**Komparasi Algoritma Random Forest dan SVM terhadap Pendaftar Beasiswa Universitas Merdeka**

Marta Yulianti1, \*, Victoria Alysha F. S. 2

1,2*Program Studi Informatika Fakultas Sains dan Teknologi,*

*Universitas Sanata Dharma, Yogyakarta, Indonesia*

***Abstract***

*Providing scholarships at Merdeka University is an effort to assist students in pursuing higher education after considering various factors. This study aims to compare the performance of SVM and Random Forest algorithms in classifying scholarship applicants at Universitas Merdeka. Utilizing a processed dataset of 900 entries, the research findings indicate that Random Forest outperforms SVM in terms of accuracy. Specifically, Random Forest achieves an accuracy rate of 90% for the 70:30 data split, and 90.56% for the 80:20 data split. This suggests that Random Forest is a more effective choice in this context for enhancing the scholarship selection process at Merdeka University.*

*Keyword: Accuracy, Classification, Random Forest, Scholarships, Support Vector Machine*

**Abstrak**

Pemberian beasiswa di Universitas Merdeka merupakan sebuah upaya bantuan bagi mahasiswa yang memungkinkan mereka melanjutkan pendidikan tinggi setelah mempertimbangkan berbagai faktor. Penelitian ini bertujuan untuk membandingkan kinerja algoritma SVM dan Random Forest dalam mengklasifikasi pendaftar beasiswa Universitas Merdeka. Menggunakan dataset berukuran 900 yang telah diproses, hasil penelitian menunjukkan bahwa Random Forest memberikan tingkat akurasi yang lebih baik daripada SVM. Secara khusus, Random Forest mencapai tingkat akurasi sebesar 90% untuk pembagian data 70:30, dan 90.56% untuk pembagian data 80:20. Ini menunjukkan bahwa Random Forest merupakan pilihan yang lebih efektif dalam konteks ini untuk meningkatkan proses seleksi beasiswa di Universitas Merdeka.

Kata Kunci: Akurasi, Beasiswa, Klasifikasi, Random Forest, Support Vector Machine

**1. PENDAHULUAN**

Pemberian beasiswa merupakan salah satu bentuk bantuan bagi mahasiswa untuk dapat melanjutkan pendidikan tinggi. Banyak pertimbangan yang harus diperhatikan dalam memberikan keputusan pemberian beasiswa ini. Di Universitas Merdeka, proses pemberian beasiswa dilakukan dengan melihat berbagai atribut pendaftar seperti Indeks Prestasi Kumulatif (IPK), kondisi ekonomi keluarga, dan lain sebagainya. Universitas Merdeka memiliki sejumlah besar data pendaftar beasiswa yang dapat dianalisis untuk membantu proses seleksi ini.

Untuk mengolah data tersebut secara efektif, diperlukan algoritma machine learning yang dapat mengklasifikasikan dan memprediksi pendaftar yang memenuhi kriteria beasiswa dengan akurasi tinggi. Beberapa algoritma seperti Random Forest dan Support Vector Machine (SVM) telah terbukti mampu menyelesaikan masalah klasifikasi dalam penelitian sebelumnya. Namun, kinerja kedua algoritma ini masih perlu dibandingkan untuk mendapatkan algoritma yang paling tepat diterapkan pada masalah pemberian beasiswa ini.

Salah satu penelitian sebelumnya yang relevan adalah perbandingan antara algoritma Random Forest dan SVM pada analisis sentimen Twitter terkait seorang selebgram yang melarikan diri dari karantina [1]. Hasilnya menunjukkan Random Forest memberikan nilai evaluasi lebih tinggi dengan akurasi 94%, sedangkan SVM mendapatkan akurasi 93%. Selain itu, perbandingan kedua algoritma dalam mengklasifikasi ulasan Shopee menunjukkan bahwa SVM mencapai akurasi sebesar 84,71%, sedangkan Random Forest memiliki akurasi 82,21% [2].

