Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 5, No 3, June 2026 | Page 295–303
DOI : 10.47065/bulletinds.v5i3.10251
https://ejurnal.seminar-id.com/index.php/bulletinds

Pemanfaatan NLP Untuk Ekstrasi Profil Kompetensi Dari Portofolio Digital Sebagai Dasar Sistem Rekomendasi

Moch Fahmi, Danang Arbian Sulistyo*

Prodi Teknik Informatika, Institut Teknologi dan Bisnis, Malang, Indonesia
Email: [1]muhammadimhaf22@gmail.com, [2,*]danangarbian@email.com
Email Penulis Korespondensi: danangarbian@email.com

Abstrak−Seleksi penerima beasiswa merupakan proses penting dalam mendukung pemerataan akses pendidikan bagi mahasiswa yang memiliki potensi akademik maupun nonakademik. Namun, proses seleksi yang masih dilakukan secara konvensional sering menghadapi kendala berupa banyaknya dokumen yang harus dianalisis serta keterbatasan dalam mengidentifikasi profil kompetensi mahasiswa secara komprehensif. Penelitian ini mengusulkan pemanfaatan Natural Language Processing (NLP) untuk mengekstraksi profil kompetensi dari portofolio digital mahasiswa sebagai dasar pengembangan sistem rekomendasi beasiswa. Data yang digunakan berupa 1.042 dokumen biodata dan portofolio mahasiswa dalam format PDF yang memuat informasi akademik, pengalaman organisasi, aktivitas kemahasiswaan, serta kondisi sosial ekonomi keluarga. Tahapan penelitian meliputi ekstraksi teks dari dokumen, pembersihan dan normalisasi data, pembentukan representasi fitur menggunakan metode Term Frequency-Inverse Document Frequency (TF-IDF), serta integrasi fitur tekstual, numerik, dan kategorikal untuk menghasilkan profil kompetensi yang lebih representatif. Selanjutnya, algoritma Random Forest diterapkan untuk mengklasifikasikan kelayakan penerima beasiswa dan menghasilkan skor rekomendasi bagi setiap kandidat. Evaluasi model dilakukan menggunakan skema pembagian data sebesar 75% untuk pelatihan dan 25% untuk pengujian. Hasil penelitian menunjukkan bahwa model mampu mencapai tingkat akurasi sebesar 74%, nilai F1-Score sebesar 0,46 pada kelas penerima beasiswa, dan nilai Average Precision sebesar 0,526. Temuan tersebut mengindikasikan bahwa pendekatan NLP efektif dalam mengidentifikasi dan mengolah informasi kompetensi yang tersebar pada portofolio digital sehingga dapat mendukung proses seleksi beasiswa yang lebih objektif, cepat, dan terukur. Selain itu, sistem yang dikembangkan mampu menghasilkan pemeringkatan kandidat berdasarkan tingkat kelayakan yang diperoleh dari kombinasi aspek akademik, pengalaman organisasi, serta kondisi sosial ekonomi, sehingga dapat menjadi alternatif solusi dalam meningkatkan kualitas pengambilan keputusan pada program beasiswa berbasis data.

Kata Kunci: Natural Language Processing; profil kompetensi; portofolio digital; sistem rekomendasi beasiswa; Random Forest

Abstract−Scholarship recipient selection is an essential process in supporting equitable access to education for students with academic and non-academic potential. However, conventional selection procedures often face challenges due to the large volume of documents that must be reviewed and the difficulty of comprehensively identifying students' competency profiles. This study proposes the utilization of Natural Language Processing (NLP) to extract competency profiles from students' digital portfolios as the foundation for a scholarship recommendation system. The dataset consisted of 1,042 student biodata and portfolio documents in PDF format containing academic records, organizational experiences, extracurricular activities, and socio-economic background information. The research process included text extraction from PDF documents, data cleaning and normalization, feature representation using the Term Frequency–Inverse Document Frequency (TF-IDF) method, and the integration of textual, numerical, and categorical features to generate more representative competency profiles. Furthermore, the Random Forest algorithm was employed to classify scholarship eligibility and generate recommendation scores for each candidate. Model evaluation was conducted using a train-test split scheme, with 75% of the data allocated for training and 25% for testing. The experimental results demonstrated that the proposed model achieved an accuracy of 74%, an F1-Score of 0.46 for the scholarship recipient class, and an Average Precision value of 0.526. These findings indicate that the NLP-based approach is effective in identifying and processing competency-related information embedded within digital portfolios, thereby supporting a more objective, efficient, and data-driven scholarship selection process. In addition, the developed system is capable of ranking candidates according to their eligibility scores derived from a combination of academic performance, organizational involvement, and socio-economic conditions. Therefore, the proposed approach can serve as an alternative solution for improving decision-making quality in data-driven scholarship recommendation systems.

Keywords: Natural Language Processing; competency profile extraction; digital portfolio; scholarship recommendation system; Random Forest

1. PENDAHULUAN

Perkembangan teknologi digital telah menimbulkan perubahan besar dalam banyak aspek pendidikan tinggi, termasuk dalam cara pengelolaan data mahasiswa dan proses pengambilan keputusan akademik. Saat ini, perguruan tinggi tidak hanya menggunakan data akademik, seperti Indeks Prestasi Kumulatif (IPK), sebagai dasar evaluasi mahasiswa, tetapi juga mulai memperhatikan berbagai faktor non-akademik yang lebih lengkap dalam mencerminkan kemampuan individu. Informasi tersebut dapat meliputi pengalaman organisasi, sertifikasi, keterampilan teknis, pencapaian akademik maupun non-akademik, serta berbagai aktivitas pengembangan diri lainnya yang tercatat dalam portofolio digital [1]. Adanya portofolio digital, gambaran mengenai potensi mahasiswa menjadi lebih menyeluruh dibandingkan hanya dengan mengandalkan data akademik saja [2].

Portofolio digital kini menjadi salah satu alat yang banyak dipakai untuk mencatat keberhasilan mahasiswa selama mereka menempuh pendidikan. Dengan menggunakan portofolio digital, mahasiswa mampu mengumpulkan dan memamerkan berbagai bukti kapasitas yang dimiliki, seperti dokumen, sertifikat, laporan aktivitas, serta karya-karya yang pernah dibuat. Data tersebut sangat berharga dalam mendukung evaluasi dan pengambilan keputusan di institusi pendidikan tinggi. Namun, sebagian besar informasi yang terdapat dalam portofolio digital masih berbentuk teks yang

Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 5, No 3, June 2026 | Page 295–303
DOI : 10.47065/bulletinds.v5i3.10251
https://ejurnal.seminar-id.com/index.php/bulletinds

tidak terstruktur, sehingga menyulitkan analisis otomatis berdasarkan metode tradisional. Proses penentuan kompetensi mahasiswa yang dilakukan secara manual juga memerlukan waktu yang lama serta bisa menimbulkan subjektivitas dalam penilaian [1].

