Terjemahan artikel 10.22146/v14i3.20516

Interpretable Machine Learning untuk Prediksi Penempatan Kerja: Analisis Fitur Berbasis SHAP

Swono Sibagariang

Politeknik Negeri Batam, Kota Batam, Kepulauan Riau 2946, Indonesia

[Diserahkan: 25 April 2025, Direvisi: 18 Juni 2025, Diterima: 14 Juli 2025]
Penulis Korespondensi: Swono Sibagariang (email: swono@polibatam.ac.id)

INTISARI — Pemodelan prediktif berperan penting dalam analisis capaian kinerja lulusan, terutama untuk meramalkan penempatan kerja berdasarkan capaian akademik dan mata kuliah. Penelitian ini bertujuan untuk meningkatkan akurasi prediktif dan interpretabilitas dalam klasifikasi penempatan kerja menggunakan model pemelajaran mesin canggih dan analisis *SHapley Additive exPlanations* (SHAP). Dengan memanfaatkan data akademik berupa nilai mata kuliah, indeks prestasi kumulatif (IPK), dan durasi magang, penelitian ini mengimplementasikan beberapa model klasifikasi, termasuk *decision tree*, *random forest*, *extreme gradient boosting* (XGBoost), *light gradient-boosting machine* (LightGBM), CatBoost, dan regresi logistik. Metrik evaluasi menunjukkan bahwa sebagian besar model mencapai presisi sebesar 92%, *recall* sebesar 92%, dan *F*1-*score* sebesar 92%, dengan akurasi 85%, sedangkan regresi logistik unggul dengan *recall* 100%, *F*1-*score* 96%, dan akurasi 92%. Analisis SHAP mengidentifikasi fitur-fitur utama yang paling berpengaruh dalam memprediksi penempatan kerja, yaitu Administrasi, Organisasi Komputer, Sistem Informasi, Kewirausahaan, Etika Profesional, dan Pemrograman Web. Fitur lain, seperti Pengantar Teknologi Informasi, Rekayasa Perangkat Lunak II, dan Penambangan Data, juga memberikan kontribusi signifikan, meskipun dengan pengaruh yang relatif lebih rendah. Kegiatan ekstrakurikuler dan pengalaman magang juga turut berperan membentuk prospek karier lulusan, menekankan pentingnya unsur akademik dan nonakademik. Temuan ini menyoroti perlunya menyediakan mata kuliah tertentu untuk mempersiapkan mahasiswa menghadapi pasar kerja dengan lebih baik. Temuan ini juga menekankan pentingnya model pemelajaran mesin yang dapat diinterpretasikan dalam prakiraan karier, yang memungkinkan institusi pendidikan untuk mengoptimalkan desain kurikulum dan meningkatkan daya kerja lulusan. Penelitian di masa mendatang sebaiknya mengeksplorasi teknik pemilihan fitur, analisis temporal, dan sistem rekomendasi yang dipersonalisasi untuk menyempurnakan akurasi prediktif.

KATA KUNCI — Pemelajaran Mesin, Prediksi Kesesuaian Kerja, *Shapley Additive Explanations* (SHAP), Penempatan Kerja Lulusan.

I. PENDAHULUAN

Persaingan di pasar kerja bagi lulusan universitas makin ketat, terutama dengan adanya transformasi industri yang pesat dan permintaan keterampilan yang terus berkembang di berbagai sektor. Salah satu tantangan utama yang dihadapi lulusan adalah menemukan pekerjaan yang sesuai dengan bidang studinya. Berdasarkan laporan World Economic Forum tahun 2023, tercatat bahwa lebih dari 30% lulusan universitas memerlukan waktu lebih dari enam bulan untuk mendapatkan pekerjaan pertama [1]. Salah satu penyebab utama masalah ini adalah kesenjangan antara keterampilan akademik yang diperoleh selama kuliah dan keterampilan yang dibutuhkan oleh industri. Kurikulum akademik sering kali tidak selaras dengan kemajuan teknologi dan kebutuhan pasar kerja, sehingga membuat lulusan belum sepenuhnya siap untuk bersaing di dunia profesional.

Saat ini, sebagian besar institusi pendidikan masih mengandalkan metode tradisional dalam memberikan bimbingan karier dan memprediksi peluang kerja lulusan. Pendekatan manual ini memiliki keterbatasan dalam mengintegrasikan berbagai faktor akademik dan nonakademik yang memengaruhi keberhasilan karier lulusan. Oleh karena itu, pendekatan berbasis data menggunakan teknik pemelajaran mesin muncul sebagai solusi potensial untuk memahami pola yang menghubungkan kinerja akademik, pengalaman, dan kesuksesan karier.

Beberapa penelitian telah mengeksplorasi penerapan pemelajaran mesin dalam memprediksi penempatan kerja lulusan perguruan tinggi. Penelitian sebelumnya telah menunjukkan bahwa teknik pemelajaran mesin dapat secara efektif menganalisis berbagai faktor akademik dan mengklasifikasikan status pekerjaan alumni dengan akurasi tinggi. Sebagai contoh, sebuah penelitian yang menggunakan algoritma *random forest* mencapai akurasi sebesar 98% dalam memprediksi penempatan kerja berdasarkan atribut akademik, seperti indeks prestasi kumulatif (IPK), pengalaman magang, dan keikutsertaan dalam kegiatan ekstrakurikuler [2]. Selain itu, faktor-faktor yang berkaitan dengan magang, seperti durasi dan kinerja selama pelatihan, telah terbukti berperan penting dalam memprediksi penempatan kerja, dengan pendekatan pemelajaran mesin *ensemble* (*stacking*) mencapai akurasi hingga 91% [3].

Meskipun penelitian sebelumnya menunjukkan hasil yang menjanjikan, beberapa area masih belum dieksplorasi. Sebagian besar penelitian hanya mempertimbangkan faktor akademik umum, seperti IPK dan pengalaman magang, tanpa mengkaji korelasi antara kinerja pada mata kuliah tertentu dengan jalur karier lulusan. Selain itu, model yang digunakan umumnya hanya berfokus pada metrik akurasi tanpa melakukan analisis interpretabilitas yang mendalam untuk memahami faktor-faktor kunci yang berkontribusi terhadap prediksi.