Penelitian ini bertujuan untuk membandingkan kinerja algoritma Random Forest dan Support Vector Machine dalam mengklasifikasikan data pendaftar beasiswa Universitas Merdeka. Dengan melakukan komparasi ini, diharapkan dapat diperoleh wawasan yang lebih dalam mengenai algoritma mana yang lebih efektif dan efisien dalam konteks ini. Hasil dari penelitian ini diharapkan dapat membantu pihak universitas dalam meningkatkan proses seleksi beasiswa, sehingga lebih tepat sasaran dan adil.

**2. METODOLOGI PENELITIAN**

Penelitian ini diawali dengan tahap pengumpulan data yang akan digunakan sebagai data mentah. Data yang terkumpul akan melalui proses pengolahan terlebih dahulu untuk memastikan kualitas dan relevansinya. Setelah data diolah, langkah berikutnya adalah pelatihan model menggunakan dua algoritma yang berbeda, yaitu Support Vector Machine (SVM) dan Random Forest. Pelatihan model dilakukan untuk membangun sistem yang dapat melakukan prediksi atau klasifikasi berdasarkan data yang telah diolah. Setelah proses pelatihan selesai, model tersebut akan diuji untuk menilai kinerjanya.

**A diagram of data processing

Description automatically generated**

**Gambar 1.** Metodologi Penelitian

1. Pengumpulan Data

Dataset yang digunakan dalam penelitian ini adalah dataset Pendaftar Beasiswa Universitas Merdeka. Dataset ini mencakup 1107 data dengan 17 atribut. Atribut-atribut tersebut dijelaskan lebih lanjut pada **Tabel 1**.

**Tabel 1**. Atribut Dataset

|  |  |
| --- | --- |
| Atribut | Keterangan |
| Periode | Periode atau tahun pendaftaran beasiswa |
| Angkatan | Tahun masuk mahasiswa |
| Kabupaten | Kabupaten asal mahasiswa |
| Propinsi | Provinsi asal mahasiswa |
| Semester | Semester saat ini yang sedang ditempuh oleh mahasiswa |
| IPK | Indeks Prestasi Kumulatif (IPK) mahasiswa |
| Srt Ket. Tidak Mampu | Surat Keterangan Tidak Mampu |
| Penghasilan Bapak | Penghasilan ayah mahasiswa |
| Penghasilan Ibu | Penghasilan ibu mahasiswa |
| Tgh. Listrik | Tagihan bulanan rumah mahasiswa |
| Jml. Tagunggan masih studi | Jumlah saudara yang masih bersekolah atau studi |
| Tunggakan | Jumlah tunggakan biaya mahasiswa |
| Keadaan Orang Tua | Kondisi orang tua |
| Jumlah Adik | Jumlah adik kandung mahasiswa |
| Jumlah Kakak | Jumlah kakak kandung mahasiswa |
| Jml. Point | Jumlah poin yang diperoleh berdasarkan kriteria beasiswa |
| Status beasiswa | Status penerimaan beasiswa (diterima/tidak diterima) |

1. *Preprocessing* Data

*Preprocessing* data adalah tahap penting dalam mempersiapkan data mentah agar siap digunakan dalam proses analisis dengan lebih mudah dan akurat [3]. Tahapan preprocessing mencakup beberapa keputusan krusial, seperti penanganan *missing value* untuk mengisi atau menghapus data yang hilang, menghilangkan duplikasi, seleksi data untuk memilih fitur-fitur yang relevan, transformasi data, *balancing* data, serta normalisasi data.

1. *Splitting* Data

*Splitting* datamerupakan proses pembagian dataset menjadi dua bagian: data *training* dan data *testing*. Data training digunakan untuk melatih model klasifikasi, sementara data testing digunakan untuk menguji kinerja model yang telah dilatih. Pembagian data ini membantu dalam membandingkan kinerja model dan mengevaluasi parameter-parameter yang berbeda. Dalam penelitian ini, data dibagi menggunakan dua rasio, yaitu 80:20 dan 70:30 untuk *training* dan *testing*, secara berurutan.