Salah satu tahapan yang memerlukan penilaian menyeluruh terhadap kemampuan mahasiswa adalah pemilihan penerima beasiswa. Umumnya, proses pemilihan beasiswa didasarkan pada beberapa faktor, termasuk prestasi akademis, kegiatan organisasi, keadaan sosial ekonomi, kemampuan kepemimpinan, dan berbagai indikator lainnya. Peningkatan jumlah pelamar sering membuat evaluasi menjadi lebih rumit karena banyaknya dokumen yang perlu diperiksa oleh penyelenggara. Situasi ini menekankan pentingnya adanya sistem yang dapat mempercepat proses pemilihan, menjadikannya lebih objektif dan berbasis data, sehingga rekomendasi yang dihasilkan menjadi lebih tepat dan transparan [3].

Kemajuan dalam teknologi kecerdasan buatan, terutama dalam Natural Language Processing (NLP), membuka jalan untuk menyelesaikan masalah ini. NLP adalah bagian dari kecerdasan buatan yang memungkinkan komputer untuk memahami, mengolah, dan mengambil informasi dari bahasa alami manusia, baik dalam teks maupun suara.[4] Di dunia pendidikan, NLP telah diterapkan dalam berbagai cara, termasuk pengelompokan dokumen, analisis emosi, pengambilan informasi, sistem penilaian otomatis, dan pengembangan sistem pendukung keputusan. Kemampuan NLP untuk mengubah data teks yang tidak terstruktur menjadi informasi yang lebih teratur memungkinkan identifikasi kompetensi mahasiswa dilakukan secara otomatis dan lebih efisien [5].

Dalam penanganan teks, tahap persiapan awal memegang peranan penting untuk memperbaiki kualitas informasi sebelum di telaah. Proses ini meliputi penyaringan data, pemecahan kata, konversi ke huruf kecil, pengurangan kata dasar, dan transformasi data ke bentuk angka. Salah satu cara yang sering digunakan untuk menyajikan teks ialah Term Frequency-Inverse Document Frequency (TF-IDF) yang mampu memberi bobot pada kata-kata krusial dalam suatu naskah [6]. Melalui pendekatan ini, beragam informasi dari kumpulan digital dapat diekstraksi menjadi penanda yang menggambarkan kompetensi mahasiswa, sehingga dapat dijadikan landasan untuk pengelompokan dan saran [7].

Kompetensi mahasiswa tidak hanya ditentukan oleh prestasi akademik, tetapi juga keterlibatan organisasi, kepemimpinan, sertifikasi, pelatihan, dan berbagai pencapaian lainnya. Oleh karena itu, identifikasi kompetensi secara tepat menjadi faktor penting dalam mendukung pengambilan keputusan di lingkungan pendidikan tinggi [8].

Dalam beberapa tahun terakhir, Educational Data Mining dan machine learning banyak dimanfaatkan untuk menganalisis data pendidikan. Berbagai penelitian menunjukkan bahwa teknik tersebut mampu mendukung proses prediksi, klasifikasi, dan rekomendasi secara lebih objektif dibandingkan metode konvensional [9].

Penelitian mengenai sistem rekomendasi beasiswa telah dilakukan oleh Nururrohman dan Setyati [3] dengan memanfaatkan algoritma XGBoost berdasarkan data akademik mahasiswa. Selain itu, Wijaya dan Yuniarto [7] menyebutkan bahwa penerapan machine learning dapat membantu meningkatkan objektivitas dan efisiensi dalam proses pengambilan keputusan. Meskipun demikian, sebagian besar penelitian masih menggunakan data yang terstruktur, sedangkan penelitian ini memanfaatkan portofolio digital yang bersifat tidak terstruktur melalui pendekatan Natural Language Processing (NLP).

Riset-riset terdahulu telah membuktikan bahwa integrasi machine learning dapat menyempurnakan proses penentuan kebijakan di sektor pendidikan. Sejumlah studi telah mengaplikasikan machine learning guna memproyeksikan capaian akademis mahasiswa, mendeteksi potensi , dan merancang sistem saran yang bertumpu pada informasi Lebih jauh lagi, kajian terkait sistem rekomendasi pemberian beasiswa berbasis machine learning mengungkap bahwa penggunaan klasifikasi algoritma dapat menunjang proses pemilihan penerima beasiswa agar lebih adil dibandingkan cara konvensional. Kendati demikian, mayoritas riset yang ada masih terpusat pada pemanfaatan data yang terorganisir rapi, contohnya nilai perkuliahan dan data kependudukan. Sementara itu, pemanfaatan data yang tidak terstruktur dari portofolio digital belum banyak dieksplorasi.Portofolio digital juga semakin sering digunakan sebagai sumber informasi untuk merekam perjalanan akademik dan nonakademik mahasiswa. Melalui sistem ini, berbagai capaian mahasiswa dapat terdokumentasi secara lebih lengkap, mulai dari prestasi akademik, pengalaman kepemimpinan, keterlibatan organisasi, hingga aktivitas pengembangan kompetensi lainnya. Keberagaman data tersebut memungkinkan terbentuknya gambaran yang lebih menyeluruh mengenai potensi dan kualitas mahasiswa [10].

Walaupun penelitian terkait NLP, text mining, dan sistem rekomendasi telah banyak dilakukan, sebagian besar penelitian masih berfokus pada penggunaan data terstruktur seperti nilai akademik, IPK, dan data demografis mahasiswa. Penelitian yang memanfaatkan portofolio digital sebagai sumber utama ekstraksi kompetensi masih relatif terbatas. Selain itu, integrasi antara NLP untuk ekstraksi profil kompetensi dan algoritma Random Forest sebagai dasar sistem rekomendasi beasiswa belum banyak dieksplorasi. Oleh karena itu, penelitian ini berupaya mengisi kesenjangan tersebut dengan menggabungkan data tekstual dari portofolio digital dan data pendukung lainnya untuk menghasilkan rekomendasi beasiswa yang lebih komprehensif [11][12].

Untuk menghasilkan model klasifikasi yang mampu menangani berbagai jenis data, penelitian ini menggunakan algoritma Random Forest. Algoritma tersebut dipilih karena memiliki kemampuan yang baik dalam mengolah data dengan karakteristik yang beragam, mampu mengurangi risiko overfitting, serta menghasilkan tingkat akurasi yang tinggi pada berbagai kasus klasifikasi Selain itu, Random Forest juga dapat digunakan untuk mengidentifikasi fitur-fitur penting yang berkontribusi terhadap hasil prediksi sehingga mendukung proses interpretasi model secara lebih baik [13].