Penelitian ini menyajikan pendekatan baru dengan menitikberatkan pada analisis terperinci mata kuliah yang memengaruhi prospek pekerjaan lulusan, tidak hanya berdasarkan IPK. Melalui penerapan teknik pemilihan fitur dan *SHapley Additive exPlanations* (SHAP), penelitian ini bertujuan untuk mengidentifikasi faktor akademik dan

nonakademik yang paling berkontribusi pada kesuksesan karier alumni [4], [5]. Penelitian sebelumnya umumnya berfokus pada variabel agregat seperti IPK, status magang, atau kegiatan ekstrakurikuler sebagai prediktor utama, tanpa mengkaji kontribusi spesifik mata kuliah tertentu terhadap karier [2], [3]. Meskipun SHAP telah terbukti membantu memahami prediksi kesuksesan akademik dan kepuasan kerja [4], [6], penggunaannya untuk menilai relevansi kurikulum terhadap keberhasilan memperoleh pekerjaan setelah lulus masih terbatas. Dengan demikian, penelitian ini tidak hanya menawarkan pendekatan yang lebih terfokus pada pemilihan fitur, tetapi juga mengevaluasi model berdasarkan akurasi dan menekankan pentingnya interpretabilitas dalam memberikan wawasan mendalam kepada pembuat kebijakan dalam meningkatkan kesiapan kerja lulusan.

Penelitian ini berkontribusi signifikan dalam mendukung institusi pendidikan tinggi dalam menciptakan kebijakan berbasis data yang relevan dengan tuntutan pasar kerja. Dengan hasil prediksi dan interpretasi model yang lebih transparan, institusi dapat merancang kurikulum dan pelatihan yang lebih terarah. Dalam hal program dan layanan bimbingan karier, penelitian ini berpotensi membantu lulusan mendapatkan pekerjaan yang sesuai dengan latar belakang pendidikan dengan lebih cepat dan efisien, sehingga meningkatkan daya saing lulusan di pasar kerja global dan mengurangi tingkat pengangguran terdidik.

Tujuan utama penelitian ini adalah mengembangkan model pemelajaran mesin yang mampu memprediksi kemampuan kerja lulusan berdasarkan data akademik dan pengalaman nonakademik. Untuk mencapai hal tersebut, penelitian ini mengkaji efektivitas berbagai algoritma pemelajaran mesin, termasuk *decision tree*, *random forest*, *extreme gradient boosting* (XGBoost), *light gradient-boosting machine* (LightGBM), CatBoost, *gradient boosting machine* (GBM), dan regresi logistik, dalam mengklasifikasikan status kerja lulusan. Model yang dikembangkan dievaluasi menggunakan metrik seperti akurasi, presisi, *recall*, dan *F*1-*score* untuk memastikan keandalan prediksi.

Metodologi penelitian ini mencakup pengumpulan data akademik dan nonakademik dari alumni, seperti informasi pekerjaan, nilai mata kuliah, IPK, pengalaman magang, dan keterlibatan organisasi. Data yang terkumpul diproses melalui teknik prapemrosesan, termasuk normalisasi nilai, pengenkodean fitur kategoris, dan pemilihan fitur, untuk mengidentifikasi variabel yang paling memengaruhi penempatan kerja. Selain itu, model prediktif diuji menggunakan teknik validasi silang dan dianalisis menggunakan SHAP untuk menginterpretasikan faktor-faktor kunci yang berkontribusi terhadap prediksi penempatan kerja.

II. PENELITIAN TERKAIT

Prediksi terhadap hasil penempatan kerja lulusan perguruan tinggi telah menjadi fokus utama pada berbagai penelitian, seiring dengan meningkatnya penerapan pengambilan keputusan berbasis data dalam pendidikan tinggi dan manajemen sumber daya manusia. Metode prediksi tradisional umumnya mengandalkan penilaian subjektif, tren ketenagakerjaan historis, dan evaluasi manual kualifikasi mahasiswa. Namun, pendekatan-pendekatan ini sering kali terbatas dalam hal akurasi, skalabilitas, dan objektivitas. Sebaliknya, teknik pemelajaran mesin menawarkan alternatif yang lebih unggul dengan kemampuan memproses *dataset* besar dan mengungkap pola-pola kompleks yang memengaruhi hasil penempatan kerja. Sejumlah penelitian telah mengkaji penerapan berbagai algoritma pemelajaran mesin dalam domain ini dan melaporkan hasil yang menjanjikan terkait akurasi dan reliabilitas prediktif.

Beberapa studi telah meneliti penggunaan model pemelajaran mesin untuk memprediksi hasil penempatan kerja bagi lulusan. Sebuah studi menguji prediksi penempatan mahasiswa melalui beragam metodologi pemelajaran mesin. Studi tersebut menunjukkan bahwa jaringan saraf mencapai tingkat akurasi 85% dalam meramalkan pekerjaan lulusan dalam kurun waktu enam bulan setelah kelulusan [7]. Sebuah studi terpisah menggunakan regresi logistik, yang mengintegrasikan kinerja akademik dengan data demografi, dan mencapai akurasi prediksi sebesar 87% [8]. Sebuah studi menggunakan model *decision tree* untuk meramalkan hasil penempatan kerja berdasarkan prestasi akademik dan kegiatan ekstrakurikuler dan mencapai akurasi sebesar 89% [9]. Studi-studi ini menggarisbawahi keandalan metodologi pemelajaran mesin dalam menilai kemampuan kerja lulusan dan meningkatkan prakiraan penempatan kerja.

Temuan ini diperkuat oleh beberapa studi lanjutan. Sebuah studi menunjukkan bahwa pengklasifikasi *decision tree* dapat secara akurat memprakirakan kapasitas kerja mahasiswa berdasarkan data prestasi akademik [10]. Sebuah studi terpisah menunjukkan bahwa penggunaan model XGBoost pada data akademik lulusan dapat secara akurat memprakirakan jalur karier mahasiswa [11]. Di sisi lain, pendekatan berbasis *recurrent neural network* (RNN) digunakan dalam sistem rekomendasi pekerjaan. Data masukan diubah menjadi representasi vektor melalui doc2Vec untuk mempertahankan signifikansi semantik [12]. Sebuah model gabungan *convolutional neural netwrok* (CNN)-RNN yang dilengkapi dengan mekanisme *self-attention* telah dikembangkan untuk memprakirakan kinerja karyawan. Model ini secara efektif merangkum informasi semantik dan struktural, sehingga menghasilkan akurasi prediktif yang lebih baik [13].