1. Support Vector Machine (SVM)

SVM adalah algoritma yang digunakan untuk klasifikasi dan regresi [4]. Tujuannya adalah menemukan *hyperplane* terbaik, yang berfungsi memisahkan dua kelas data dan terletak di tengah-tengah antara dua objek dari dua kelas yang berbeda. *Hyperplane* ini bertujuan untuk memaksimalkan margin atau jarak antara data pelatihan dan batas keputusan. Proses optimasi SVM yang ingin memaksimalkan nilai margin dapat dilakukan dengan cara meminimalkan pembaginya, yaitu ||w||, dengan rumus:

Minimal: (1)

Dengan syarat:

( + *b*) ≥ 1, 𝑖 = 1, 2, …, 𝑁 (2)

Keterangan:

: Bobot vektor.

: Bilangan skalar yang menyatakan nilai bias.

: Fitur vektor data latih ke-*i*.

: Data ke-*i, i* = 1, 2, …, *n*.

: Label kelas data latih ke-*i* (+1/-1).

Tahapan dalam melakukan klasifikasi dengan algoritma SVM meliputi:

1. Menentukan *hyperplane* atau garis pembatas antara dua support vector.
2. Menentukan margin atau garis jarak antara support vector dan *hyperplane.*
3. Memetakan support vector ke dalam kelas dalam dimensi yang sama.
4. Random Forest

Random Forest adalah metode yang dapat meningkatkan akurasi dengan secara acak memilih atribut untuk setiap node [5]. Metode ini terdiri dari sekumpulan Decision Tree yang digunakan untuk mengklasifikasikan data ke dalam suatu kelas. Setiap Decision Tree dibangun dengan menentukan node akar dan berakhir dengan beberapa node daun untuk menghasilkan keputusan akhir.

Pembentukan masing-masing pohon didasarkan pada prinsip *boostrapping* dan pembangkitan atribut acak. Indeks *Gini* digunakan untuk memilih atribut mana yang akan menjadi simpul pembagi dalam setiap node pohon. Indeks *Gini* dihitung menggunakan rumus:

) = 1 - (3)

Keterangan:

: Frekuensi relatif kelas dalam set.

: Kelas untuk 𝑖=1, …, 𝑐−1.

𝑐 : Jumlah kelas yang telah ditentukan.

Pada proses pembentukan pohon keputusan dalam algoritma Random Forest, kualitas pembagian (*split*) pada fitur k menjadi subset dianalisis dengan memperhatikan jumlah sampel yang termasuk dalam setiap kelas . Evaluasi ini menggunakan indeks *Gini* dari subset yang dihasilkan setelah pembagian. Secara matematis dapat dinyatakan dengan rumus :

= (4)

Keterangan:

𝑛𝑖 : Jumlah sampel dalam subset 𝑆𝑖 setelah di split.

n : Jumlah sampel di node yang diberikan.

1. Evaluasi

Evaluasi adalah tahap yang bertujuan untuk mengukur kualitas model dengan menilai keakuratan hasil yang dicapai oleh model tersebut. Evaluasi kinerja model dapat diukur dengan *confusion matrix* yang mampu menunjukan jumlah ketepatan klasifikasi terhadap kelas tersebut. *Confusion matrix* juga digunakan dalam mengukur akurasi [6][7]*.*

**Tabel 2.** Confusion Matrix dari Dua Kelas

|  |  |  |  |
| --- | --- | --- | --- |
| **Confusion Matrix** | | **Nilai Aktual** | |
| **True** | **False** |
| **Nilai Prediksi** | **True** | TP (*True Positive)* | FP (*False Positive)* |
| **False** | FN (*False Negative)* | TN (*True Negative)* |

Akurasi adalah ukuran yang menunjukkan seberapa dekat nilai prediksi dengan nilai aktual.

(5)

**3.** **HASIL DAN ANALISIS**

Data mentah diperoleh dari dataset Pendaftar Beasiswa Universitas Merdeka yang mencakup 17 atribut dan 1107 baris data. Proses *preprocessing* dimulai dengan seleksi atribut, di mana atribut yang dianggap tidak relevan seperti ‘Periode’, ‘Kabupaten’, dan ‘Provinsi’ dihapus. Selanjutnya, pengecekan missing value dilakukan, dan nilai kosong diisi pada kolom ‘Srt Ket. Tidak Mampu’ sementara baris yang memiliki missing value lainnya dihapus. Transformasi data diterapkan dengan mengubah atribut kategorikal menjadi numerik. Selain itu, dilakukan penghapusan *outlier*, penyeimbangan data dengan mengambil sampel sebanyak 450 data per kelas, serta normalisasi menggunakan metode *min-max scaler*.

