

Optimasi Prediksi Kelulusan Tepat Waktu: Studi Perbandingan Algoritma Random Forest dan Algoritma K-NN Berbasis PSO

Indra Irawan^{1*}, M Riski Qisthiano², Muhammad Syahril³, Pamuji M. Jakak⁴

^{1,4}Informatika, Universitas Nurul Huda, ²Sistem Informasi, Institut Teknologi dan Bisnis Nasional (ITBN), ³Teknik Informatika, Politeknik Sekayu

Penulis Korespondensi: Indrairawan@unuha.ac.id^{*}, thiano72@gmail.com², muhammadsyahril@polsky.ac.id³, jakak@unuha.ac.id⁴

Informasi Artikel:

Received: 10 Agustus 2023, Accepted: 22 Agustus 2023, Published: 1 Oktober 2023

Abstract

Predicting on-time graduation for students involves various measurement techniques, including criteria such as majors, class types, and semester grade achievements. These factors play a crucial role in determining whether students will complete their studies within the designated timeframe. In line with this, a model has been developed to forecast the probability of timely graduation. This model leverages the Random Forest and k-nearest Neighbor (K-NN) algorithms as tools to classify students into appropriate groups. Optimization is uses the Particle Swarm Optimizer (PSO) algorithm to enhance prediction accuracy. The data used originates from alumni of various Universities in Palembang. This model utilizes multiple attributes, such as majors, university origins, class types, and semester grade records up to the fourth semester. Other attributes encompass the year of graduation and year of enrollment. Data management and processing are conducted using Rapidminer. Validation is performed by splitting the dataset into training and testing groups through the split validation method. Based on research and testing, the Random Forest algorithm achieves an accuracy of 95.79% with an Area Under Curve (AUC) of 0.991. After optimization with PSO, accuracy increases to 97.89% with an AUC of 0.993. Meanwhile, the k-NN algorithm achieves an accuracy of 93.49% with an AUC of 0.975; after optimization with PSO, accuracy rises to 96.74% with an AUC of 0.986.

Keywords: Random Forest, K-NN, PSO, Rapidminer, University, Students, Alumni

1. Pendahuluan

Pendidikan merupakan salah satu pilar penting dalam pembangunan suatu bangsa. Kelulusan tepat waktu dari pendidikan tinggi menjadi salah satu indikator kualitas sistem pendidikan suatu negara. Dalam era digital dan global seperti saat ini, upaya untuk memahami faktor-faktor yang memengaruhi kelulusan tepat waktu dan mengembangkan metode prediksi yang akurat menjadi semakin relevan [1]. Prediksi kelulusan tepat waktu memiliki potensi untuk memberikan panduan kepada lembaga pendidikan dalam merancang program-program pendidikan yang lebih efektif, memahami perubahan tren, dan merancang intervensi yang sesuai untuk meningkatkan tingkat kelulusan [2].

Untuk melaksanakan studi ini, diperlukan kumpulan data komprehensif berhubungan dengan

informasi kelulusan mahasiswa. Data ini harus mencakup rincian tentang mahasiswa seperti Nomor Induk Mahasiswa (NIM), nama, bidang studi, fakultas, tahun awal kuliah, dan tahun kelulusan. Dalam konteks ini, dataset contoh diambil dari beberapa lembaga Perguruan Tinggi ternama di kota Palembang. Baik itu dari Universitas, Institut, atau Sekolah Tinggi, data ini menjadi penentu dalam meramalkan tingkat kelulusan tepat waktu mahasiswa. Secara umum, seorang mahasiswa dianggap lulus tepat waktu jika dapat menyelesaikan pendidikannya dalam jangka waktu empat (4) tahun sesuai masa studi yang telah ditentukan. Jika melebihi batas waktu tersebut, mahasiswa tersebut dikategorikan sebagai lulusan di luar waktu yang telah ditetapkan.

Penelitian sebelumnya telah mengeksplorasi berbagai pendekatan prediksi kelulusan tepat waktu menggunakan berbagai teknik seperti algoritma Random Forest dan algoritma K-NN (K-Nearest Neighbors). Namun, akurasi prediksi menjadi faktor penting dalam memilih algoritma yang paling sesuai untuk tujuan ini. Optimasi juga menjadi aspek kritis dalam memastikan bahwa model prediksi yang dihasilkan memiliki kinerja maksimal [3].