Meskipun banyak penelitian telah memvalidasi efektivitas teknik pemelajaran mesin dalam memprediksi hasil penempatan kerja, variasi akurasi dan kinerjanya bergantung pada beberapa faktor, termasuk pemilihan fitur, kualitas *dataset*, dan pemilihan model. Penelitian ini bertujuan untuk mengembangkan model prediktif yang lebih baik dengan mengintegrasikan berbagai pendekatan pemelajaran mesin, termasuk *decision tree*, *random forest*, XGBoost, LightGBM, CatBoost, GBM, dan regresi logistik. Penelitian ini bertujuan untuk menyediakan kerangka kerja yang kuat dan dapat diinterpretasikan untuk memprediksi hasil kerja lulusan dengan memanfaatkan berbagai fitur yang mencakup atribut akademik, ekstrakurikuler, dan personal.

Secara keseluruhan, literatur yang ada menekankan potensi pemelajaran mesin dalam mengoptimalkan prediksi penempatan kerja. Namun, diperlukan penelitian lebih lanjut untuk menyempurnakan model-model ini dan meningkatkan penerapannya di berbagai disiplin ilmu akademik dan konteks geografis. Kontribusi utama dari studi ini terletak pada eksplorasi kinerja komparatif beberapa algoritma pemelajaran mesin dan mengusulkan model optimal yang dapat membantu perguruan tinggi dan perusahaan dalam menyederhanakan proses penempatan kerja.

III. METODOLOGI

Penelitian ini dilakukan menggunakan metodologi terstruktur yang mencakup tahap pengumpulan data,

prapemrosesan, pengembangan model, evaluasi, dan analisis interpretabilitas. Perangkat lunak yang digunakan dalam penelitian ini mencakup berbagai perangkat dan pustaka dalam ekosistem Python yang mendukung analisis data dan implementasi algoritma pemelajaran mesin. Data yang terkumpul diproses menggunakan pustaka Pandas, yang memungkinkan impor dan manipulasi data dalam format *spreadsheet* secara efisien. Untuk analisis numerik, penelitian ini menggunakan NumPy, yang menyediakan berbagai fungsi matematika untuk memproses data secara optimal.

A. PENGUMPULAN DATA

Studi ini menggunakan data lulusan program D3 Teknik Informatika, Politeknik Negeri Batam tahun 2019 melalui survei, catatan institusi, dan data dari *tracer study* alumni. Data ini mencakup berbagai faktor akademik dan nonakademik yang dapat memengaruhi penempatan kerja lulusan.

Faktor akademik dalam penelitian ini menunjukkan capaian akademik mahasiswa selama masa studi. Data yang digunakan berupa IPK dan nilai berbagai mata kuliah yang relevan dengan kurikulum D3 Teknik Informatika. Mata kuliah yang dianalisis meliputi Tugas Akhir I dan II, Bahasa Inggris I dan II, Etika Profesi, Pelaporan Kerja, Administrasi, Pemrograman Dasar, Penambangan Data, Jaringan Komputer, Keselamatan & Kesehatan Kerja, Kewirausahaan, Matematika, Pemrograman Basis Data, Pemrograman Perangkat Keras, Pemrograman Berorientasi Objek, Pemrograman Web, Pengantar Basis Data, Pengantar Teknologi Informasi, Sistem Informasi, Statistika, Pancasila, Kecerdasan Buatan, Multimedia, Pemrograman Perangkat Seluler, Organisasi Komputer, Sistem Operasi, Jaringan Komputer Lanjut, dan Rekayasa Perangkat Lunak I dan II. Nilai dari mata kuliah ini dikonversi ke dalam skala numerik (0–4).

Kategori informasi pekerjaan mencakup berbagai data tentang status pekerjaan alumni setelah lulus. Data ini meliputi informasi lulusan yang bekerja, menganggur, wiraswasta, magang, atau menjadi pekerja lepas. Informasi mengenai tempat kerja, jenis perusahaan (seperti sektor industri, lembaga pendidikan, atau organisasi nirlaba), serta skala perusahaan (lokal, nasional, atau multinasional) juga dianalisis. Faktor-faktor lain yang juga dikaji meliputi pendapatan dalam rupiah, waktu tunggu untuk mendapatkan pekerjaan setelah lulus, dan tingkat relevansi pekerjaan dengan bidang studi yang ditempuh.

Penelitian ini juga menggunakan data nonakademik yang dapat memengaruhi keberhasilan lulusan dalam mendapatkan pekerjaan. Data ini mencakup pengalaman organisasi yang telah diikuti oleh alumni, termasuk peran alumni dalam organisasi tersebut: anggota, administrator, ketua, atau memegang peran lain. Pengalaman magang juga merupakan aspek penting dalam kategori ini, termasuk lama magang dan bidang industri tempat magang dilakukan, seperti teknologi informasi, manufaktur, atau keuangan.

B. PRAPEMROSESAN DATA

Prapemrosesan data merupakan langkah penting dalam pengembangan model pemelajaran mesin. Langkah ini memastikan data yang digunakan berkualitas tinggi dan siap untuk analisis lebih lanjut. Tahapan utama dalam prapemrosesan data meliputi pembersihan data, pengenkodean fitur, dan normalisasi.

Pembersihan data (*data cleaning*) adalah proses mengidentifikasi dan menangani anomali dalam suatu *dataset*, seperti nilai yang hilang, duplikat, atau inkonsisten. Langkah ini berperan penting untuk meningkatkan akurasi model dan mencegah bias dalam analisis. Teknik yang umum digunakan meliputi imputasi nilai yang hilang, penghapusan duplikat, dan koreksi kesalahan ketik. Pembersihan data yang efektif meningkatkan kinerja model pemelajaran mesin secara signifikan [14].

Dataset sering kali berisi variabel kategoris yang perlu diubah ke dalam format numerik, sehingga algoritma pemelajaran mesin dapat memprosesnya. Metode pengenkodean yang umum digunakan adalah *label encoding* dalam pemelajaran mesin dan prapemrosesan data. *Label encoding* adalah metode yang mengubah data kategoris yang direpresentasikan sebagai label teks menjadi format numerik. Dengan kata lain, metode ini menetapkan nilai numerik unik untuk setiap kategori atau label dalam variabel kategoris [15]. Pengenkodean satu aktif (*one hot encoding*) dalam pemrosesan data dan pemelajaran mesin mengacu pada teknik yang digunakan untuk merepresentasikan variabel kategoris sebagai vektor biner. Setiap kategori atau label diubah menjadi vektor biner dengan panjang yang sama, dengan jumlah total kategori yang berbeda dalam variabel tersebut. Semua elemen vektor memiliki nilai nol, kecuali indeks yang sesuai dengan kategori, yang direpresentasikan oleh angka 1 [16]. Pemilihan metode pengenkodean yang tepat dapat memengaruhi kinerja model secara keseluruhan.