Berdasarkan permasalahan tersebut, penelitian ini mengusulkan pemanfaatan Natural Language Processing untuk mengekstraksi profil kompetensi mahasiswa dari portofolio digital sebagai dasar sistem rekomendasi beasiswa. Informasi

Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 5, No 3, June 2026 | Page 295–303
DOI : 10.47065/bulletinds.v5i3.10251
https://ejurnal.seminar-id.com/index.php/bulletinds

hasil ekstraksi akan dikombinasikan dengan data akademik, aktivitas organisasi, dan kondisi sosial ekonomi mahasiswa untuk membentuk profil kompetensi yang lebih komprehensif.[14] Selanjutnya, algoritma Random Forest digunakan untuk melakukan klasifikasi dan menghasilkan rekomendasi penerima beasiswa. Dengan pendekatan tersebut, diharapkan sistem mampu mendukung proses seleksi beasiswa secara lebih objektif, transparan, dan berbasis data sehingga dapat membantu pihak penyelenggara dalam menentukan penerima beasiswa yang paling sesuai dengan kriteria yang ditetapkan [15].

Berdasarkan hal tersebut, penelitian ini mengusulkan pendekatan yang menggabungkan teknik NLP dengan algoritma Random Forest untuk mengekstraksi profil kompetensi mahasiswa dari portofolio digital dan menggunakannya sebagai dasar sistem rekomendasi beasiswa. Kebaruan penelitian ini terletak pada integrasi data tekstual dari portofolio digital dengan data numerik dan kategorikal dalam proses pemodelan. Dengan pendekatan ini, diharapkan sistem dapat menghasilkan rekomendasi penerima beasiswa yang lebih objektif, transparan, dan merepresentasikan kompetensi mahasiswa secara menyeluruh [16][17].

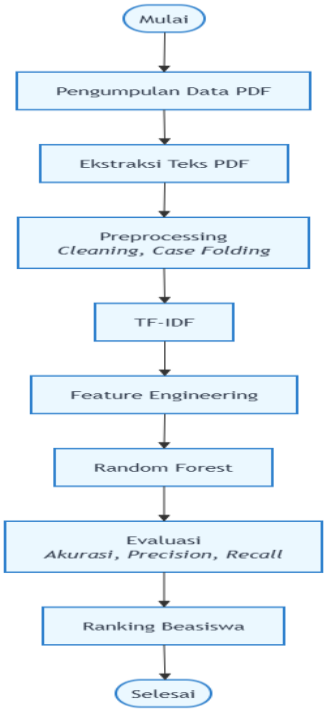
Sistem ini dirancang untuk membantu pengelola pendidikan dalam mengubah data yang kompleks menjadi informasi yang lebih terstruktur sehingga dapat digunakan dalam proses pengambilan keputusan, termasuk seleksi beasiswa. Pendekatan berbasis data dan algoritma machine learning semakin banyak digunakan untuk meningkatkan akurasi hasil keputusan[18].

Seiring perkembangan teknologi, teknik text mining semakin penting dalam pengolahan data tidak terstruktur. Teknik ini memungkinkan sistem untuk mengekstraksi informasi dari kumpulan dokumen dalam jumlah besar, termasuk dokumen akademik dan portofolio digital mahasiswa [19]. Proses seperti tokenisasi, stemming, dan pembobotan kata digunakan untuk mengubah teks menjadi representasi numerik yang dapat diproses oleh model machine learning [20].

2. METODOLOGI PENELITIAN

Penelitian ini menerapkan pendekatan kuantitatif dengan memanfaatkan teknik *Natural Language Processing* (NLP) dan algoritma pembelajaran mesin untuk membangun sistem rekomendasi beasiswa berbasis profil kompetensi mahasiswa. Pendekatan ini dipilih karena mampu mengolah data dalam jumlah besar secara otomatis dan menghasilkan rekomendasi yang lebih objektif dibandingkan proses seleksi konvensional. Tahapan penelitian dirancang secara sistematis mulai dari pengumpulan data, ekstraksi informasi dari dokumen digital, pengolahan data, pembentukan fitur, pembangunan model klasifikasi, hingga evaluasi kinerja sistem.

Untuk memberikan gambaran yang lebih jelas mengenai tahapan penelitian yang dilakukan, alur proses penelitian disajikan pada Gambar 1.



Gambar 1. Tahapan Penelitian Sistem Rekomendasi Beasiswa

Gambar 1 memperlihatkan rangkaian tahapan penelitian yang digunakan dalam pengembangan sistem rekomendasi beasiswa berbasis NLP. Proses diawali dengan pengumpulan dokumen portofolio mahasiswa dalam format PDF yang kemudian diekstraksi menjadi data teks. Data tersebut selanjutnya melalui tahap preprocessing dan pembentukan fitur menggunakan metode TF-IDF sebelum diproses pada model Random Forest. Hasil klasifikasi

Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 5, No 3, June 2026 | Page 295–303
DOI : 10.47065/bulletinds.v5i3.10251
https://ejurnal.seminar-id.com/index.php/bulletinds

kemudian dievaluasi menggunakan beberapa metrik pengujian dan digunakan sebagai dasar dalam menentukan peringkat kandidat penerima beasiswa.

Sumber data penelitian berupa dokumen portofolio digital mahasiswa yang tersimpan dalam format PDF. Dokumen tersebut berisi berbagai informasi yang mencerminkan kompetensi dan karakteristik mahasiswa, seperti identitas pribadi, program studi, riwayat pendidikan, prestasi akademik, pengalaman organisasi, aktivitas kemahasiswaan, serta kondisi sosial ekonomi keluarga. Sebanyak 1.042 dokumen digunakan sebagai dataset penelitian yang kemudian diproses untuk menghasilkan informasi yang relevan dalam mendukung proses rekomendasi penerima beasiswa.

Tahap pertama penelitian adalah ekstraksi data dari dokumen PDF. Proses ini dilakukan untuk mengonversi informasi yang terdapat pada portofolio digital menjadi data yang dapat dibaca dan diolah oleh sistem. Setelah proses ekstraksi selesai, dilakukan tahap prapemrosesan (*preprocessing*) yang mencakup pembersihan data, normalisasi teks, penghapusan karakter yang tidak diperlukan, serta penyeragaman format penulisan. Langkah ini bertujuan untuk meningkatkan kualitas data dan meminimalkan kesalahan yang dapat memengaruhi hasil analisis.

Selanjutnya, dilakukan proses pembentukan fitur (*feature engineering*) untuk menghasilkan representasi data yang dapat digunakan dalam proses pembelajaran mesin. Informasi berbentuk teks, seperti nama, program studi, asal sekolah, pengalaman organisasi, kegiatan kemahasiswaan, dan pekerjaan orang tua, digabungkan menjadi satu korpus teks. Data tersebut kemudian ditransformasikan menggunakan metode *Term Frequency–Inverse Document Frequency* (TF-IDF) sehingga dapat direpresentasikan dalam bentuk numerik. Selain fitur tekstual, penelitian ini juga memanfaatkan fitur numerik berupa IPK, jumlah SKS, dan jumlah tanggungan keluarga. Adapun data kategorikal, seperti jenis kelamin dan program studi, dikonversi menggunakan teknik *One-Hot Encoding* agar dapat diproses oleh model klasifikasi.