Optimasi performa prediksi masih menjadi area yang dapat ditingkatkan. Oleh karena itu, dalam penelitian ini, kami melakukan studi perbandingan antara algoritma Random Forest dan algoritma K-NN berbasis PSO (Particle Swarm Optimization) dalam konteks prediksi kelulusan tepat waktu. PSO merupakan teknik optimasi yang berfokus pada pencarian solusi terbaik dalam ruang pencarian yang kompleks [3].

Dengan mempertimbangkan perkembangan teknologi dan metode-metode baru yang tersedia, penelitian ini bertujuan untuk mengambil langkah lebih maju dalam memahami dan meningkatkan prediksi kelulusan tepat waktu. Diharapkan bahwa hasil dari penelitian ini dapat memberikan wawasan baru tentang pendekatan yang lebih baik untuk prediksi kelulusan serta memberikan kontribusi bagi pengembangan strategi pendidikan yang lebih efektif di masa depan.

2. Tinjauan Literatur

2.1 Data Mining

Data mining adalah hasil perpaduan antara beberapa bidang ilmu yang menggabungkan teknik-teknik dari machine learning, pengenalan pola, statistik, manajemen basis data, serta visualisasi. Tujuan utamanya adalah mengatasi permasalahan dalam mengambil informasi berharga dari kumpulan data besar dalam sebuah database yang signifikan [4].

2.2 Particle Swarm Optimization (PSO)

Particle Swarm Optimization (PSO) adalah metode optimasi global yang diperkenalkan oleh Kennedy dan Eberhart pada tahun 1995, terinspirasi dari penelitian tentang perilaku kawanan burung dan ikan. Setiap partikel dalam PSO memiliki kecepatan yang menggerakkan mereka di ruang pencarian, dan kecepatan tersebut secara dinamis disesuaikan berdasarkan perilaku historis partikel tersebut [5].

Metode Particle Swarm Optimization (PSO) digunakan dalam beragam konteks masalah skala besar dengan tujuan menemukan solusi optimal. Pendekatan ini mencapai hal tersebut dengan mengkomputasi bobot optimal untuk setiap data record dan kemudian menghitung rata-rata dari seluruh sampel [6].

2.3 Algoritma Random Forest

Algoritma *Random Forest* (RF) merupakan suatu pendekatan yang berasal dari *Classification*

dan *Regression Tree* (CART) yang telah dikembangkan. Pengembangan ini melibatkan penerapan seleksi fitur acak dan teknik bagging (*bootstrap aggregating*). Random Forest adalah sebuah klasifikasi yang terdiri dari sekumpulan pohon keputusan $\{h(x, S^b), b = 1, \dots, B\}$ di mana $\{S^b\}$ adalah vektor-vektor acak yang independen dan memiliki distribusi yang identik untuk setiap pohon. Pohon-pohon ini memberikan suara (vote) untuk kelas yang paling umum pada input x [7].

2.3 Algoritma *k*-Nearest Neighbor (K-NN)

K-Nearest Neighbor (K-NN) adalah algoritma untuk mengklasifikasikan data berdasarkan data training yang diperoleh dari tetangga. Metode ini hanya menyimpan vektor-vektor fitur dan klasifikasi dari data yang digunakan untuk pelatihan. Pada tahap klasifikasi, atribut-atribut yang sama dihitung untuk data uji (yang kelasnya belum diketahui). Jarak antara vektor baru ini dan semua vektor dalam data pelatihan dihitung, dan dipilih sejumlah k vektor terdekat. Label untuk vektor baru ditentukan berdasarkan mayoritas label dari vektor-vektor tersebut. Nilai optimal untuk k dalam algoritme ini bergantung pada karakteristik data. Umumnya, nilai k yang besar akan mengurangi dampak noise dalam klasifikasi, tetapi dapat menyebabkan batas antara label menjadi kurang jelas. Nilai k yang sesuai dapat dipilih melalui optimasi parameter, contohnya melalui proses validasi silang [8].

2.3 Penelitian Terdahulu

Penelitian oleh Yuda melakukan analisis prediksi kelulusan tepat waktu menggunakan algoritma Random Forest. Hasil penelitian tersebut menunjukkan akurasi sebesar 98%, mengungkapkan potensi algoritma Random Forest dalam meramalkan kelulusan mahasiswa dengan baik [9]. Penelitian dengan judul Klasifikasi Ketepatan Lama Studi Mahasiswa Dengan Algoritma Random Forest Dan Gradient Boosting dengan Hasil uji coba sampel pada 590 data, algoritma random forest 10 k-fold mendapatkan akurasi 82,64% [10]. Di sisi lain, penelitian oleh Nasution menggunakan pendekatan algoritma K-NN untuk prediksi kelulusan tepat waktu dengan atribut nilai bahasa Inggris, nilai cekak logika dan mencapai akurasi sebesar 99,5% [11].

