan dalam analisis. Seleksi ini didasarkan pada kepentingan variabel-variabel tersebut terhadap kelulusan tepat waktu siswa SMK. Setelah seleksi data, langkah berikutnya adalah "pembagian data," di mana dataset dibagi menjadi subset pelatihan (training set) dan subset pengujian (test set). Pembagian ini penting untuk menguji kinerja model pada data yang belum pernah dilihat sebelumnya, menghindari overfitting. Fase berikutnya, "penerapan model random forest," terbagi menjadi dua komponen utama: "pelatihan model" dan "pengujian model." Selama pelatihan model, algoritma random forest mengenali pola dalam data pelatihan, sedangkan pengujian model menguji

keberlanjutan model pada data yang belum pernah dilihat sebelumnya. Setelah itu, gambar menunjukkan "evaluasi model," yang melibatkan analisis kritis terhadap kinerja model. Evaluasi ini dapat mencakup metrik seperti akurasi, presisi, dan recall untuk mengevaluasi sejauh mana model dapat memprediksi kelulusan tepat waktu dengan tepat. Penjelasan ini, yang didasarkan pada flowchart pada gambar 1, memberikan pandangan yang rinci dan terstruktur tentang setiap langkah dalam penelitian. Alur kerja yang terdefinisi dengan baik ini membantu memastikan bahwa setiap tahap dilakukan dengan cermat dan meminimalkan potensi kesalahan, sehingga menghasilkan temuan dan kesimpulan penelitian yang lebih andal dan berarti.

Load Data

Istilah 'pemuatan data' merujuk pada tahap kritis dalam pengembangan perangkat lunak dan analisis data. Proses ini merupakan fondasi bagi penelitian data yang efektif dan pembangunan model analitis yang akurat. Pemuatan data melibatkan pengambilan atau pembacaan kumpulan data dan informasi dari berbagai sumber, seperti file lokal, basis data, dan antarmuka pemrograman aplikasi (API) web, lalu memasukkannya ke dalam program atau sistem yang akan digunakan [7]. Dalam konteks penelitian ini, pemuatan data menjadi langkah awal yang sangat penting, karena menentukan kualitas data yang akan digunakan dalam pengembangan model prediktif. Fungsi utama dari proses pemuatan data adalah menyediakan data yang diperlukan untuk analisis lebih lanjut atau untuk melatih model pembelajaran mesin, seperti algoritma random forest yang digunakan dalam penelitian ini. Pemahaman yang mendalam terhadap sumber data yang beragam memungkinkan peneliti untuk mengakses dan mengelola informasi yang relevan, menciptakan landasan yang kuat untuk analisis selanjutnya. Proses ini tidak hanya mencakup pengambilan data tetapi juga mencakup pembersihan data (data cleaning) dan transformasi data (data transformation) jika diperlukan untuk memastikan keakuratan dan konsistensi dataset yang akan digunakan dalam pengembangan model [14]. Pentingnya proses pemuatan data dalam konteks penelitian ini dapat dilihat dari gambar 1, di mana langkah "load data" menjadi awal dari alur kerja yang komprehensif. Dengan memahami dan menguasai tahap pemuatan data dengan baik, penelitian ini dapat memastikan bahwa data yang diolah dan dianalisis selanjutnya memiliki kualitas yang optimal, memberikan dasar yang solid untuk pengembangan model prediktif yang akurat dan relevan.

Exploratory Data Analysis (EDA)

Analisis Data Eksplorasi (EDA) mencakup serangkaian teknik dan metode analisis awal yang ditujukan untuk memperoleh pemahaman yang mendalam tentang struktur dan properti suatu kumpulan data [15] Salah satu tujuan utama EDA adalah menemukan pola-pola penting dalam data. Ini mencakup identifikasi distribusi, hubungan antarvariabel, outlier, dan tren yang dapat memberikan wawasan penting terkait dengan fenomena yang sedang diamati. Dengan melakukan EDA dengan seksama, peneliti dapat mengidentifikasi potensi asosiasi dan korelasi antarvariabel, yang dapat membimbing tahap selanjutnya dalam analisis data. [16]. Selain itu, hasil dari EDA memberikan landasan yang kuat untuk pengembangan model analitis dan pengambilan keputusan tingkat lanjut. Informasi yang ditemukan selama proses ini dapat membentuk dasar bagi pemilihan atribut yang relevan untuk model prediktif, membantu dalam pemilihan metode analisis yang sesuai, dan memberikan wawasan yang mendalam untuk merumuskan pertanyaan penelitian yang lebih tajam.