Pada tahap pemodelan, algoritma *Random Forest* digunakan untuk mengidentifikasi tingkat kelayakan mahasiswa dalam menerima beasiswa. Pemilihan algoritma ini didasarkan pada kemampuannya dalam menangani data dengan berbagai tipe atribut serta menghasilkan performa klasifikasi yang stabil. Untuk memastikan proses pengolahan data berjalan secara terintegrasi, tahapan transformasi fitur dan pelatihan model dikombinasikan menggunakan mekanisme *Pipeline* dan *ColumnTransformer*. Dataset kemudian dibagi menjadi data pelatihan dan data pengujian dengan perbandingan 75:25 menggunakan metode *stratified sampling* guna menjaga proporsi setiap kelas pada kedua kelompok data.

Kinerja model dievaluasi menggunakan sejumlah metrik pengukuran, yaitu akurasi, presisi, *recall*, *F1-score*, *Area Under the Curve* (AUC), dan *Average Precision* (AP). Selain itu, evaluasi juga dilakukan melalui analisis *confusion matrix*, kurva ROC, dan *Precision-Recall Curve* untuk mengetahui kemampuan model dalam membedakan kandidat yang layak dan tidak layak menerima beasiswa. Setelah model memperoleh hasil evaluasi yang memadai, sistem digunakan untuk menghasilkan skor probabilitas bagi setiap mahasiswa. Skor tersebut selanjutnya diurutkan untuk membentuk peringkat kandidat sehingga diperoleh rekomendasi mahasiswa dengan tingkat kelayakan tertinggi sebagai calon penerima beasiswa.

Proses penelitian tidak hanya difokuskan pada klasifikasi kelayakan penerima beasiswa, tetapi juga mencakup tahap pemeringkatan kandidat untuk menghasilkan rekomendasi yang lebih akurat. Nilai probabilitas yang dihasilkan oleh model *Random Forest* digunakan sebagai indikator tingkat kelayakan masing-masing mahasiswa dalam memperoleh beasiswa. Probabilitas tersebut mencerminkan tingkat keyakinan model terhadap status penerimaan setiap kandidat. Selanjutnya, seluruh mahasiswa diurutkan berdasarkan skor probabilitas dari nilai tertinggi hingga terendah sehingga diperoleh daftar peringkat yang menunjukkan prioritas penerima beasiswa. Pendekatan ini memungkinkan proses seleksi dilakukan secara lebih objektif karena keputusan yang dihasilkan didasarkan pada analisis data dan pola yang dipelajari oleh model. Selain itu, mekanisme pemeringkatan memberikan kemudahan bagi pihak pengelola beasiswa dalam menentukan kandidat yang paling memenuhi kriteria yang telah ditetapkan. Dengan demikian, sistem yang dikembangkan tidak hanya mampu mengidentifikasi mahasiswa yang layak menerima beasiswa, tetapi juga menyediakan urutan rekomendasi yang dapat digunakan sebagai dasar pengambilan keputusan secara transparan, konsisten, dan berbasis data.

3. HASIL DAN PEMBAHASAN

Penelitian ini bertujuan untuk mengembangkan sistem seleksi penerima beasiswa berbasis pembelajaran mesin dengan memanfaatkan dokumen biodata peserta sebagai sumber data utama. Sistem dirancang untuk mengolah dokumen biodata mulai dari tahap ekstraksi data hingga menghasilkan rekomendasi penerima beasiswa berdasarkan hasil klasifikasi dan pemeringkatan. Secara umum, tahapan penelitian meliputi ekstraksi dokumen, pembentukan dataset, prapemrosesan data, rekayasa fitur, pelatihan model klasifikasi, evaluasi performa, dan proses perangkingan kandidat.

Data penelitian berupa dokumen biodata peserta beasiswa dalam format PDF. Dokumen tersebut bersifat tidak terstruktur sehingga tidak dapat langsung digunakan dalam proses pembelajaran mesin. Oleh karena itu, dilakukan proses ekstraksi dan parsing untuk mengubah data menjadi format terstruktur. Setiap halaman dokumen diproses untuk mengidentifikasi atribut penting, seperti nama mahasiswa, program studi, jenis kelamin, IPK, asal sekolah, pekerjaan orang tua, penghasilan keluarga, dan jumlah tanggungan. Variasi penulisan atribut dinormalisasi menggunakan mekanisme alias mapping sehingga seluruh informasi dapat ditempatkan pada kolom yang sesuai. Hasil proses ekstraksi menghasilkan dataset sebanyak 1.042 data dengan 15 atribut utama.

Tahap berikutnya adalah prapemrosesan data (preprocessing) yang bertujuan untuk meningkatkan kualitas dataset. Proses ini meliputi pembersihan data dari karakter khusus, simbol yang tidak relevan, penghapusan spasi berlebih,

Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 5, No 3, June 2026 | Page 295–303
DOI : 10.47065/bulletinds.v5i3.10251
https://ejurnal.seminar-id.com/index.php/bulletinds

serta penanganan nilai yang hilang. Setelah data dibersihkan, dilakukan rekayasa fitur untuk mengubah data yang bersifat tekstual, numerik, dan kategorikal ke dalam representasi numerik yang dapat diproses oleh algoritma pembelajaran mesin.

Pada fitur tekstual, beberapa atribut seperti nama, program studi, asal sekolah, organisasi, unit kegiatan mahasiswa, dan pekerjaan orang tua digabungkan menjadi satu representasi teks yang kemudian dikonversi menggunakan metode Term Frequency–Inverse Document Frequency (TF-IDF). Sementara itu, atribut numerik seperti IPK, jumlah SKS, dan jumlah tanggungan digunakan secara langsung setelah dilakukan normalisasi. Adapun atribut kategorikal dikonversi menggunakan teknik One-Hot Encoding sehingga dapat digunakan dalam proses klasifikasi.

Dataset yang telah diproses kemudian dibagi menjadi data pelatihan dan data pengujian. Data pelatihan digunakan untuk membangun model klasifikasi, sedangkan data pengujian digunakan untuk mengevaluasi kemampuan model dalam mengenali data yang belum pernah dipelajari sebelumnya. Pada penelitian ini digunakan algoritma Random Forest sebagai metode klasifikasi utama karena mampu menangani kombinasi fitur teks, numerik, dan kategorikal serta memiliki ketahanan yang baik terhadap data yang mengandung noise.

Kinerja model dievaluasi menggunakan beberapa metrik, yaitu akurasi, precision, recall, F1-score, dan Area Under Curve (AUC). Hasil probabilitas yang dihasilkan oleh model kemudian digunakan sebagai skor kelayakan penerimaan beasiswa. Seluruh kandidat diurutkan berdasarkan skor probabilitas secara menurun sehingga diperoleh daftar penerima beasiswa yang direkomendasikan oleh sistem. Kandidat dengan nilai tertinggi ditempatkan pada peringkat teratas sebagai calon penerima beasiswa yang paling layak berdasarkan hasil prediksi model.