Penelitian lain dengan Judul Komparasi Metode K-Nearest Neighbor Dan Neural Network Untuk Menentukan Tingkat Akurasi Dalam Prediksi Kelulusan Mahasiswa menghasilkan akurasi model K-NN 57,89% [12]. Penelitian yang lain melaksanakan penerapan teknik data mining untuk meramalkan durasi studi mahasiswa dengan menerapkan pendekatan Algoritma *K-Nearest Neighbor* (KNN). Dalam penelitiannya, atribut yang dipertimbangkan adalah prestasi akademik dari semester 1 hingga 4, serta total skor kredit semester 1 hingga semester 4. Hasil dari penelitian ini menunjukkan tingkat keakuratan sekitar 75% dalam meramalkan masa studi mahasiswa [13].

Pada penelitian yang dilakukan oleh Ahmad Fauzi, dengan judul Optimasi Algoritma Klasifikasi *Naive Bayes*, *Decision Tree*, *K – Nearest Neighbor*, dan *Random Forest* menggunakan Algoritma *Particle Swarm Optimization* pada Diabetes Dataset menghasilkan akurasi Algoritma *K-Nearest Neighbor* sebelum optimasi 85.99% dan setelah optimasi 94.51% [3].

3. Metode Penelitian

Tahapan yang dilakukan pada penelitian ini meliputi penentuan dataset sampai dengan pengujian data training dan testing tingkat kelulusan mahasiswa dengan klasifikasi tepat waktu

dan tidak tepat waktu.

3.1 Jenis Penelitian

Penelitian ini bertujuan untuk menguji dampak dari algoritma optimasi *Particel Swarm Optimizer (PSO)* pada peningkatan kinerja algoritma *Random Forest* dan K-NN. Hasil penelitian akan dibandingkan untuk melihat tingkat akurasi tertinggi dari masing-masing metode yang digunakan.

3.2 Dataset

Untuk mendukung penelitian ini, diperlukan data penelitian yang terdiri dari sampel data, teknik pengumpulan data, jenis dan sumber data, serta analisis data. Metode pengumpulan data terdiri dari dua teknik berikut:

1. Pengambilan data dari beberapa Universitas, Institut, atau Sekolah Tinggi.
2. Studi Literatur (Literature Research): Penelitian dilakukan dengan mempelajari buku, jurnal, dan bahan bacaan lainnya yang relevan dengan judul penelitian.

Tabel 1. Dataset

Tepat Waktu	Tidak Tepat Waktu	Total
909	830	1739

3.2 Metode Analisis dan Pengujian Data

Dalam penelitian ini, data alumni digunakan untuk melakukan analisis guna mengetahui tingkat keberhasilan mahasiswa dalam menyelesaikan studi tepat waktu. Data tersebut telah diambil dari beberapa Universitas dan institusi pendidikan, kemudian diolah dan diuji menggunakan algoritma K-NN dan *Random Forest* dengan optimasi PSO.

Tabel 2. Jumlah Sampel Berdasarkan PT

No	Perguruan Tinggi	Jumlah Data
1	UIN Raden Fatah Palembang	513
2	Universitas Bina Darma Palembang	700
3	Universitas PGRI Palembang	90
4	STISIPOL Candradimuka Palembang	436

Kinerja klasifikasi dievaluasi menggunakan beberapa metrik seperti akurasi, error rate, specificity, sensitivity, dan presisi. Penelitian ini mengadopsi *metode Cross-Standard Industry for Data Mining (CRISP-DM)* sebagai kerangka kerja, yang terdiri dari 6 tahap

proses sebagai berikut:

a. *Business Understanding*:

Penelitian ini bertujuan untuk menguji pengaruh algoritma optimasi Particle Swarm Optimizer (PSO) terhadap peningkatan performa algoritma Random Forest dan K-NN dalam hal akurasi dan AUC.

b. *Data Understanding*:

Pada tahap *Data Understanding*, dilakukan pengumpulan data dari beberapa Universitas, Institut, atau Sekolah Tinggi di Kota Palembang.