Deskripsi Variabel

Gambar 2 merupakan EDA (Exploratory Data Analysis) deskripsi variabel yang digunakan untuk menghitung jumlah data lulus dan tidak lulus pada dataset kelulusan siswa SMK Teknik Komputer MBM Rawalo. Proses EDA ini membantu memberikan gambaran yang lebih rinci tentang distribusi data pada variabel target, yaitu status kelulusan siswa. Dengan menghitung jumlah data yang lulus dan tidak lulus, penelitian ini dapat mengidentifikasi sebaran kelas pada dataset, memahami proporsi kelulusan, dan merinci informasi yang relevan untuk analisis lebih lanjut. Langkah ini sangat penting karena memberikan dasar statistik awal untuk memahami karakteristik data dan distribusi variabel target. Hasilnya dapat memberikan wawasan awal tentang sejauh mana dataset mencerminkan situasi aktual pada SMK Teknik Komputer MBM Rawalo terkait tingkat kelulusan siswa. Informasi ini kemudian dapat menjadi dasar untuk perencanaan dan pemilihan variabel prediktif yang lebih lanjut, serta membantu mengarahkan analisis lanjutan terhadap faktor-faktor yang dapat memengaruhi kelulusan siswa.

EDA merupakan deskripsi variabel yang digunakan untuk mengecek informasi yang terkandung dalam dataset kelulusan siswa. Data ini terdiri dari 8 kolom data objek dan 11 kolom data integer, dengan jumlah keseluruhan data sebanyak 160 data [17]. EDA menjadi langkah awal yang penting untuk memahami struktur

Jurnal Manajemen Informatika (JAMIKA)
Volume 14 Nomor 1 Edisi April 2024
E ISSN: 2655-6960 | P ISSN: 2088-4125
OJS: https://ojs.unikom.ac.id/index.php/jamika

dan karakteristik dataset, membantu identifikasi tipe variabel, serta memastikan konsistensi dan kevalidan data. Selanjutnya, EDA dalam deskripsi variabel juga digunakan untuk mengecek deskripsi statistik data yang ada pada dataset kelulusan siswa. Analisis statistik ini memberikan gambaran yang lebih mendalam tentang distribusi nilai-nilai dalam dataset, mencakup nilai rata-rata, median, kuartil, dan lainnya. Dengan memeriksa statistik deskriptif, peneliti dapat memastikan bahwa data valid dan sesuai dengan ekspektasi. Pentingnya menegaskan bahwa nilai minimum tidak ada yang di bawah 0 dalam dataset menunjukkan bahwa data tersebut konsisten dan sesuai dengan logika konteks pendidikan. Hal ini penting untuk memastikan bahwa data memiliki integritas yang tinggi dan dapat diandalkan dalam pengembangan model machine learning. Dengan demikian, EDA tidak hanya memberikan pemahaman awal tentang distribusi dan karakteristik variabel, tetapi juga memvalidasi keandalan data, memberikan dasar yang kokoh untuk analisis lebih lanjut dan pengembangan model prediktif.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 160 entries, 0 to 159
Data columns (total 20 columns):
 #   Column                     Non-Null Count  Dtype
---  ------                     --------------  -----
 0   NO_ABSEN                   160 non-null    int64
 1   NISN                       160 non-null    int64
 2   NAMA_SISWA                 160 non-null    object
 3   KELAS                      160 non-null    object
 4   JENIS_KELAMIN              160 non-null    object
 5   JURUSAN                    160 non-null    object
 6   PRESENSI                   160 non-null    object
 7   PRESTASI                   160 non-null    object
 8   STATUS_EKONOMI             160 non-null    object
 9   DAFTAR_NILAI_PEN.AGAMA     160 non-null    int64
 10  DAFTAR_NILAI_PPKN          160 non-null    int64
 11  DAFTAR_NILAI_BINDO         160 non-null    int64
 12  DAFTAR_NILAI_B.ING         160 non-null    int64
 13  DAFTAR_NILAI_MATEMATIKA    160 non-null    int64
 14  DAFTAR_NILAI_SENI_BUDAYA   160 non-null    int64
 15  DAFTAR_NILAI_PENJASORKEST  160 non-null    int64
 16  DAFTAR_NILAI_PRODUKTIF     160 non-null    int64
 17  JUMLAH                     160 non-null    int64
 18  RATA-RATA                  160 non-null    float64
 19  STATUS                     160 non-null    object
dtypes: float64(1), int64(11), object(8)
memory usage: 25.1+ KB
```

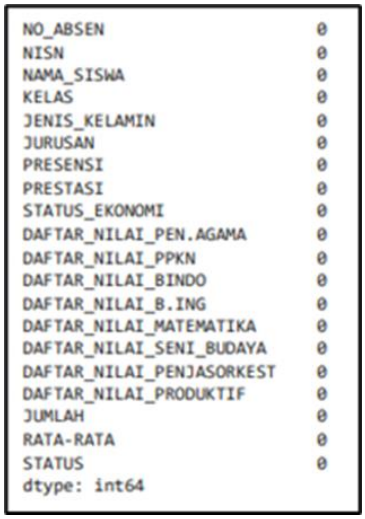
Gambar 2. Informasi Dataset

Mengecek Data pada Missing Value

Gambar 3 merupakan EDA yang digunakan untuk mengecek data pada missing value. Pengecekan terhadap keberadaan nilai yang hilang (missing value) menjadi langkah penting dalam memastikan integritas dan kualitas data sebelum proses analisis lebih lanjut. EDA ini melibatkan identifikasi setiap variabel dan pemeriksaan apakah terdapat nilai yang tidak lengkap atau kosong pada setiap baris data. Proses analisis ini dapat melibatkan teknik visualisasi, seperti pembuatan grafik batang atau heatmap, yang memperlihatkan distribusi missing value pada setiap variabel. Jika terdapat sejumlah besar nilai yang hilang, langkah-langkah selanjutnya dapat mencakup penggunaan metode imputasi atau bahkan pertimbangan untuk menghapus kolom atau baris tertentu dari dataset. Melalui EDA ini, peneliti dapat memahami sejauh mana dataset bersih dari missing value, memberikan gambaran tentang kualitas data yang digunakan dalam analisis, dan memungkinkan penentuan langkah-langkah lanjutan untuk menangani nilai yang hilang dengan tepat. Dengan menggunakan EDA untuk pengecekan missing value, penelitian ini memastikan keakuratan dan kehandalan data sebelum proses lanjutan seperti pemodelan dan evaluasi dilakukan.