Penelitian ini mengkomparasikan kinerja model SVM dan Random Forest dengan membagi data menjadi dua proporsi, yaitu 80:20 dan 70:30. Hasil pengujian menunjukkan bahwa Random Forest mencapai akurasi tertinggi di kedua skema pembagian data.

**Figur 1.** Komparasi SVM dan Random Forest

Berdasarkan hasil perbandingan yang tertera dalam **Figur 1**, terlihat bahwa Random Forest berhasil mencapai tingkat akurasi sebesar 89.63% untuk pembagian data 70:30, dan 91.11% untuk pembagian data 80:20, dalam proses klasifikasi pendaftar beasiswa di Universitas Merdeka.

**4. KESIMPULAN**

Setelah menerapkan metode SVM dan Random Forest untuk mengklasifikasikan pendaftar beasiswa di Universitas Merdeka dengan menggunakan dataset berukuran 900 data, hasilnya menunjukkan bahwa Random Forest menunjukkan kinerja yang lebih unggul dibandingkan dengan SVM. Random Forest mencapai tingkat akurasi sebesar 89.63% untuk pembagian data 70:30, dan 91.11% untuk pembagian data 80:20. Oleh karena itu, dapat disimpulkan bahwa Random Forest merupakan pilihan yang lebih efektif dalam konteks ini, karena mampu memberikan hasil klasifikasi yang lebih akurat.

**REFERENSI**

[1] Wahyuningtias, P., Warih Utami, H., Ahda Raihan, U., Nur Hanifah, H., & Nicholas Adanson, Y. (2022). Comparison of Random Forest and Support Vector Machine Methods on Twitter Sentiment Analysis (case study: internet selebgram rachel vennya escape from quarantine). *Jurnal Teknik Informatika (JUTIF)*, *3*(1), 141–145. <https://doi.org/10.20884/1.jutif.2022.3.1.168>

[2] Suswadi, S., & Erkamim, Moh. (2023). Sentiment Analysis of Shopee App Reviews Using Random Forest and Support Vector Machine. *ILKOM Jurnal Ilmiah*, *15*(3), 427–435. <https://doi.org/10.33096/ilkom.v15i3.1610.427-435>

[3] Agung, A., Daniswara, A., Kadek, I., & Nuryana, D. (2023). Data Preprocessing Pola Pada Penilaian Mahasiswa Program Profesi Guru. *Journal of Informatics and Computer Science*, *05*.

[4] Tantika, R. S., & Kudus, A. (2022). Penggunaan Metode Support Vector Machine Klasifikasi Multiclass pada Data Pasien Penyakit Tiroid. *Bandung Conference Series: Statistics*, *2*(2), 159–166. <https://doi.org/10.29313/bcss.v2i2.3590>

[5] Suci Amaliah, Nusrang, M., & Aswi, A. (2022). Penerapan Metode Random Forest Untuk Klasifikasi Varian Minuman Kopi di Kedai Kopi Konijiwa Bantaeng. *VARIANSI: Journal of Statistics and Its Application on Teaching and Research*, *4*(3), 121–127. <https://doi.org/10.35580/variansiunm31>

[6] Qadrini, L., Seppewali, A., & Aina, A. (2021). Decision Tree dan Adaboost pada Klasifikasi Penerima Program Bantuan Sosial. *Jurnal Inovasi Penelitian*, *2*(7), 1959–1966. https://doi.org/10.47492/jip.v2i7.1046

[7] Azhari, M., Situmorang, Z., & Rosnelly, R. (2021). Perbandingan Akurasi, Recall, dan Presisi Klasifikasi pada Algoritma C4.5, Random Forest, SVM dan Naive Bayes. *JURNAL MEDIA INFORMATIKA BUDIDARMA*, *5*(2), 640. https://doi.org/10.30865/mib.v5i2.2937