c. *Data Preparation:*

Pada tahap ini, dilakukan persiapan data menggunakan program komputer Jupyter Notebook untuk mengolah data yang ada. Sedangkan dalam pemodelan, digunakan software Rapidminer. Dataset yang digunakan terdiri dari 1739 data dengan 10 atribut dan 1 atribut target.

d. *Modelling:*

Proses optimasi menggunakan algoritma PSO dengan menerapkan algoritma K-NN dan *Random Forest*. Penelitian ini akan membandingkan performa algoritma K-NN dan Random Forest setelah dioptimasi dengan menggunakan PSO.

e. *Evaluation:*

Pada fase ini, akan dilakukan evaluasi terhadap hasil akurasi dari algoritma yang telah dioptimasi. Evaluasi ini akan melibatkan confusion matrix dan Grafik AUC untuk menentukan nilai dari model yang digunakan dan mengukur kesesuaian model dengan tujuan awal penelitian.

f. *Deployment:*

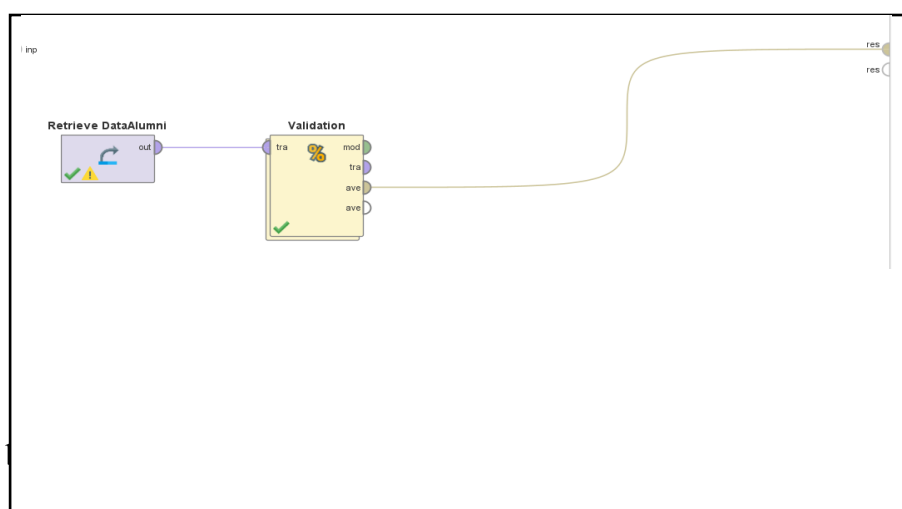
Hasil dari penerapan model yang dilakukan dengan algoritma klasifikasi Random Forest dan K-NN yang telah dioptimasi dengan PSO akan digunakan untuk menentukan tingkat kelulusan tepat waktu mahasiswa serta menentukan metode yang memiliki tingkat akurasi paling baik.

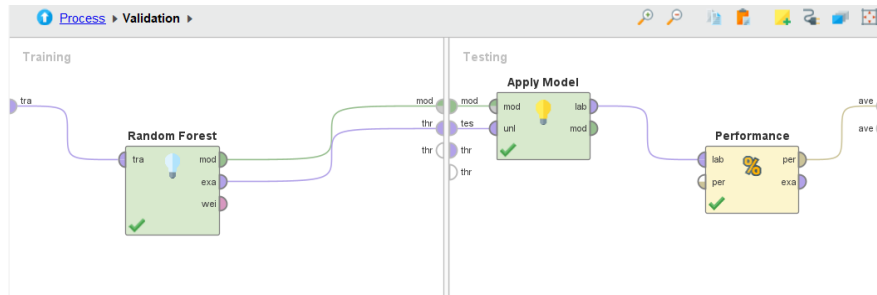
4. Hasil dan Pembahasan

Penelitian ini bertujuan untuk menciptakan sebuah model yang menggunakan algoritma K-NN dan Random Forest yang dioptimasi dengan PSO. Model ini akan memungkinkan untuk memprediksi data baru berdasarkan data yang telah ada dan diproses, dengan tujuan mencapai kinerja algoritma yang memiliki akurasi tertinggi.