3.1 Pembagian Data Training dan Testing

Gambar 2 menggambarkan rancangan sistem yang digunakan untuk proses penentuan penerima beasiswa. Tahapan awal dimulai dengan pengumpulan data biodata mahasiswa dalam format PDF, yang kemudian diubah menjadi data terstruktur melalui proses ekstraksi dan parsing. Pada tahap ini, berbagai atribut penting seperti identitas mahasiswa, informasi akademik, aktivitas organisasi, serta kondisi ekonomi keluarga diekstraksi dan dipetakan ke kolom yang sesuai. Data hasil ekstraksi kemudian melalui proses validasi untuk memastikan kesesuaian dan konsistensi setiap atribut sebelum disimpan sebagai dataset terstruktur. Setelah dataset terbentuk, dilakukan tahapan praproses dan rekayasa fitur untuk menghasilkan data numerik yang siap digunakan oleh algoritma machine learning. Proses ini meliputi pembersihan data, transformasi fitur teks menggunakan TF-IDF, pengolahan fitur numerik, serta pengkodean variabel kategorikal dengan metode One-Hot Encoding. Dataset yang telah diproses kemudian digunakan sebagai dasar dalam pembangunan model klasifikasi

Gambar 2. Alur Proses Ekstraksi Fitur hingga Perangkingan Beasiswa

.

Data penelitian dipisahkan menjadi data pelatihan dan data pengujian dengan komposisi 80% dan 20%. Data pelatihan dimanfaatkan untuk membentuk model Random Forest, sedangkan data pengujian digunakan untuk menilai kemampuan model dalam mengenali pola kelayakan penerima beasiswa. Penilaian performa dilakukan menggunakan confusion matrix, accuracy, precision, recall, dan F1-score. Setelah proses evaluasi selesai, model dilatih kembali menggunakan seluruh data untuk menghasilkan skor probabilitas yang digunakan dalam proses pemeringkatan kandidat penerima beasiswa.

3.2 Implementasi Model Random Forest

Model Random Forest dibangun menggunakan data pelatihan yang telah melalui proses ekstraksi, pembersihan, dan pembentukan fitur. Selanjutnya, model dievaluasi menggunakan data pengujian untuk mengetahui kemampuan klasifikasi terhadap data yang belum pernah diproses sebelumnya. Evaluasi dilakukan menggunakan confusion matrix, accuracy, precision, recall, dan F1-score sehingga kinerja model dapat diukur secara menyeluruh.

3.3 Klasifikasi Menggunakan Random Forest

Random Forest dipilih sebagai model klasifikasi utama karena mampu menangani data berdimensi tinggi yang terdiri atas kombinasi fitur teks, numerik, dan kategorikal. Metode ini bekerja dengan membangun sejumlah pohon keputusan (decision tree) yang dilatih menggunakan sampel data dan subset fitur yang dipilih secara acak. Setiap pohon menghasilkan prediksi independen, kemudian hasil akhir ditentukan berdasarkan mekanisme voting mayoritas atau rata-rata probabilitas seluruh pohon. Pada penelitian ini, model Random Forest dibangun menggunakan 400 pohon keputusan (n_estimators = 400) dengan parameter class_weight = balanced untuk mengatasi ketidakseimbangan distribusi kelas penerima dan nonpenerima beasiswa

Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 5, No 3, June 2026 | Page 295–303
DOI : 10.47065/bulletinds.v5i3.10251
https://ejurnal.seminar-id.com/index.php/bulletinds

3.4 Hasil Klasifikasi Menggunakan Random Forest

Penelitian ini menggunakan dataset biodata mahasiswa sebanyak 1.042 record yang terdiri atas 15 atribut, meliputi data akademik, aktivitas kemahasiswaan, dan kondisi ekonomi keluarga. Variabel target yang digunakan adalah Status Beasiswa dengan dua kategori, yaitu *Terima* dan *Tidak Terima*. Model Random Forest digunakan untuk mempelajari pola kelayakan penerima beasiswa berdasarkan atribut-atribut tersebut. Probabilitas kelayakan setiap kandidat diperoleh dari rata-rata prediksi seluruh pohon keputusan dalam model, yang dirumuskan sebagai:

$$P(y = 1|x) = \frac{1}{T}\sum_{t=1}^{T} h_t(x) \quad (1)$$

Keterangan pada rumus 1 menunjukkan bahwa P(y=1|x) merupakan probabilitas seorang mahasiswa dinyatakan layak menerima beasiswa berdasarkan variabel atau kriteria input yang dianalisis. Variabel (T) menyatakan jumlah pohon keputusan (*decision tree*) yang digunakan dalam model. Sementara itu, $h_t(x)$ merupakan prediksi yang dihasilkan oleh pohon keputusan ke-(t) terhadap data input (x).

Output model berupa probabilitas kelayakan penerimaan beasiswa pada rentang nilai 0–1. Semakin tinggi probabilitas yang dihasilkan, semakin besar peluang kandidat untuk direkomendasikan sebagai penerima beasiswa. Namun demikian, hasil pengujian menunjukkan bahwa banyak kandidat memiliki nilai probabilitas yang relatif berdekatan. Kondisi ini menyulitkan proses penentuan prioritas penerima beasiswa secara objektif apabila hanya mengandalkan probabilitas dari Random Forest. Probabilitas yang dihasilkan oleh Random Forest kemudian digunakan sebagai dasar dalam proses pemeringkatan kandidat penerima beasiswa. Semakin tinggi nilai probabilitas yang diperoleh, semakin besar tingkat kelayakan mahasiswa untuk direkomendasikan sebagai penerima beasiswa.

Tabel 1. Profil Data Mahasiswa yang Digunakan dalam Prediksi Beasiswa

Fitur	Nilai
IPK	3.72
SKS	144
Tanggungan	4
Penghasilan (juta)	1.5

Tabel 1 menampilkan profil akademik dan kondisi ekonomi mahasiswa yang digunakan dalam proses penilaian penerimaan beasiswa. Mahasiswa memiliki nilai IPK sebesar 3,72 dengan total 144 SKS yang telah ditempuh, menunjukkan capaian akademik yang baik. Dari aspek ekonomi keluarga, mahasiswa memiliki 4 orang tanggungan dengan penghasilan keluarga sekitar Rp1,5 juta per bulan. Data tersebut menjadi salah satu dasar dalam menentukan tingkat kelayakan mahasiswa untuk memperoleh bantuan beasiswa berdasarkan aspek akademik dan kondisi sosial ekonomi.