Jurnal Manajemen Informatika (JAMIKA)
Volume 14 Nomor 1 Edisi April 2024
E ISSN: 2655-6960 | P ISSN: 2088-4125
OJS: https://ojs.unikom.ac.id/index.php/jamika

Dalam konteks Analisis Data Eksplorasi (EDA), pemeriksaan terhadap keberadaan nilai yang kosong atau missing value menjadi langkah yang sangat penting. Pada tahapan ini, dilakukan penelitian untuk memastikan bahwa dataset kelulusan tidak memiliki nilai yang hilang yang dapat memengaruhi keakuratan dan validitas analisis. EDA untuk mengecek data pada missing value menganalisis setiap variabel yang terlibat dalam penelitian ini dan mengidentifikasi apakah terdapat entri yang kosong atau tidak lengkap. Hasil dari analisis ini menyimpulkan bahwa tidak terdapat data kosong pada dataset penelitian. Kondisi ini sangat menguntungkan karena memastikan integritas dan konsistensi data yang akan digunakan dalam proses analisis lebih lanjut. Kehadiran data yang lengkap memungkinkan peneliti untuk merancang dan menerapkan model prediktif dengan keyakinan yang tinggi, serta meminimalkan risiko bias atau kesalahan hasil yang dapat terjadi akibat keberadaan missing value.



Gambar 3. Data *Missing Value*

Menghitung Jumlah Data Lulus dan Tidak Lulus

Dibawah ini merupakan EDA yang digunakan untuk menghitung jumlah data lulus dan tidak lulus pada dataset kelulusan. Analisis ini melibatkan pemahaman mendalam terhadap distribusi kategori kelulusan siswa dalam dataset. Melalui EDA, peneliti dapat mengidentifikasi proporsi siswa yang lulus dan tidak lulus, memberikan wawasan awal tentang distribusi kelas pada dataset tersebut. Hasil dari perhitungan ini penting untuk memastikan keseimbangan antara kedua kategori kelulusan, yang menjadi faktor kritis dalam pembangunan model prediktif. Selain itu, EDA juga memberikan informasi mengenai variabilitas nilai-nilai pada setiap kategori kelulusan, yang dapat memberikan gambaran tentang sebaran skor siswa. Hal ini berguna untuk mengevaluasi apakah ada perbedaan signifikan dalam distribusi nilai antara siswa yang lulus dan tidak lulus. Dengan demikian, EDA bukan hanya sekadar menghitung jumlah data, tetapi juga memberikan pandangan yang lebih kaya tentang karakteristik dataset yang dapat membantu peneliti dalam pengambilan keputusan selanjutnya terkait dengan pemilihan variabel, pemodelan, dan evaluasi hasil prediksi. Pentingnya EDA dalam menghitung jumlah data lulus dan tidak lulus menjadi langkah awal yang sangat penting dalam memahami landasan data dan memastikan bahwa analisis selanjutnya dapat dilakukan dengan akurat. Dengan informasi ini, peneliti dapat memastikan bahwa model prediktif yang dibangun dapat mengenali dengan baik kedua kelas kelulusan dan memberikan hasil yang lebih dapat diandalkan.

Gambar 4. Jumlah Data Lulus dan Tidak Lulus

Melalui EDA yang telah dilakukan, dapat disimpulkan bahwa dari total 160 data pada dataset kelulusan, sebanyak 152 siswa berhasil lulus tepat waktu, sementara hanya 8 siswa yang tidak lulus. Analisis ini memberikan gambaran yang jelas tentang distribusi kelulusan siswa di dalam dataset. Proporsi yang tinggi dari siswa yang berhasil lulus menunjukkan adanya potensi keberhasilan dalam proses pendidikan di tingkat

sekolah menengah kejuruan. umlah siswa yang tidak lulus yang relatif kecil dapat memberikan pandangan awal bahwa tingkat kelulusan secara keseluruhan cukup baik. Namun, fokus pada siswa yang tidak lulus juga penting untuk mengidentifikasi faktor-faktor penyebab dan merancang strategi intervensi yang lebih efektif. Selain itu, rasio antara jumlah siswa yang lulus dan tidak lulus memberikan gambaran awal tentang sejauh mana keberhasilan sistem pendidikan dalam mendukung kesuksesan siswa.