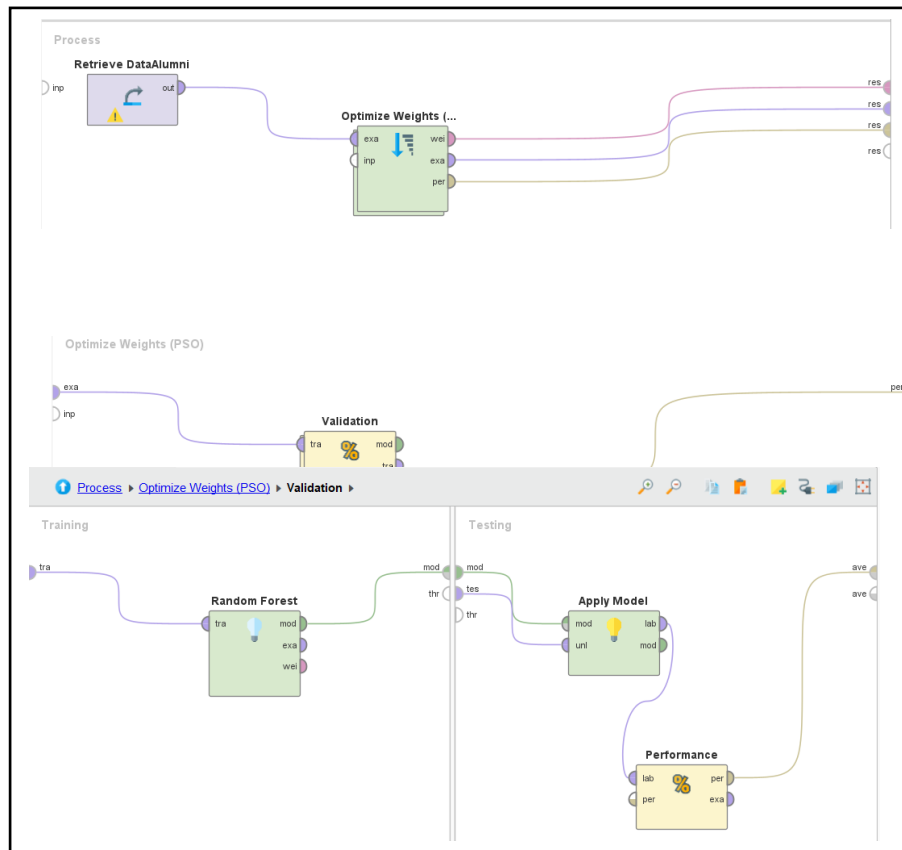
4.1 Proses Pemodelan

Setelah melalui proses data preparation tahap selanjutnya adalah pemodelan. Pada penelitian ini, digunakan dua algoritma yaitu K-Nearest Neighbor (K-NN) dan Random Forest yang dioptimasi dengan menggunakan *Particle Swarm Optimization (PSO)*. Pengoptimalan ini dilakukan menggunakan *RapidMiner Studio* sebagai platform analisis dan pemodelan data.