Normalisasi adalah proses penskalaan fitur numerik ke dalam rentang tertentu, biasanya antara 0 dan 1, untuk memastikan bahwa setiap fitur berkontribusi secara proporsional terhadap model. Salah satu metode normalisasi yang umum adalah penskalaan *min-max*, yang dirumuskan seperti pada (1).

$$x' \frac{x - x_{min}}{x_{max} - x_{min}}. \quad (1)$$

Dalam proses normalisasi, x' mewakili nilai yang dinormalisasi, x merupakan nilai asli, sedangkan x_{min} dan x_{max} merupakan nilai minimum dan maksimum fitur [17]. Normalisasi yang tepat mencegah fitur dengan rentang numerik yang lebih besar mendominasi analisis [18].

C. PEMELAJARAN MESIN

Pengembangan model pemelajaran mesin berperan penting dalam membangun sistem prediksi yang andal. Berbagai algoritma pemelajaran mesin diterapkan dan dibandingkan untuk mengidentifikasi model dengan kinerja terbaik dalam memprediksi penempatan kerja lulusan. Model yang digunakan meliputi *decision tree*, *random forest*, XGBoost, LightGBM, CatBoost, GBM, dan regresi logistik.

Decision tree adalah algoritma yang memiliki struktur berbasis aturan yang menyerupai pohon dengan fitur-fitur yang disusun secara hierarkis [19]. Algoritma ini membagi data menjadi subset berdasarkan atribut tertentu, sehingga membentuk struktur seperti pohon yang membantu pengambilan keputusan. Keunggulan utama *decision tree* adalah kesederhanaannya dan kemampuannya untuk diinterpretasikan dengan mudah.

Random forest adalah teknik pemelajaran ensambel yang membangun beberapa *decision tree* selama pelatihan dan menggabungkan hasilnya untuk memperoleh keputusan akhir. Metode ini secara efektif menangani berbagai variabel dan data dalam jumlah besar, serta dapat mengurangi *overfitting* dengan menggabungkan prediksi dari beberapa *decision tree*, sehingga dihasilkan model yang lebih kuat dan lebih tergeneralisasi [20].

XGBoost adalah algoritma *boosting* yang dikenal karena akurasinya yang tinggi. XGBoost bekerja dengan membangun

model secara iteratif, dengan setiap model baru dirancang untuk memperbaiki kesalahan model sebelumnya. Teknik ini telah terbukti unggul dalam berbagai kompetisi dan aplikasi pemelajaran mesin, terutama dalam menangani data yang kompleks dan beragam [21].

LightGBM adalah kerangka kerja *gradient boosting* yang dioptimalkan untuk efisiensi dan skalabilitas dalam aplikasi pemelajaran mesin. Kerangka kerja ini menggunakan pemelajaran berbasis histogram untuk mempercepat pelatihan, sehingga memungkinkan pengelolaan *dataset* berukuran besar dengan penggunaan memori yang lebih rendah. Berbeda dengan teknik *boosting* konvensional, LightGBM membangun pohon secara *leaf-wise*, sehingga dapat meningkatkan akurasi dan mengurangi *overfitting* [22].

CatBoost adalah algoritma *gradient boosting* yang dirancang khusus untuk menangani fitur kategorikal secara efisien. Berbeda dengan algoritma *gradient boosting* tradisional yang memerlukan prapemrosesan ekstensif untuk variabel kategorikal, CatBoost menggabungkan mekanisme pengenkodean canggih, yang memungkinkan data kategorikal diproses langsung tanpa mengurangi daya prediktifnya. Fitur utama CatBoost adalah *ordered boosting*, yang berfungsi mengurangi *target leakage* dan *overfitting* dengan memastikan bahwa setiap data pelatihan hanya menggunakan informasi historis dalam proses prediksi [23].

GBM adalah metodologi pemelajaran ensambel yang membangun model prediktif dengan menggabungkan beberapa pemelajar lemah secara bertahap, biasanya dalam bentuk *decision tree*. Setiap model selanjutnya dilatih untuk mengoreksi kesalahan residual dari ensambel sebelumnya, sehingga meningkatkan akurasi prediktif secara keseluruhan [24].

Regresi logistik adalah teknik statistik yang banyak digunakan untuk memodelkan hasil biner, dengan variabel dependen dapat mengambil salah satu dari dua kemungkinan nilai, biasanya dilambangkan sebagai 0 atau 1. Regresi logistik menggunakan fungsi logistik (*sigmoid*) pada kombinasi linear data masukan untuk menentukan kemungkinan suatu masukan tertentu dikaitkan dengan kelas tertentu. Tidak seperti regresi linear yang memprediksi nilai kontinu, regresi logistik menghasilkan skor probabilitas berkisar antara 0 hingga 1. Setiap karakteristik masukan diberi bobot (koefisien) dan jumlah terbobot tersebut ditransformasikan menggunakan fungsi *sigmoid* untuk menghasilkan probabilitas. Ambang batas, yang sering ditetapkan sebesar 0,5, kemudian digunakan untuk mengklasifikasikan observasi ke dalam salah satu dari dua kategori.

Pendekatan ini secara efektif memprediksi probabilitas terjadinya suatu peristiwa berdasarkan satu atau lebih variabel prediktor. Salah satu keunggulan utama regresi logistik adalah interpretabilitasnya; koefisien model menunjukkan kekuatan dan arah hubungan antara setiap variabel dan hasil yang diprediksi. Regresi logistik didasarkan pada asumsi bahwa variabel independen terhubung secara linear dengan *log odd* peluang variabel dependen. Namun, asumsi ini tidak berlaku pada situasi yang kompleks atau nonlinear. Selain itu, regresi logistik mensyaratkan observasi bersifat independen dan mengasumsikan multikolinearitas minimal di antara para prediktor [25].

D. SHAPLEY ADDITIVE EXPLANATIONS (SHAP)

Pendekatan SHAP adalah metode yang digunakan untuk menginterpretasikan model prediksi pemelajaran mesin *black box* atau yang sulit dipahami [26]. Tujuan SHAP adalah menguraikan prediksi fitur xxx dengan menghitung kontribusi setiap fitur terhadap prediksi [27]. Metode ini menghitung nilai Shapley dengan cara yang serupa dengan teori *coalition game* [28]. Pada konteks ini, nilai fitur dari data dianggap sebagai pemain dalam sebuah koalisi. Nilai Shapley mendistribusikan kontribusi prediksi ke fitur-fitur secara adil. Setiap pemain mewakili nilai fitur individual, sehingga memungkinkan pemahaman yang lebih mendalam tentang faktor-faktor yang memengaruhi hasil prediksi.