Seleksi Data

Seleksi data menjadi tahap krusial dalam penggunaan algoritma Random Forest untuk prediksi kelulusan siswa SMK Teknik Komputer MBM Rawalo. Proses analisis dimulai dengan EDA, yang memberikan pemahaman mendalam tentang karakteristik variabel dan distribusi data yang terlibat. EDA menjadi langkah awal yang penting dalam pemahaman terhadap dataset, membuka jalan untuk mengeksplorasi pola-pola yang mungkin muncul. Dalam konteks ini, EDA tidak hanya memberikan gambaran visual tentang data tetapi juga membimbing seleksi variabel yang relevan untuk dimasukkan dalam model prediktif. Proses ini menjadi kunci karena memastikan bahwa hanya variabel-variabel yang paling berpengaruh terhadap prediksi kelulusan siswa yang diikutsertakan, mengoptimalkan kinerja model dan mencegah overfitting. Langkah selanjutnya setelah EDA adalah melakukan korelasi variabel. Proses ini penting untuk mengidentifikasi hubungan yang signifikan antara variabel-variabel tertentu dengan variabel target, yaitu status kelulusan. Korelasi ini membantu menilai sejauh mana variabel-variabel tersebut dapat memberikan kontribusi dalam memprediksi kelulusan siswa. Variabel-variabel dengan korelasi tinggi dapat dianggap sebagai prediktor yang kuat dan relevan dalam membantu algoritma Random Forest membuat keputusan yang akurat. Dengan demikian, seleksi data ini bukan hanya tentang memilih variabel secara sembarang, tetapi melibatkan pemahaman mendalam tentang karakteristik dataset dan dampaknya terhadap tujuan prediksi. Keseluruhan proses ini menjadi landasan yang kokoh untuk membangun model prediktif yang dapat memberikan hasil yang akurat dan dapat diandalkan.

Pembagian Data

Pentingnya membagi data menjadi set latih dan uji terletak pada proses pemodelan machine learning [18]. Set latih memiliki peran krusial sebagai sumber untuk melatih model, menjadi dasar utama dalam pembentukan model tersebut. Setelah itu, data uji dimanfaatkan untuk mengevaluasi kinerja model. Penggunaan data uji menjadi penting karena memungkinkan model untuk dinilai kinerjanya pada data yang belum pernah diakses sebelumnya [19]. Penjelasan tentang pemanfaatan fungsi train_test_split pada pustaka sklearn.model_selection untuk melakukan pembagian dataset menjadi dua bagian, yaitu data latihan dan data uji. Contoh yang diberikan menetapkan proporsi pembagian data sebesar 80% untuk data latih dan 20% untuk data uji. Pembagian dilaksanakan secara acak dan proporsional, dengan menerapkan stratified random sampling. Pendekatan ini dilakukan untuk memastikan representasi yang seimbang antara kedua set data latih dan uji, kelas outcome yang dimaksud.

Penerapan Model Random Forest

Setelah berhasil membangun model Random Forest, langkah berikutnya adalah mengevaluasi performa model menggunakan data uji. Seperti yang telah dijelaskan sebelumnya, penggunaan data uji yang tidak pernah diakses oleh model selama proses pelatihan bertujuan untuk memastikan evaluasi yang lebih obyektif dan mencegah overfitting [20]. Prediksi hasil pada data uji kemudian dibandingkan dengan nilai sebenarnya untuk menilai seberapa baik model dapat memprediksi data tersebut [21]. Variabel y_pred digunakan untuk menyimpan prediksi kelas pada setiap observasi dalam data uji secara berurutan [22]. Untuk mengukur efektivitas model dalam melakukan prediksi, perbandingan dilakukan dengan label kelas sebenarnya yang tersimpan dalam variabel y_test. Proses evaluasi ini dilakukan dengan menghitung berbagai metrik kinerja. Pengukuran akurasi, yang mengindikasikan proporsi prediksi yang tepat terhadap total jumlah observasi, memberikan gambaran umum tentang performa model. Presisi, yang mengukur proporsi prediksi positif yang benar-benar positif, penting untuk menilai kualitas prediksi model. Recall, atau sensitivitas, mengindikasikan kemampuan model untuk mengidentifikasi semua kasus positif yang sebenarnya. Skor F1, yang merupakan harmonis rata-rata dari presisi dan recall, memberikan keseimbangan antara kedua metrik tersebut dan seringkali lebih berguna dalam situasi dimana distribusi kelas tidak seimbang. Support mengacu pada jumlah sampel untuk setiap kelas yang sebenarnya, yang membantu memahami seberapa representatif sampel tersebut untuk setiap kelas.