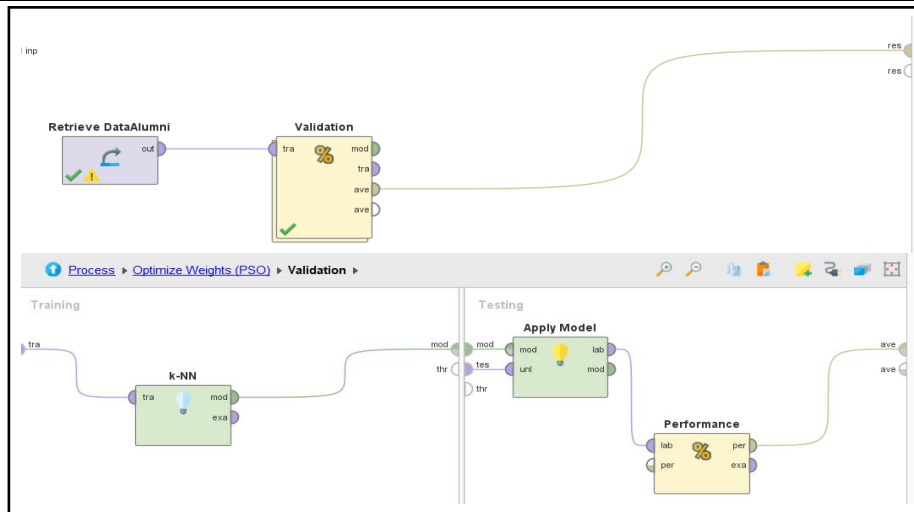


Gambar 1. Model Pengujian Algoritma *Random Forest*

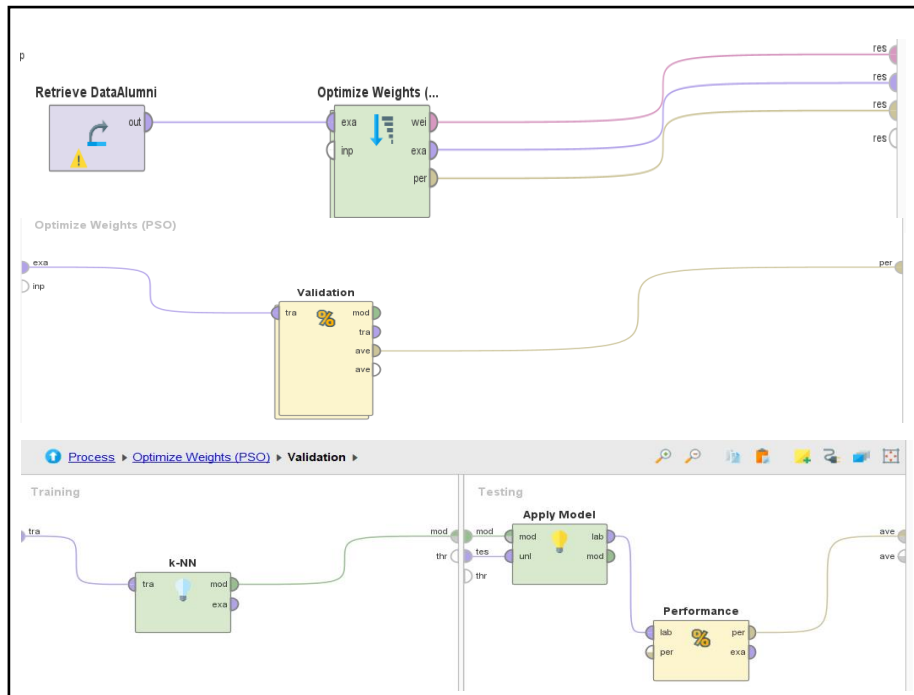


Gambar 2. Model Pengujian Algoritma *Random Forest* Berbasis PSO

Hasil pengujian algoritma *Random Forest* berbasis PSO pada Gambar 2 menunjukkan nilai akurasi sebesar 97.89% dan AUC = 0.993 dengan menggunakan metode split validation, yang membagi data menjadi 70% data training dan 30% data testing. Di sisi lain, ketika algoritma *Random Forest* diuji secara terpisah tanpa menggunakan PSO, menghasilkan akurasi sebesar 95.79% dan AUC 0.991.



Gambar 3. Model Pengujian Algoritma K-NN



Gambar 4. Model Pengujian Algoritma K-NN Berbasis PSO

Hasil Pengujian algoritma K-NN berbasis PSO pada Gambar 4 menunjukkan hasil akurasi sebesar 96.74% dan AUC = 0.986 ketika menggunakan nilai $k=3$ dan population size = 5 dengan metode split validation, yang membagi data menjadi 70% data training dan 30% data testing. Di sisi lain, ketika algoritma Random Forest diuji secara terpisah tanpa menggunakan PSO, menghasilkan akurasi sebesar 93.49% dan AUC 0.975.

4.2 Proses Evaluasi

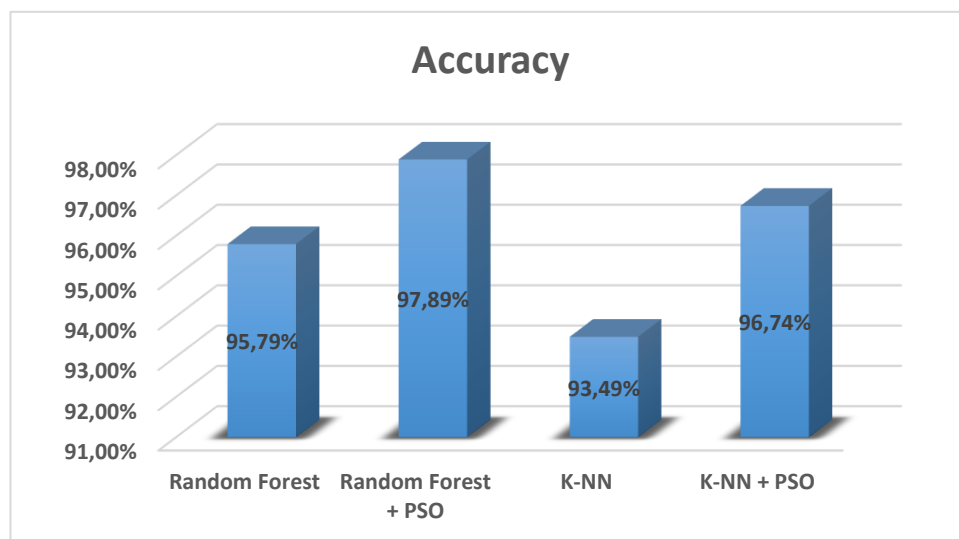
Diperlukan sebuah evaluasi untuk melihat hasil dari validasi akurasi dan AUC. Di bawah ini adalah perbandingan antara nilai Accuracy dan AUC:

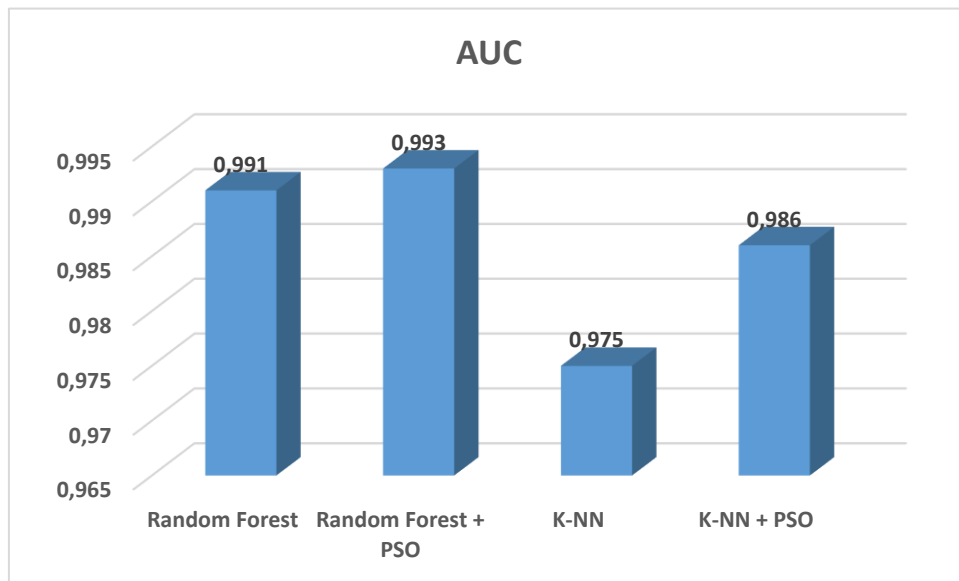
Tabel 3. Komparasi Accuracy dan AUC

Algoritma	Accuracy	AUC
<i>Random Forest</i>	95.79%	0.991
<i>Random Forest + PSO</i>	97.89%	0.993
K-NN	93.49%	0.975
K-NN + PSO	96.74%	0.986

Dalam penelitian ini, hasil perhitungan menggunakan metode Random Forest menunjukkan nilai Accuracy sebesar 95.79% dan AUC sebesar 0.991, sedangkan ketika metode Random Forest+PSO digunakan, menghasilkan nilai Accuracy sebesar 97.89% dan AUC sebesar 0.993. Selain itu, pengujian juga mencakup metode k-NN dan k-NN berbasis PSO. Hasil perhitungan dari pengujian dengan metode k-NN menunjukkan nilai Accuracy sebesar 93.49% dan AUC sebesar 0.975. Namun, menariknya, metode k-NN berbasis PSO menghasilkan nilai Accuracy yang menurun menjadi 96.74% dan AUC sebesar 0.986.

Penerapan *Particle Swarm Optimization (PSO)* dalam penelitian ini terbukti dapat meningkatkan akurasi algoritma *Random Forest* dan K-NN. Berikut ini adalah grafik untuk memvisualisasikan nilai accuracy dan AUC:


Gambar 5. Grafik Accuracy



Gambar 6. Grafik AUC

5. Kesimpulan

Berdasarkan hasil penelitian dan pengujian yang telah dilakukan dengan menggunakan algoritma *K-Nearest Neighbor* (K-NN) dan *Random Forest* berbasis *Particle Swarm Optimization* (PSO) dalam menentukan tingkat kelulusan tepat waktu mahasiswa, ditemukan bahwa algoritma *Random Forest* memiliki akurasi sebesar 95,79% dan AUC sebesar 0,991. Di sisi lain, ketika algoritma *Random Forest* dioptimasi dengan PSO, akurasinya meningkat menjadi 97,89% dan AUC menjadi 0,993. Sedangkan untuk algoritma k-NN, akurasinya adalah 93,49% dengan AUC sebesar 0,975. Kemudian setelah dioptimasi dengan PSO, akurasinya menjadi 96,74% dengan AUC sebesar 0,986.