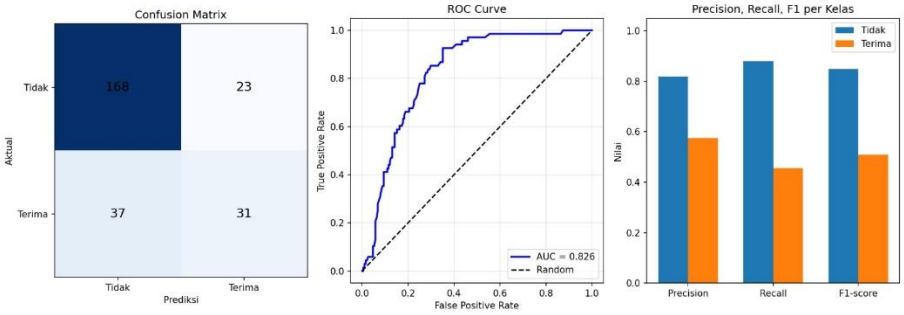
Tabel 2. Sepuluh Kandidat dengan Skor Tertinggi

Rank	Nama Lengkap	Prodi	IPK	Skor
1	Juieta Foxwell	S1 Psikolog | Reguler	3.39	0.9425
2	Patty Crowe	S1 Teknik Elektro | Reguler	3.50	0.9300
3	Jareb Skule	S1 Teknik Elektro | Reguler	3.68	0.9300
4	Eloisa O'Mahoney	S1 Teknik Elektro | Reguler	3.16	0.9250
5	Barbaraanne Abrahamowitcz	S1 Psikolog | Reguler	3.10	0.9225
6	Valerie Gerrard	S1 Teknik Elektro | Reguler	3.16	0.9200
7	Alfonse Bootes	S1 Teknik Elektro | Reguler	3.54	0.9175
8	Ethyl Josiah	S1 Keperawatan | Reguler	3.68	0.9150
9	Rouvin Gres	S1 Teknik Mesin | Reguler	3.41	0.9075
10	Jodie Hought	S1 Teknik Elektro | Reguler	3.50	0.9050

Tabel 2 menunjukkan bahwa kandidat dengan skor tertinggi tidak hanya dipengaruhi oleh nilai akademik, tetapi juga kombinasi faktor non-akademik dan ekonomi yang dipelajari oleh model Random Forest dalam proses klasifikasi..

Hasil evaluasi model menggunakan confusion matrix, kurva ROC, serta metrik precision, recall, dan F1-score ditunjukkan pada Gambar 3

Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 5, No 3, June 2026 | Page 295–303
DOI : 10.47065/bulletinds.v5i3.10251
https://ejurnal.seminar-id.com/index.php/bulletinds

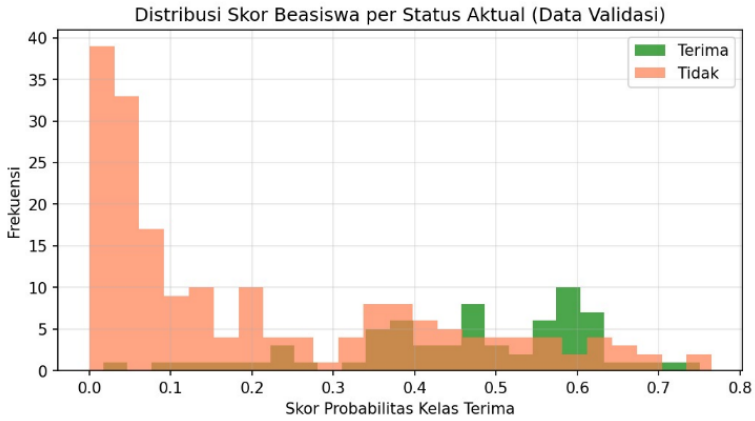


Gambar 3. Hasil Confusion Matrix, Kurva ROC, dan Evaluasi Precision Recall

Gambar 3 memperlihatkan performa model dalam melakukan klasifikasi terhadap mahasiswa penerima beasiswa. Berdasarkan hasil confusion matrix, model mampu mengklasifikasikan dengan benar sebanyak 168 data mahasiswa yang tidak menerima beasiswa dan 31 data mahasiswa yang menerima beasiswa. Namun demikian, masih ditemukan beberapa kesalahan prediksi, yaitu 23 data yang sebenarnya tidak menerima beasiswa tetapi diprediksi menerima, serta 37 data yang sebenarnya menerima beasiswa namun diprediksi tidak menerima.

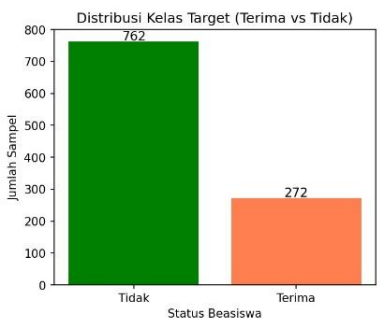
Pada kurva ROC, diperoleh nilai AUC sebesar 0,826 yang menunjukkan bahwa model memiliki kemampuan yang cukup baik dalam membedakan antara mahasiswa yang layak dan tidak layak menerima beasiswa. Di sisi lain, nilai precision, recall, dan F1-score pada kelas “Tidak Menerima” lebih tinggi dibandingkan dengan kelas “Menerima”. Hal ini mengindikasikan bahwa model lebih akurat dalam mengidentifikasi mahasiswa yang tidak menerima beasiswa dibandingkan dengan yang menerima.

Secara keseluruhan, hasil evaluasi menunjukkan bahwa model telah memberikan performa yang cukup baik dalam proses klasifikasi penerima beasiswa. Meskipun demikian, masih diperlukan peningkatan lebih lanjut agar kemampuan model dalam mendeteksi mahasiswa yang benar-benar layak menerima beasiswa dapat menjadi lebih optimal.



Gambar 4. Distribusi Skor Probabilitas Penerimaan Beasiswa

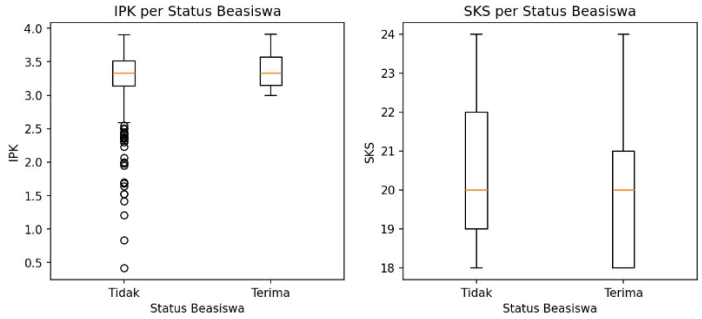
Gambar 4 menggambarkan sebaran skor probabilitas penerimaan beasiswa pada data validasi berdasarkan status aktual mahasiswa. Mahasiswa yang benar-benar menerima beasiswa umumnya memiliki nilai probabilitas yang lebih tinggi, dengan distribusi yang banyak berada pada kisaran 0,4 hingga 0,7. Sebaliknya, mahasiswa yang tidak menerima beasiswa cenderung memiliki skor yang lebih rendah, yaitu pada rentang 0,0 hingga 0,2. Namun demikian, masih terlihat adanya area tumpang tindih antara kedua kelompok pada rentang skor menengah, yang mengindikasikan bahwa model belum sepenuhnya mampu membedakan beberapa kandidat secara tegas. Secara keseluruhan, hasil ini menunjukkan bahwa model Random Forest memiliki performa yang cukup baik dalam memisahkan calon penerima dan bukan penerima beasiswa berdasarkan nilai probabilitas yang dihasilkan.



Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 5, No 3, June 2026 | Page 295–303
DOI : 10.47065/bulletinds.v5i3.10251
https://ejurnal.seminar-id.com/index.php/bulletinds

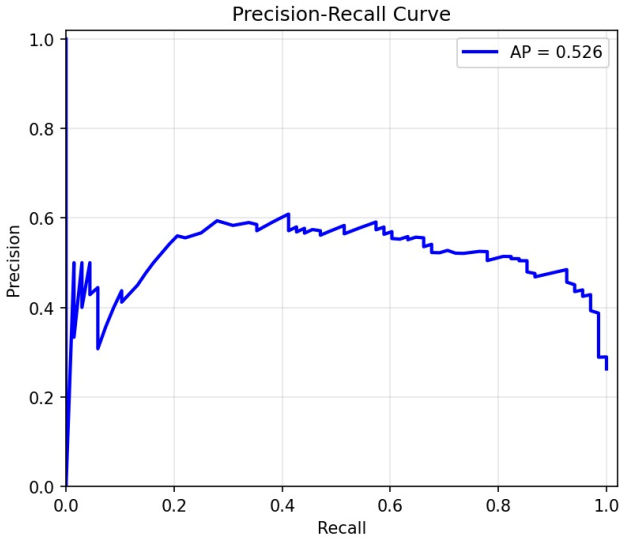
Gambar 5. Distribusi Data Berdasarkan Status Beasiswa

Gambar 5 memperlihatkan distribusi jumlah data berdasarkan status penerimaan beasiswa. Sebanyak 762 mahasiswa tercatat tidak menerima beasiswa, sedangkan 272 mahasiswa lainnya menerima beasiswa. Ketimpangan jumlah yang cukup signifikan antara kedua kelas tersebut menunjukkan bahwa dataset yang digunakan bersifat tidak seimbang (imbalanced). Kondisi ini berpotensi memengaruhi kinerja model dalam mengenali kelas mahasiswa penerima beasiswa secara optimal. Oleh karena itu, diperlukan strategi penanganan data tidak seimbang, seperti penggunaan pengaturan *class weight* serta evaluasi model yang lebih komprehensif dengan mempertimbangkan metrik Precision, Recall, dan F1-Score.



Gambar 6. Perbandingan IPK dan SKS Berdasarkan Status Beasiswa

Gambar 6 menunjukkan perbandingan nilai IPK dan SKS berdasarkan status penerimaan beasiswa. Pada variabel IPK, mahasiswa yang menerima beasiswa cenderung memiliki nilai yang sedikit lebih tinggi dan lebih konsisten dibandingkan mahasiswa yang tidak menerima beasiswa. Sementara itu, pada variabel SKS, kedua kelompok memiliki median yang relatif sama, yaitu sekitar 20 SKS, dengan rentang distribusi yang juga tidak jauh berbeda. Hasil ini mengindikasikan bahwa IPK memiliki pengaruh yang lebih kuat dalam membedakan status penerimaan beasiswa dibandingkan jumlah SKS yang ditempuh mahasiswa.



Gambar 7. Precision Recall Curve Model Random Forest

Gambar 7 menunjukkan hubungan antara nilai *precision* dan *recall* dalam memprediksi penerima beasiswa. Kurva menghasilkan nilai Average Precision (AP) sebesar 0,526, yang menunjukkan bahwa model memiliki kemampuan cukup baik dalam mengidentifikasi mahasiswa yang layak menerima beasiswa meskipun data bersifat tidak seimbang. Secara umum, nilai precision cenderung berada pada kisaran 0,5 ketika recall meningkat, sehingga model masih mampu mempertahankan tingkat ketepatan prediksi yang cukup stabil. Hasil ini menunjukkan bahwa Random Forest cukup efektif dalam mendeteksi kandidat penerima beasiswa, terutama pada kondisi distribusi kelas yang tidak seimbang.

Hasil yang diperoleh menunjukkan bahwa pendekatan machine learning mampu mendukung proses seleksi penerima beasiswa secara lebih objektif. Temuan ini sejalan dengan penelitian Nururrohman dan Setyati [4] yang menerapkan algoritma XGBoost dalam sistem rekomendasi beasiswa. Selain itu, Wijaya dan Yuniarto [10] juga menjelaskan bahwa metode klasifikasi dapat membantu meningkatkan kualitas pengambilan keputusan berbasis data. Perbedaan penelitian ini terletak pada pemanfaatan data portofolio digital yang diolah menggunakan pendekatan Natural Language Processing (NLP), sehingga informasi kompetensi mahasiswa dapat diekstraksi secara lebih komprehensif sebagai dasar rekomendasi beasiswa.

Bulletin of Data Science
ISSN 2745-7966 (Media Online)
Vol 5, No 3, June 2026 | Page 295–303
DOI : 10.47065/bulletinds.v5i3.10251
https://ejurnal.seminar-id.com/index.php/bulletinds

4. KESIMPULAN

Penelitian ini menegaskan bahwa integrasi Natural Language Processing (NLP) dengan algoritma Random Forest dapat menjadi solusi yang efektif dalam mendukung proses seleksi penerima beasiswa berbasis data. Pendekatan ini memungkinkan data dari portofolio digital mahasiswa yang awalnya tidak terstruktur untuk diolah menjadi informasi yang lebih sistematis dan siap dianalisis. Melalui proses ekstraksi informasi, berbagai aspek kompetensi mahasiswa seperti prestasi akademik, keterlibatan organisasi, keterampilan, serta kondisi sosial ekonomi dapat digali secara lebih menyeluruh. Hal ini membuat proses penilaian tidak hanya bertumpu pada indikator kuantitatif seperti IPK, tetapi juga mempertimbangkan data kualitatif yang mencerminkan potensi mahasiswa secara lebih lengkap. Penggunaan metode TF-IDF dalam tahap pengolahan fitur membantu meningkatkan kualitas representasi teks sehingga informasi penting dalam dokumen dapat diidentifikasi dengan lebih optimal. Selanjutnya, penerapan algoritma Random Forest dalam proses klasifikasi memberikan kemampuan untuk mengelola data yang memiliki variasi tinggi serta menghasilkan prediksi yang konsisten dan akurat. Output dari proses ini kemudian dapat dikonversi menjadi skor rekomendasi yang digunakan sebagai dasar dalam menentukan peringkat calon penerima beasiswa secara lebih objektif. Secara keseluruhan, sistem yang dikembangkan melalui penelitian ini memberikan kontribusi penting dalam pengembangan sistem pendukung keputusan di bidang pendidikan berbasis kecerdasan buatan. Pendekatan ini mampu meningkatkan efisiensi, transparansi, dan objektivitas dalam proses seleksi beasiswa. Selain itu, penggabungan antara data teks, numerik, dan kategorikal menghasilkan gambaran yang lebih representatif terhadap kondisi nyata mahasiswa. Dengan demikian, penelitian ini dapat menjadi landasan bagi pengembangan sistem seleksi beasiswa yang lebih cerdas, adaptif, dan berbasis data di masa yang akan datang.