Jurnal JAMIKA Manajemen Informatika
Jurnal Manajemen Informatika (JAMIKA)
Volume 14 Nomor 1 Edisi April 2024
E ISSN: 2655-6960 | P ISSN: 2088-4125
OJS: https://ojs.unikom.ac.id/index.php/jamika

III. HASIL DAN PEMBAHASAN

Dataset yang diterapkan dalam penelitian ini merupakan kumpulan data kelulusan yang terdiri dari 160 entri dan mencakup 19 atribut, menunjukkan kompleksitas yang tidak dapat efektif diolah hanya dengan menggunakan tools seperti MS Excel/spreadsheet. Alat seperti excel mungkin memadai untuk analisis data dasar dan tugas pengolahan data sederhana, tetapi kekurangannya menjadi jelas ketika melakukan proses pemuatan data, EDA (Exploratory Data Analysis), dan pembagian data yang merupakan tahapan penting dalam pembentukan model machine learning. Misalnya, ketika melakukan EDA, diperlukan langkah seperti deskripsi variabel, pengecekan deskripsi statistik, analisis data missing value, dan penghitungan jumlah data lulus dan tidak lulus, yang melibatkan manipulasi dan analisis data yang kompleks dan detail. Selanjutnya, proses seleksi data untuk prediksi dan pembagian dataset menjadi data latih dan uji, yang memiliki peran krusial, menuntut kapabilitas yang melampaui batas alat spreadsheet tradisional. Penggunaan fungsi train_test_split dari pustaka sklearn.model_selection untuk membagi dataset secara acak dan proporsional (stratified random sampling) menegaskan perlunya alat analisis data yang lebih canggih. Keterbatasan Excel dalam mengelola proses pembagian data dan evaluasi model, seperti memastikan representasi yang seimbang antara kelas outcome pada data latih dan uji, serta mengevaluasi model menggunakan data uji untuk menghindari overfitting, menekankan pentingnya alat analitik yang lebih maju. Oleh karena itu, aplikasi seperti Python dan pustakanya yang kaya, termasuk Scikit-Learn untuk machine learning, menjadi solusi yang sangat diperlukan untuk mengatasi tantangan ini, memastikan penelitian dapat dilakukan dengan ketepatan dan efisiensi yang lebih tinggi.. Fungsi `train_test_split` pada pustaka `sklearn.model_selection` dapat digunakan untuk membagi dataset, dengan contoh proporsi pembagian sekitar 80% untuk data latih dan 20% untuk data uji. Proses pembagian ini dilakukan secara acak dan proporsional (stratified random sampling) untuk menjaga representasi yang seimbang antara kedua kelas outcome pada data latih dan uji. Setelah berhasil memperoleh model Random Forest, langkah berikutnya adalah mengevaluasi model menggunakan data uji, sesuai dengan prinsip bahwa penggunaan data uji yang tidak terlibat dalam proses pelatihan memberikan evaluasi yang lebih objektif dan menghindari overfitting. Dibawah ini terdapat tabel yang menunjukkan proses pemuatan data.

TABEL 1
PEMUATAN DATA

No	NISN	Nama Siswa	JK	Jurusan	Jumlah Nilai	Rata-rata	Status
1	1133445566	Divina Rea Egistina	P	TJKT	644	80,5	Lulus
2.	1144556677	Fatar Afril Shahila	L	TJKT	624	78	Lulus
3	1155667788	Fatih Dzaki Ramadhan	L	TKJT	689	86,12	Lulus
4	1166778899	Ibnu Rifki Al Wansyah	L	TJKT	643	80,37	Lulus
5	1177890010	Ibnue Ibrahim Setiaji	L	TKJT	643	80,37	Lulus
...	....	...	...	....	....	....	....
...	....	....	...	...	...	...	....