Pendekatan SHAP ditunjukkan pada (2):

$$\phi_i = \sum_{S\subseteq\{1,\dots,p\}\setminus\{i\}} \frac{|S|!(p-|S|-1)!}{p!} \times [Val(S \cup \{i\}) - Val(S)] \quad (2)$$

dengan ϕ_i merupakan nilai Shapley dari anggota fitur sehubungan dengan hasil prediksi, $Val(S)$ merupakan keluaran model pemelajaran mesin yang akan dijelaskan menggunakan serangkaian fitur S, dan p merupakan jumlah total fitur.

Kontribusi akhir atau nilai Shapley dari suatu fitur $i(\phi i)$ juga didefinisikan sebagai rata-rata kontribusi marginalnya di semua kemungkinan permutasi himpunan fitur. Proses ini menggunakan perhitungan nilai Shapley dengan menggabungkan semua kemungkinan kombinasi fitur dan mengukur berubahnya kontribusi setiap fitur seiring dengan perubahan fitur lainnya.

IV. HASIL DAN PEMBAHASAN

Bagian ini menganalisis secara rinci faktor-faktor yang memengaruhi peluang kerja lulusan dan membandingkan algoritma pemelajaran mesin yang diterapkan untuk mengidentifikasi model-model berkinerja terbaik dalam memprediksi penempatan kerja lulusan.

A. DESKRIPSI DATASET

Data untuk model ini berasal dari *dataset* alumni. Variabel yang digunakan dalam studi ini mencakup semua parameter yang memengaruhi penempatan kerja lulusan. Dalam *dataset* ini, terdapat 43 atribut yang mencakup informasi akademik spesifik tentang lulusan, seperti informasi pekerjaan, kinerja akademik, keterlibatan dalam organisasi, dan pengalaman magang. Informasi *dataset* ini dapat diperlihatkan pada Tabel I.

Pengenkodean diterapkan pada beberapa atribut kategoris guna mempersiapkan data untuk pemodelan. *Label encoding* diterapkan pada fitur biner atau ordinal seperti status pekerjaan, relevansi pekerjaan dengan studi, magang industri, dan aktivitas organisasi. Metode ini mengubah nilai kategoris menjadi representasi numerik sederhana, seperti “Ya” menjadi 1 dan “Tidak” menjadi 0. Sementara itu, pengenkodean satu aktif diterapkan pada fitur dengan beberapa kategori nominal tanpa urutan inheren, seperti tempat kerja, jenis perusahaan, skala perusahaan, posisi dalam organisasi, durasi magang, dan bidang magang. Pengenkodean satu aktif mengubah setiap kategori unik dalam fitur-fitur ini menjadi kolom biner terpisah, yang memungkinkan algoritma pemelajaran mesin untuk memproses informasi tanpa menyiratkan hubungan ordinal apa pun. Langkah prapemrosesan ini memastikan bahwa semua data masukan berada dalam format numerik yang sesuai untuk model pemelajaran mesin yang digunakan dalam studi.

B. METODE EVALUASI

Studi ini menilai kinerja model melalui *confusion matrix*, yang merupakan instrumen penting untuk mengukur akurasi model dalam klasifikasi data. Matriks ini memungkinkan

TABEL I
INFORMASI *DATASET*

No.	Nama Atribut	Deskripsi Variabel	Tipe Atribut
1	NIM	Nomor induk mahasiswa	*String*
2	Nama	Nama lengkap siswa	*String*
3	IPK	Indeks prestasi kumulatif	Numerik
4	Tugas Akhir I	Nilai Tugas Akhir I (0-4)	Numerik
5	Tugas Akhir II	Nilai Tugas Akhir II (0-4)	Numerik
6	Bahasa Inggris I	Nilai Bahasa Inggris I (0-4)	Numerik
7	Bahasa Inggris II	Nilai Bahasa Inggris II (0-4)	Numerik
8	Etika Profesional	Nilai Etika Profesional (0-4)	Numerik
9	Pelaporan Pekerjaan	Nilai Pelaporan Pekerjaan (0-4)	Numerik
10	Administrasi	Nilai Administrasi (0-4)	Numerik
11	Dasar-Dasar Pemrograman	Nilai Dasar-Dasar Pemrograman (0-4)	Numerik
12	*Data Mining*	Nilai Data Mining (0-4)	Numerik
13	Jaringan Komputer	Nilai Jaringan Komputer (0-4)	Numerik
14	Kesehatan & Keselamatan Kerja	Nilai Kesehatan & Keselamatan Kerja (0-4)	Numerik
15	Kewirausahaan	Nilai Kewirausahaan (0-4)	Numerik
16	Matematika	Nilai Matematika (0-4)	Numerik
17	Pemrograman Basis Data	Nilai Pemrograman Basis Data (0-4)	Numerik
18	Pemrograman Perangkat Keras	Nilai Pemrograman Perangkat Keras (0-4)	Numerik
19	Pemrograman Berorientasi Objek	Nilai Pemrograman Berorientasi Objek (0-4)	Numerik
20	Pemrograman Web	Nilai Pemrograman Web (0-4)	Numerik
21	Pengantar Basis Data	Nilai Pengantar Basis Data (0-4)	Numerik
22	Pengantar TI	Nilai Pengantar TI (0-4)	Numerik
23	Sistem Informasi	Nilai Sistem Informasi (0-4)	Numerik
24	Statistik	Nilai Statistik (0-4)	Numerik
25	Pancasila	Nilai Pancasila (0-4)	Numerik
26	Kecerdasan Buatan	Nilai Kecerdasan Buatan (0-4)	Numerik
27	Multimedia	Nilai Multimedia (0-4)	Numerik
28	Pemrograman Seluler	Nilai Pemrograman Seluler (0-4)	Numerik
29	Organisasi Komputer	Nilai Organisasi Komputer (0-4)	Numerik
30	Sistem Operasi	Nilai Sistem Operasi (0-4)	Numerik
31	Jaringan Komputer Lanjutan	Nilai Jaringan Komputer Lanjutan (0-4)	Numerik
32	Rekayasa Perangkat Lunak I	Nilai Rekayasa Perangkat Lunak I (0-4)	Numerik
33	Rekayasa Perangkat Lunak II	Nilai Rekayasa Perangkat Lunak II (0-4)	Numerik
34	Apakah Anda pernah bergabung dengan suatu organisasi?	Bergabung tidaknya mahasiswa pada organisasi	Boolean (Ya/Tidak)
35	Jabatan di Organisasi	Peran di organisasi (anggota, pengurus, ketua, lainnya)	*String*
36	Industri Magang	Magang tidaknya mahasiswa	Boolean (Ya/Tidak)
37	Durasi Magang	Lama magang (bulan)	Numerik
38	Bidang Magang	Bidang magang (TI, manufaktur, keuangan, dll.)	*String*
39	Status Pekerjaan	Status pekerjaan saat ini (bekerja, tidak bekerja)	*String*
40	Tempat Kerja	Nama tempat bekerja	*String*
41	Jenis Perusahaan	Sektor industri (industri, institusi pendidikan, nonprofit, dll.)	*String*
42	Skala Perusahaan	Ukuran perusahaan (lokal, nasional, multinasional/internasional)	*String*
43	Relevansi Pekerjaan dengan Studi	Tingkat relevansi pekerjaan terhadap bidang studi (relevan, tidak relevan)	*String*