Dengan demikian, hasil menunjukkan bahwa algoritma *Random Forest* yang dioptimasi dengan PSO lebih akurat dibandingkan dengan algoritma k-NN yang juga dioptimasi dengan PSO.

Referensi

- [1] A. Yandi Saputra and Y. Primadasa, "Penerapan Teknik Klasifikasi Untuk Prediksi Kelulusan Mahasiswa Menggunakan Algoritma K-Nearest Neighbour Implementation of Classification Method to Predict Student Graduation Using K-Nearest Neighbor Algorithm," 2018.
- [2] "105-Article Text-142-1-10-20190125".
- [3] A. Fauzi, A. Heri, and Y. #2, "JEPIN (Jurnal Edukasi dan Penelitian Informatika) Optimasi Algoritma Klasifikasi Naive Bayes, Decision Tree, K-Nearest Neighbor, dan Random Forest menggunakan Algoritma Particle Swarm Optimization pada Diabetes Dataset".
- [4] P. Bidang Komputer Sains dan Pendidikan Informatika, D. Akademi Perekam dan Informasi Kesehatan Iris Padang Jl Gajah Mada No, and S. Barat, "Jurnal Edik Informatika Data Mining : Klasifikasi Menggunakan Algoritma C4.5 Yuli Mardi".

- [5] Wikipedia contributors, "Particle swarm optimization. In Wikipedia, The Free Encyclopedia," *en.wikipedia.org*, Aug. 10, 2023. https://en.wikipedia.org/w/index.php?title=Particle_swarm_optimization&oldid=1168805143 (accessed Sep. 20, 2023).
- [6] M. Husni Rifqo and N. David Veronica, "Implementasi Algoritme Naïve Bayes Berbasis Particle Swarm Optimization Dalam Penentuan Pemberian Kredit," 2019. [Online]. Available: www.ejournal.unib.ac.id/index.php/pseudocode
- [7] G. S. Saragih and F. Ui, "Universitas Indonesia Prediksi Kebangkrutan Bank Dengan Menggunakan Random Forest Skripsi Glori Stephani Saragih 1406558153 Fakultas Matematika Dan Ilmu Pengetahuan Alam Program Studi Sarjana Matematika Depok Mei 2018."
- [8] Wikipedia contributors, "K-nearest neighbors algorithm," *en.wikipedia.org*, Aug. 10, 2023. https://en.wikipedia.org/w/index.php?title=Knearest_neighbors_algorithm&oldid=1168216736 (accessed Sep. 20, 2023).
- [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, Dec. 2022, doi: 10.33372/stn.v8i2.885.
- [10] M. Labib Mu'tashim *et al.*, "Klasifikasi Ketepatan Lama Studi Mahasiswa Dengan Algoritma Random Forest Dan Gradient Boosting (Studi Kasus Fakultas Ilmu Komputer Universitas Pembangunan Nasional Veteran Jakarta)," 2023.
- [11] T. Nasution, "Implementasi Algoritma K-Nearest Neighbor Untuk Penentuan Kelulusan Mahasiswa Tepat Waktu."
- [12] Ramdani And Galih Farid, "Komparasi Metode K-Nearest Neighbor Dan Neural Network Untuk Menentukan Tingkat Akurasi Dalam Prediksi Kelulusan Mahasiswa (Studi Kasus: Program Studi Informatika Universitas Siliwangi) ," Universitas Siliwangi, 2023. Accessed: Sep. 20, 2023. [Online]. Available: <http://repositori.unsil.ac.id/10237/>
- [13] Eko Prasetyo Rohmawan, "Prediksi Kelulusan Mahasiswa Tepat Waktu Menggunakametode Desicion Tree Dan Artificial Neural Network," *Jurnal Ilmiah MATRIK*, vol. 20, no. 1, 2018.

Copyrights

Copyright for this article is retained by the author(s), with first publication rights granted to the journal.

This is an open-access article distributed under the terms and conditions of the Creative Commons Attribution license (<http://creativecommons.org/licenses/by/4.0/>)