156	1122334455	Angel Badriyah	P	TJKT	565	73,12	Tidak Lulus
157	1600112228	Rafi Ramdan	L	TJKT	588	73,5	Tidak Lulus
158	1689001116	Ahmad Saefuk Jaiz	L	TJKT	585	73,12	Tidak Lulus
159	2166778889	Siti Atiqoh	P	TO	588	73,5	Tidak Lulus
160	2255667777	Fawaz Prasetyo	L	TO	585	73,12	Tidak Lulus

Pada tabel 1 terdapat 7 variable yang digunakan sebagai atribut, yaitu NISN, nama siswa, jenis kelamin, jurusan, jumlah nilai, rata-rata dan status kelulusan. Pada status kelulusan di ambil dari nilai rata-rata, jika lebih dari 76 maka dinyatakan lulus sebaliknya jika kurang dari 76 maka status kelulusannya tidak lulus

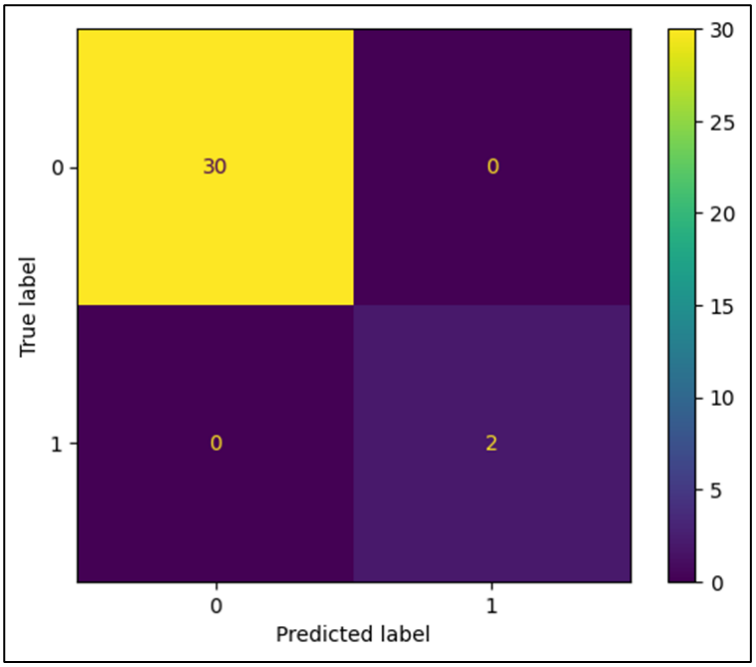
Jurnal Manajemen Informatika (JAMIKA)
Volume 14 Nomor 1 Edisi April 2024
E ISSN: 2655-6960 | P ISSN: 2088-4125
OJS: https://ojs.unikom.ac.id/index.php/jamika

	precision	recall	f1-score	support
0	1.00	1.00	1.00	30
1	1.00	1.00	1.00	2
accuracy			1.00	32
macro avg	1.00	1.00	1.00	32
weighted avg	1.00	1.00	1.00	32

Accuracy Score RandomForestClassifier : 1.0
Precision Score RandomForestClassifier : 1.0
Recall Score RandomForestClassifier : 1.0
F1 Score RandomForestClassifier : 1.0

Gambar 6. akurasi, precision, recall dan F1 Score

Berdasarkan hasil evaluasi menggunakan Gambar 6 di atas, model berhasil mencapai akurasi, precision, recall, dan F1 Score sebesar 1.0. Angka-angka ini menunjukkan bahwa model berhasil memprediksi semua data uji dengan benar, mencapai tingkat keakuratan 100 persen. Akurasi mencerminkan rasio prediksi yang benar secara keseluruhan, sedangkan precision mengukur sejauh mana prediksi positif model benar, recall menilai seberapa baik model dapat menangkap keseluruhan instance yang positif, dan F1 Score merupakan harmonisasi antara precision dan recall.



Gambar 7. *Confussion Matrix*

Berdasarkan confusion matrix dari 30 data uji dengan label 0 (Lulus), dapat disimpulkan bahwa model berhasil melakukan prediksi secara sempurna tanpa adanya kesalahan. Dari total 30 data uji yang merupakan kasus "Lulus", model berhasil mengklasifikasikan semuanya dengan benar, tidak ada prediksi yang salah. Sementara itu, pada 2 data uji yang memiliki label 1 (Tidak Lulus), model juga berhasil memprediksi dengan benar pada semua kasus, tanpa adanya prediksi yang salah. Model berhasil mengidentifikasi keduanya secara tepat, tanpa menghasilkan kesalahan prediksi. Hasil ini menggambarkan kinerja model yang sangat baik dalam membedakan antara kategori "Lulus" dan "Tidak Lulus" pada dataset ini. Kedua nilai pada matriks kebingungan (confusion matrix) yang menunjukkan nol kesalahan prediksi pada kedua kategori menegaskan bahwa model dapat membedakan dengan akurat antara siswa yang lulus dan tidak lulus. Pencapaian ini memberikan keyakinan tambahan tentang keandalan model dalam mengklasifikasikan status kelulusan siswa, dan nilai akurasi, precision, recall, dan F1 Score sebesar 1.0 memberikan dukungan yang kuat terhadap keunggulan model dalam menghadapi tugas klasifikasi pada dataset ini. Meskipun hasil ini sangat positif, perlu

diingat untuk selalu melakukan evaluasi yang komprehensif dan mencakup skenario pengujian yang beragam untuk memastikan generalisasi model pada berbagai kondisi data.