penentuan jumlah prediksi benar dan salah, sehingga memudahkan perhitungan metrik-metrik penting seperti akurasi, presisi, *recall*, dan *F*1-*score* [29]. Perhitungan metrik-metrik ini menggunakan rumus-rumus yang banyak digunakan, yaitu dengan membandingkan total prediksi akurat (*true positive* dan *true negative*) dengan jumlah keseluruhan prakiraan, sebagaimana ditunjukkan dalam (3).

$$Akurasi = \frac{TP+TN+FP+FN}{TP+TN}. \quad (3)$$

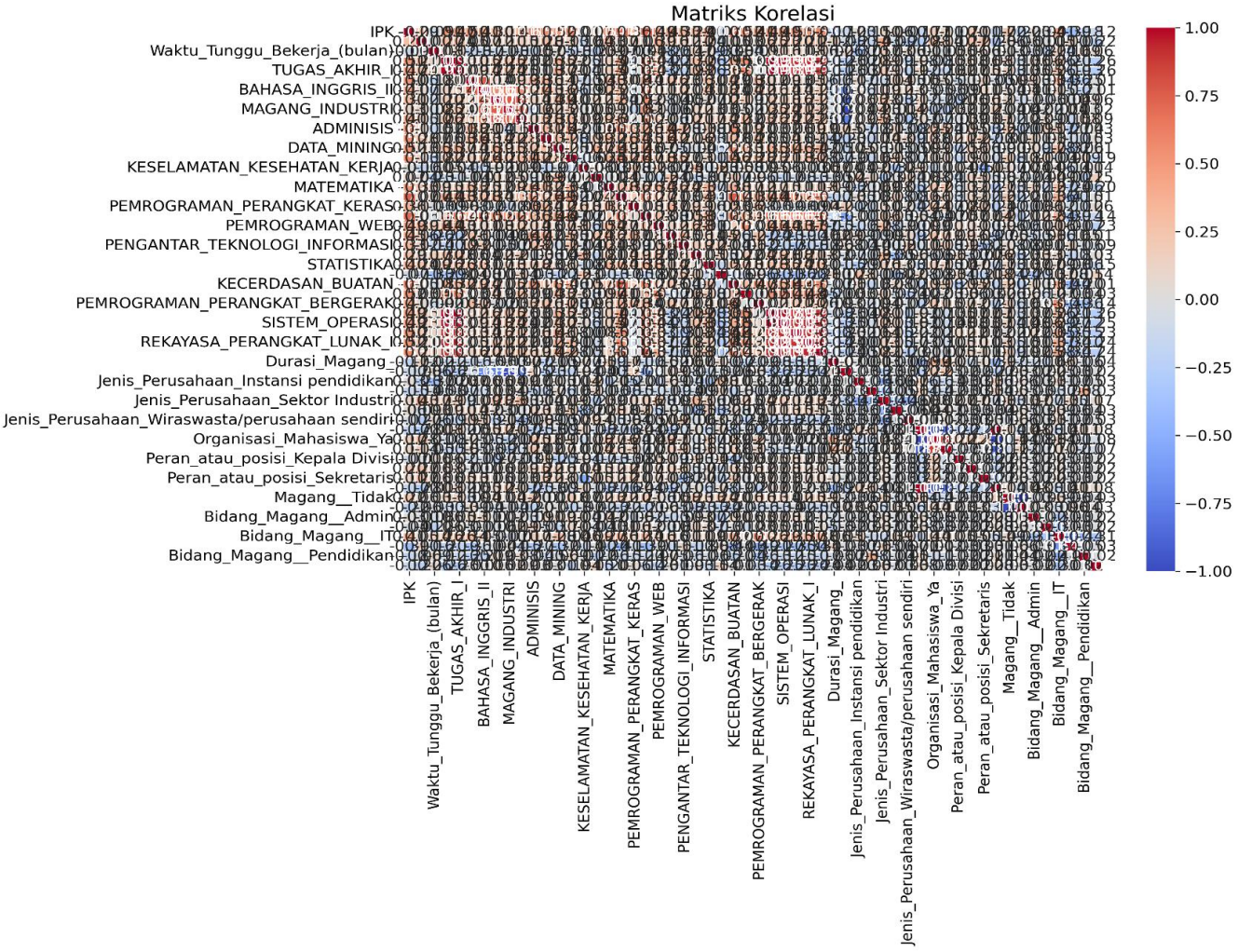
True positive (TP) menunjukkan jumlah data positif yang diidentifikasi secara akurat oleh model sebagai positif. Sebaliknya, *false positive* (FP) menunjukkan jumlah data negatif yang diklasifikasikan sebagai positif. *False negative* (FN) menunjukkan jumlah data positif yang dikategorikan oleh model sebagai negatif. *True negative* (TN) menunjukkan jumlah data negatif yang diidentifikasi secara akurat oleh model sebagai negatif [30].

C. PARAMETER MODEL

Beberapa model pemelajaran mesin diimplementasikan menggunakan Python untuk menjalankan tugas klasifikasi dan prediksi. Model-model tersebut dibangun dan dievaluasi menggunakan pustaka scikit-learn, XGBoost, LightGBM, dan

TABELII
KONFIGURASI PARAMETER

Model Klasifikasi	Konfigurasi Variabel
RandomForestClassifier	*n_estimators* = 100, *random_state* = 42
XGBClassifier	*use_label_encoder* = False, *eval_metric* = 'logloss', *random_state* = 42
DecisionTreeClassifier	*random_state* = 42
GradientBoostingClassifier	*random_state* = 42
LGBMClassifier	*random_state* = 42, *verbose* = -1
CatBoostClassifier	*verbose* = 0, *random_state* = 42
LogisticRegression	*max_iter* = 1.000, *random_state* = 42



Gambar 1. Matriks korelasi.