REFERENCES

[1] S. V. Panyukova, "Student's Digital Portfolio for Assessing and Presenting Talents," *ITM Web Conf.*, vol. 35, p. 02006, 2020, doi: 10.1051/itmconf/20203502006.
[2] S. A. Aljawarneh, "Reviewing and exploring innovative ubiquitous learning tools in higher education," *J. Comput. High. Educ.*, vol. 32, no. 1, pp. 57–73, 2020, doi: 10.1007/s12528-019-09207-0.
[3] M. T. Nururrohman and E. Setyati, "Sistem Rekomendasi Penerima Beasiswa pada Calon Mahasiswa Menggunakan Algoritma XGBoost di Universitas Islam Darul 'Ulum Lamongan," *Syntax J. Inform.*, vol. 14, no. 01, pp. 13–24, 2025, doi: 10.35706/syji.v14i01.13070.
[4] L. K. Allen, S. C. Creer, and P. Öncel, "Natural Language Processing: Towards a Multi-Dimensional View of the Learning Process," *Handb. Learn. Anal.*, pp. 46–53, 2022, doi: 10.18608/hla22.005.
[5] D. Sulistyo, F. Ahda, and V. A. Fitria, "Epistomologi dalam Natural Language Processing," *J. Inov. Teknol. dan Edukasi Tek.*, vol. 1, no. 9, pp. 652–664, 2021, doi: 10.17977/um068v1i92021p652-664.
[6] V. Ayumi, M. Purba, and S. Mailana, "Klasifikasi Teks Umpan Balik Kompetensi Sosial Dosen di Perguruan Tinggi Menggunakan Word2Vec dan CNN-1D," *JSAI (Journal Sci. Appl. Informatics)*, vol. 8, no. 2, pp. 564–569, 2025, doi: 10.36085/jsai.v8i2.8763.
[7] M. Ahmednor, Suhartono, and Imamudin, "Klasifikasi Keterampilan Kerja Menggunakan Metode Tf-Idf Dan Decision Tree Pada Data Lowongan Kerja Linkedin," *J. Apl. Dan Inov. Ipteks "Soliditas,"* vol. 8, no. 1, pp. 52–60, 2025, doi: 10.31328/js.v8i1.7152.
[8] R. Amin and A. S. F. Utami, "Prediksi Nilai Ujian Berdasarkan Kebiasaan Siswa Menggunakan Algoritma Random Forest Regressor," *Inf. Syst. Educ. Prof. J. Inf. Syst.*, vol. 10, no. 2, p. 149, 2025, doi: 10.51211/isbi.v10i2.3722.
[9] Gefy Fitry Wijaya and Dwi Yuniarto, "Tinjauan Penerapan Machine Learning pada Sistem Rekomendasi Menggunakan Model Klasifikasi," *Pop. J. Penelit. Mhs.*, vol. 3, no. 4, pp. 144–153, 2024, doi: 10.58192/populer.v3i4.2798.
[10] R. K. Nasirin and R. D. Indahsari, "Model Penilaian Jawaban Esai Berbasis Semantic Understanding Menggunakan LLaMa dan Text Similarity," *J. Edukasi dan Penelit. Inform.*, vol. 11, no. 2, pp. 326–336, 2025.
[11] E. D. Manawan and F. A. Ahda, "RNN Based Customer Service Chatbot for Information Support at Gleamore," *ICoBITS*, vol. 1, pp. 772–780, 2026, doi: 10.32664/icobits.v1.52.
[12] H. A. Salman, A. Kalakech, and A. Steiti, "Random Forest Algorithm Overview," *Babylonian J. Mach. Learn.*, vol. 2024, pp. 69–79, 2024, doi: 10.58496/BJML/2024/007.
[13] Mega Maulina, Y. P. Hiola, Indahwati, and Aam Alamudi, "Performance Evaluation of Multinomial Logistic Regression, Random Forest, and XGBoost Methods in Data Classification," *J. Math. Comput. Stat.*, vol. 8, no. 2, pp. 355–373, 2025, doi: 10.35580/jmathcos.v8i2.8459.
[14] W. D. Ahmad and A. Abu Bakar, "Classification Models for Higher Learning Scholarship Award Decisions," *Asia-Pacific J. Inf. Technol. Multimed.*, vol. 07, no. 02, pp. 131–145, 2018, doi: 10.17576/apjitm-2018-0702-10.
[15] M. Amien, "Sejarah dan Perkembangan Teknik Natural Language Processing (NLP) Bahasa Indonesia: Tinjauan tentang sejarah, perkembangan teknologi, dan aplikasi NLP dalam bahasa Indonesia," no. 2007, pp. 1–7, 2023, [Online]. Available: http://arxiv.org/abs/2304.02746
[16] N. Livia, "Kelayakan Pemberian Kredit Pada Pt Bpr Prima Multi Makmur," vol. 3, pp. 1–11.
[17] Mr. Shrikanth N G, Shraddha D S, Shreeja H S, Sanmitha, and Shravya G Naik, "Decision Support Systems and Automated Platforms for Scholarship Selection: A Comprehensive Review," *Int. J. Adv. Res. Sci. Commun. Technol.*, p. 218, 2025, doi: 10.48175/ijarsct-30424.
[18] N. Rahma *et al.*, "REVIEW OF LITERATURE REVIEW OF DECISION SUPPORT SYSTEM MODELING IN MICRO , SMALL," vol. 11, no. 2, pp. 185–190, 2023.
[19] A. Ahadi, A. Singh, M. Bower, and M. Garrett, "Text Mining in Education—A Bibliometrics-Based Systematic Review," *Educ. Sci.*, vol. 12, no. 3, 2022, doi: 10.3390/educsci12030210.
[20] B. A. Wulandari, R. Norawati, I. Anastasia, A. Ridha, and R. Heryanti, "Penggunaan Portofolio Digital Untuk Mendorong Pembelajaran Refleksi dan Mandiri," *J. Karya Abdi Masy.*, vol. 5, no. 3, pp. 356–362, 2021, [Online]. Available: https://mail.online-journal.unja.ac.id/JKAM/article/view/16220