IV. KESIMPULAN

Penelitian mengenai prediksi kelulusan siswa SMK Teknik Komputer MBM Rawalo menggunakan algoritma random forest dengan dataset kelulusan telah memberikan hasil yang sangat positif. Dari tahap awal proses load data hingga evaluasi model dengan data uji, setiap langkah dilakukan dengan cermat. Analisis exploratory data membantu memahami karakteristik dataset, sedangkan pemisahan dataset menjadi data latih dan uji memastikan kehandalan model dalam menghadapi data yang belum pernah dilihat sebelumnya. Pentingnya pembagian dataset menjadi dua bagian, yaitu data latih dan data uji, terletak pada kemampuan model untuk belajar dari data latih dan kemudian diuji pada data uji untuk mengukur kemampuan generalisasi. Pemilihan proporsi pembagian sebesar 80% untuk data latih dan 20% untuk data uji, dengan penerapan stratified random sampling, menciptakan kondisi yang ideal untuk melatih dan menguji model secara efektif. Ini juga membantu menghindari bias yang mungkin muncul jika pembagian dilakukan secara tidak merata, terutama pada kasus kelas kelulusan yang tidak seimbang. Hasil akhir menunjukkan akurasi model sebesar 1.0, dengan model mampu memprediksi kelulusan siswa secara sempurna, seperti yang tercermin dalam confusion matrix yang menunjukkan keberhasilan prediksi untuk kedua kelas. Kesimpulan dari penelitian ini menyiratkan bahwa penggunaan algoritma random forest pada dataset kelulusan siswa dapat menjadi pendekatan yang efektif dalam mendukung prediksi kelulusan tepat waktu. Dengan akurasi yang tinggi dan kemampuan model untuk mengenali baik siswa yang lulus maupun tidak lulus, hasil penelitian ini memberikan kontribusi positif terhadap pemahaman dan implementasi model machine learning dalam konteks pendidikan. Implikasinya dapat merangsang pengembangan metode evaluasi dan intervensi yang lebih canggih untuk meningkatkan efisiensi dan efektivitas sistem pendidikan di masa yang akan datang.

UCAPAN TERIMA KASIH

Ucapan terima kasih yang tulus juga ingin kami sampaikan kepada semua pihak yang turut berperan serta dalam kesuksesan penelitian ini. Terima kasih kepada tim peneliti yang telah berkontribusi secara langsung dalam merancang, menjalankan, dan menganalisis penelitian ini dengan penuh dedikasi. Semua dukungan dan kerjasama yang diberikan oleh berbagai pihak telah memberikan kontribusi positif dalam menjadikan penelitian ini sukses. Terima kasih juga kepada lembaga atau institusi yang telah memberikan dukungan finansial dan fasilitas untuk penelitian ini. Tanpa dukungan tersebut, pencapaian hasil yang memuaskan tidak akan menjadi mungkin. Kami menghargai setiap bentuk dukungan yang diberikan, baik itu dalam bentuk sumber daya, bimbingan, maupun dukungan moral. Selain itu, terima kasih kepada semua responden atau peserta penelitian yang telah bersedia berpartisipasi dan menyediakan data yang sangat berharga bagi kelancaran penelitian ini. Keberhasilan penelitian ini juga merupakan hasil dari partisipasi dan kontribusi mereka. Semoga hasil penelitian ini dapat memberikan manfaat yang berkelanjutan dalam upaya meningkatkan kualitas dan efisiensi pendidikan di masa yang akan datang. Harapan kami agar temuan dan kontribusi dari penelitian ini dapat memberikan pandangan baru, solusi yang inovatif, dan memberikan arahan bagi pengembangan lebih lanjut di bidang pendidikan. Terima kasih sekali lagi kepada semua pihak yang telah berperan serta dalam perjalanan penelitian ini.

DAFTAR PUSTAKA

[1] K. B. Pso, I. Irawan, M. R. Qisthiano, M. Syahril, and P. M. Jakak, “Optimasi Prediksi Kelulusan Tepat Waktu : Studi Perbandingan Algoritma Random Forest dan Algoritma,” vol. 4, no. 4, pp. 26–36, 2023.

[2] G. A. Sandag, “Prediksi Rating Aplikasi App Store Menggunakan Algoritma Random Forest,” *CogITo Smart Journal*, vol. 6, no. 2, pp. 167–178, 2020, doi: 10.31154/cogito.v6i2.270.167-178.

[3] Yoga Religia, Agung Nugroho, and Wahyu Hadikristanto, “Klasifikasi Analisis Perbandingan Algoritma Optimasi pada Random Forest untuk Klasifikasi Data Bank Marketing,” *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 5, no. 1, pp. 187–192, 2021, doi: 10.29207/resti.v5i1.2813.

[4] N. L. Hanun, A. U. Zailani, P. Studi, T. Informatika, and U. Pamulang, “Journal of technology information,” vol. 6, no. 1, pp. 7–14, 2020.

[5] J. Khatib *et al.*, “Indonesian Journal of Computer Science,” vol. 11, no. 1, pp. 1015–1022, 2022.

[6] N. B. Putri and A. W. Wijayanto, "Analisis Komparasi Algoritma Klasifikasi Data Mining Dalam Klasifikasi Website Phishing," *Komputika : Jurnal Sistem Komputer*, vol. 11, no. 1, pp. 59–66, 2022, doi: 10.34010/komputika.v11i1.4350.

[7] N. W. S. Saraswati and N. M. L. Martarini, "Extract Transform Loading Data Absensi Stmik Stikom Indonesia Menggunakan Pentaho," *MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 19, no. 2, pp. 273–281, 2020, doi: 10.30812/matrik.v19i2.564.