CatBoost. Selain itu, kerangka kerja metrik klasifikasi digunakan untuk menilai kinerja model, sedangkan Pandas dan NumPy digunakan untuk manipulasi dan perbandingan data. Setiap model dikonfigurasi dengan parameter yang telah disesuaikan secara spesifik untuk mengoptimalkan kinerja dan memastikan konsistensi selama pelatihan dan pengujian. Konfigurasi parameter yang digunakan dalam studi ini disajikan pada Tabel II.

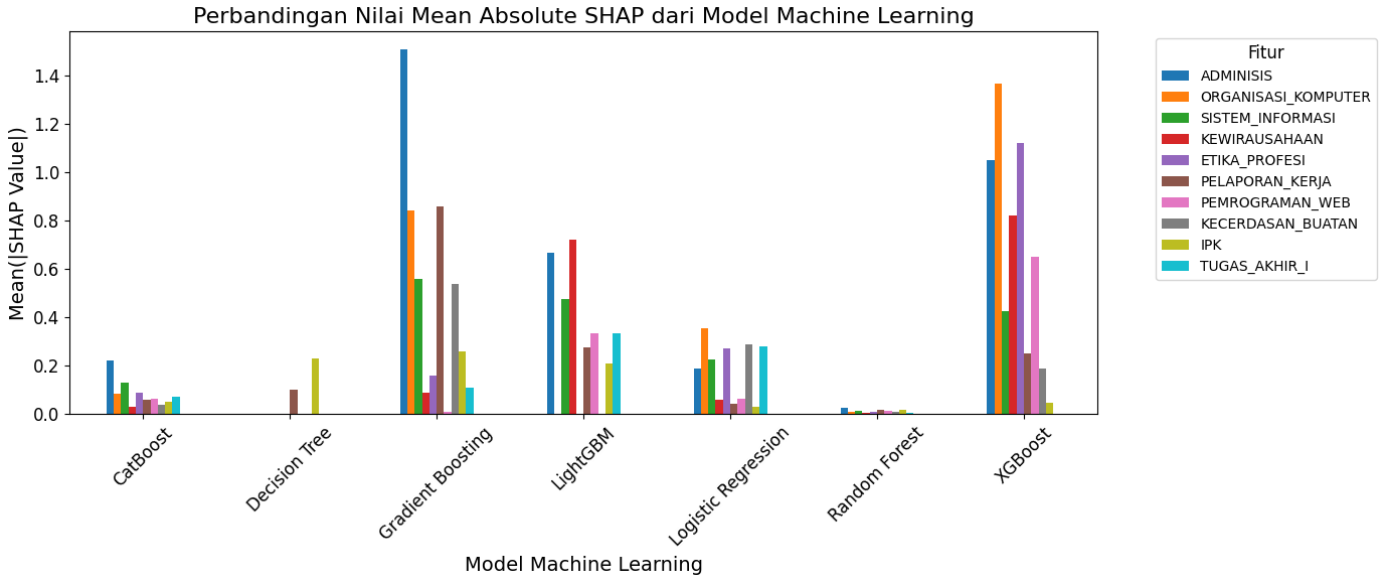
D. ANALISIS MODEL

Analisis korelasi dilakukan untuk lebih memahami hubungan antarvariabel dalam *dataset*. Langkah ini bertujuan untuk mengidentifikasi atribut yang sangat terkait dan relatif independen, sehingga memberikan wawasan tentang potensi pentingnya fitur dan *redundancy*. Hasil analisis ini divisualisasikan menggunakan *correlation heatmap*, sebagaimana ditunjukkan pada Gambar 1.