[8] F. Fallah, "Hierarchical Quadratic Random Forest Classifier," 2023, [Online]. Available: http://arxiv.org/abs/2306.01893

[9] Oon Wira Yuda, Darmawan Tuti, Lim Sheih Yee, and Susanti, "Penerapan Penerapan Data Mining Untuk Klasifikasi Kelulusan Mahasiswa Tepat Waktu Menggunakan Metode Random Forest," *SATIN - Sains dan Teknologi Informasi*, vol. 8, no. 2, pp. 122–131, 2022, doi: 10.33372/stn.v8i2.885.

[10] J. Zeniarja, A. Salam, and F. A. Ma'ruf, "Seleksi Fitur dan Perbandingan Algoritma Klasifikasi untuk Prediksi Kelulusan Mahasiswa," *Jurnal Rekayasa Elektrika*, vol. 18, no. 2, pp. 102–108, 2022, doi: 10.17529/jre.v18i2.24047.

[11] O. Baines, A. Chung, and R. Raval, "Random forest classification algorithm," *Mathematics Research Journal*, no. April, p. 500, 2020, [Online]. Available: http://orange3.readthedocs.io/en/3.4.0/widgets/classify/randomforest.html

[12] A. Darmawan, I. Yudhisari, A. Anwari, and M. Makruf, "Pola Prediksi Kelulusan Siswa Madrasah Aliyah Swasta dengan Support Vector Machine dan Random Forest," *Jurnal Minfo Polgan*, vol. 12, no. 1, pp. 387–400, 2023, doi: 10.33395/jmp.v12i1.12388.

[13] M. Syukron, R. Santoso, and T. Widiharih, "Perbandingan Metode Smote Random Forest Dan Smote Xgboost Untuk Klasifikasi Tingkat Penyakit Hepatitis C Pada Imbalance Class Data," *Jurnal Gaussian*, vol. 9, no. 3, pp. 227–236, 2020, doi: 10.14710/j.gauss.v9i3.28915.

[14] G. S. Suwardika and I. K. P. Suniantara, "Analisis Random Forest Pada Klasifikasi Cart Ketidaktepatan Waktu Kelulusan Mahasiswa Universitas Terbuka," *BAREKENG: Jurnal Ilmu Matematika dan Terapan*, vol. 13, no. 3, pp. 177–184, 2019, doi: 10.30598/barekengvol13iss3pp177-184ar910.

[15] T. A. Cardona, E. A. Cudney, J. Snyder, and R. W. Hoerl, "Predicting student degree completion using random forest," *ASEE Annual Conference and Exposition, Conference Proceedings*, vol. 2020-June, no. April, 2020, doi: 10.18260/1-2--35074.

[16] J. Yang, S. Devore, D. Hewagallage, P. Miller, Q. X. Ryan, and J. Stewart, "Using machine learning to identify the most at-risk students in physics classes," *Phys Rev Phys Educ Res*, vol. 16, no. 2, p. 20130, 2020, doi: 10.1103/PhysRevPhysEducRes.16.020130.

[17] van S. Plaosan, "Algoritma Random Forest," *Http://Learningbox.Coffeecup.Com/05_2_Randomforest.Html*, vol. 18, no. 1, pp. 10–14, 2019.

[18] M. Nachouki and M. A. Naaj, "Predicting Student Performance to Improve Academic Advising Using the Random Forest Algorithm," *International Journal of Distance Education Technologies*, vol. 20, no. 1, pp. 1–17, 2022, doi: 10.4018/IJDET.296702.

[19] R. Rismayati, I. Ismarmiaty, and S. Hidayat, "Ensemble Implementation for Predicting Student Graduation with Classification Algorithm," *International Journal of Engineering and Computer Science Applications (IJECSA)*, vol. 1, no. 1, pp. 35–42, 2022, doi: 10.30812/ijecsa.v1i1.1805.

[20] M. Nachouki and M. A. Naaj, "Predicting Student Performance to Improve Academic Advising Using the Random Forest Algorithm," *International Journal of Distance Education Technologies*, vol. 20, no. 1, pp. 1–17, 2022, doi: 10.4018/IJDET.296702.

[21] M. A. Yulianto, "Implementasi FIS Sugeno pada Algoritma C4. 5 Berbasis Particle Swarm Optimization (PSO) untuk Prediksi Prestasi Siswam," *JOAIIA: Journal of Artificial Intelligence and Innovative Applications*, vol. 1, no. 1, pp. 12–22, 2020, [Online]. Available: http://www.openjournal.unpam.ac.id/index.php/JOAIIA/article/view/4272

[22] Sudriyanto, R. Rizaldi, and M. Ainun Rofiq Hariri, "Implementasi Algoritme Decision Tree (C4.5) dengan Optimize Weights (PSO) untuk Memprediksi Kelulusan Mahasiswa Tepat Waktu," *Jurnal Informatika Universitas Pamulang*, vol. 6, no. 2, pp. 252–257, 2021, [Online]. Available: http://openjournal.unpam.ac.id/index.php/informatika252