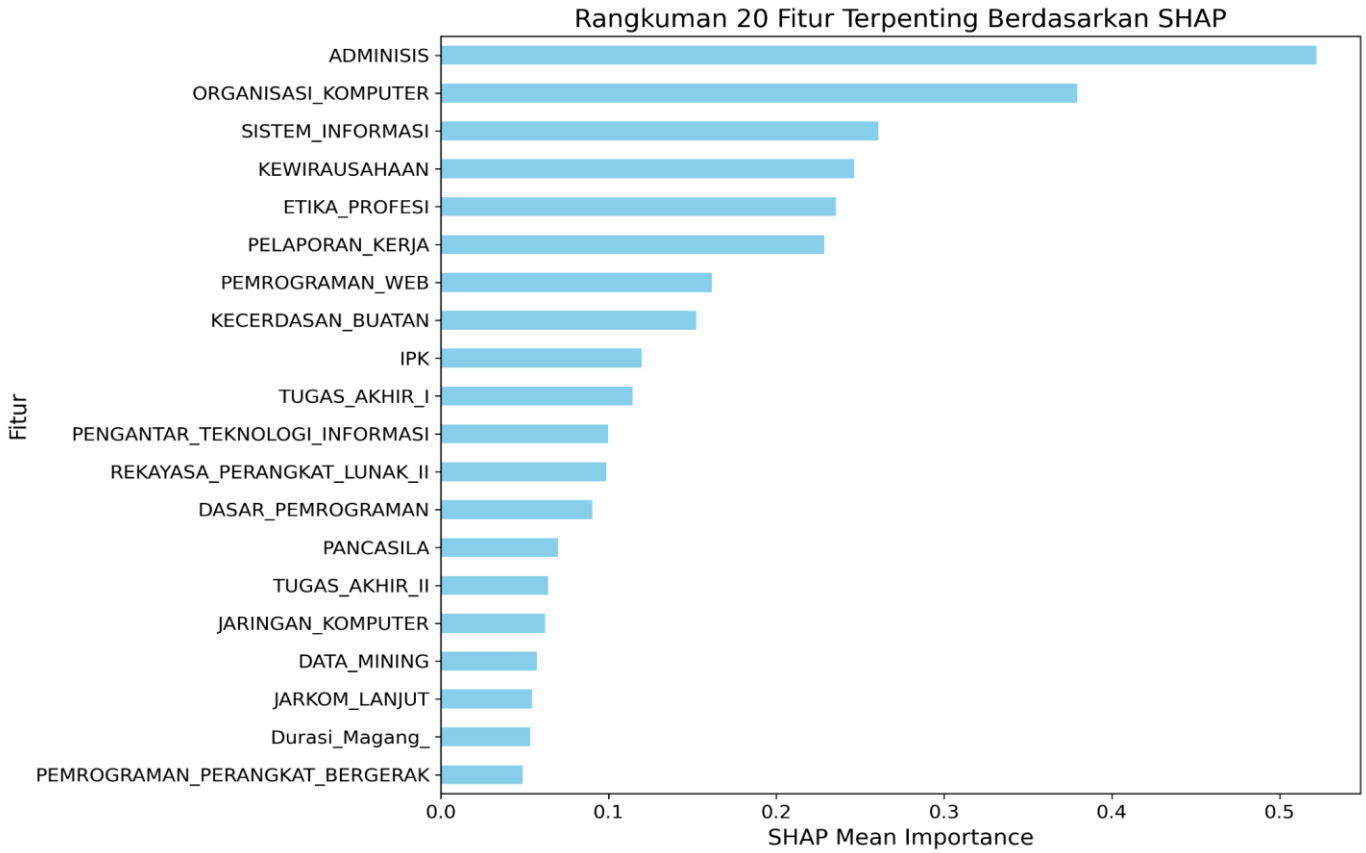
TABEL III
HASIL SETIAP MODEL EVALUASI

Model	Presisi	*Recall*	*F1-Score*	Akurasi
Decision Tree	92%	92%	92%	85%
Random Forest	92%	92%	92%	85%
Extreme Gradient Boosting	92%	92%	92%	85%
LightGBM	92%	92%	92%	85%
CatBoost	92%	92%	92%	85%
Gradient Boosting Machines	92%	92%	92%	85%
Regresi Logistik	92%	100%	96%	92%

Pada Gambar 1, *correlation heatmap* menunjukkan keterkaitan setiap atribut *dataset* dengan semua atribut lainnya. Warna yang lebih terang menunjukkan kurangnya korelasi antara dua variabel, sedangkan warna biru yang lebih gelap



Gambar 2. Hasil SHAP absolut rata-rata untuk setiap model.



Gambar 3. Fitur paling berpengaruh.

menunjukkan korelasi yang lebih kuat dalam rentang -1,0 hingga 1,0. Korelasi sampel yang meningkat dari waktu ke waktu mencerminkan kualitas *dataset* sintetis, dengan sebagian besar atribut menunjukkan korelasi yang tinggi satu sama lain di seluruh *dataset*. Kinerja model pemelajaran mesin dievaluasi berdasarkan akurasi, *recall*, presisi, dan *F*1-*score*. Hasil evaluasi detail untuk setiap model ditunjukkan pada Tabel III.

Evaluasi terhadap berbagai model pemelajaran mesin untuk tugas klasifikasi menunjukkan kinerja yang konsisten dan tinggi pada sejumlah metrik. Hasil yang diperoleh semuanya mencapai skor yang identik, dengan presisi 92%, *recall* 92%, dan *F*1-*score* 92%. Sementara itu, tingkat akurasi tetap pada angka 85%. Konsistensi ini mengindikasikan bahwa model-model ini secara efektif menyeimbangkan presisi dan *recall*, meskipun masih memiliki ruang untuk peningkatan dalam akurasi keseluruhan. Di antara model-model yang diuji, regresi logistik menunjukkan kinerja paling menonjol, dengan *recall* sempurna sebesar 100% dan presisi 92%. Model ini menghasilkan *F*1-*score* sebesar 96% dan peningkatan akurasi keseluruhan sebesar 92%, yang mengindikasikan bahwa model ini secara akurat mengidentifikasi semua kasus positif tanpa melewatkan contoh yang relevan. Berdasarkan hasil tersebut, regresi logistik tampaknya menjadi model yang paling andal untuk menangkap semua contoh yang relevan sambil

mempertahankan presisi yang substansial. Namun, *trade-off* antara *recall* dan akurasi perlu dipertimbangkan dengan cermat berdasarkan kebutuhan spesifik dari tugas klasifikasi.

E. ANALISIS INTERPRETABILITAS MODEL

Dalam pengembangan sistem prediktif berbasis pemelajaran mesin, interpretabilitas model merupakan aspek krusial untuk memahami pengaruh fitur masukan terhadap hasil prediksi. Aspek ini memungkinkan pengguna, pengambil keputusan, dan pengembang sistem untuk memahami alasan di balik setiap keputusan yang dibuat oleh model, sehingga meningkatkan kepercayaan dan akuntabilitas. Guna mendukung analisis interpretabilitas, penelitian ini menggunakan pendekatan SHAP, yaitu metode *explainable AI* (XAI) yang dirancang untuk mengilustrasikan keluaran model pemelajaran mesin, baik secara individual maupun agregat. SHAP didasarkan pada teori nilai Shapley dari teori permainan kooperatif, yang memberikan kontribusi "adil" dari setiap fitur terhadap keluaran prediktif. Pada penelitian ini, SHAP digunakan untuk mengukur kontribusi setiap fitur (seperti mata kuliah tertentu, IPK, dan durasi magang) terhadap prediksi status pekerjaan lulusan. Evaluasi dilakukan dengan menggunakan nilai SHAP absolut rata-rata, yang merepresentasikan kepentingan keseluruhan setiap fitur tanpa mempertimbangkan arah pengaruhnya (positif atau negatif). Makin tinggi nilai SHAP suatu fitur, makin besar pengaruhnya terhadap hasil prediksi model.

F. ANALISIS SHAP

Analisis interpretabilitas model dilakukan dengan menghitung nilai SHAP absolut rata-rata untuk setiap fitur. Nilai ini merepresentasikan kontribusi absolut rata-rata setiap fitur terhadap hasil prediksi model, sehingga fitur dengan pengaruh terbesar ditunjukkan dengan nilai SHAP yang lebih tinggi. Pendekatan ini tidak hanya memungkinkan identifikasi fitur yang paling berpengaruh, tetapi juga memberikan pemahaman kuantitatif mengenai besarnya kontribusi setiap fitur terhadap keseluruhan keluaran model.

Visualisasi nilai SHAP dibuat menggunakan pustaka SHAP Python, yang terintegrasi dengan pustaka visualisasi Matplotlib dan Seaborn. Seluruh proses analisis dan visualisasi dilakukan dalam lingkungan pengembangan Jupyter Notebook berbasis Python 3. Gambar 2 menampilkan nilai SHAP absolut rata-rata untuk setiap fitur yang digunakan dalam model pemelajaran mesin.

Visualisasi perbandingan nilai SHAP absolut rata-rata antarmodel pemelajaran mesin memberikan gambaran kuantitatif kontribusi setiap fitur terhadap hasil prediksi. Nilai tersebut mencerminkan kontribusi absolut rata-rata setiap fitur, terlepas dari arah pengaruhnya (positif atau negatif), sehingga fitur dengan nilai SHAP tinggi dapat diinterpretasikan sebagai variabel paling berpengaruh dalam keputusan model. Gambar 2 menunjukkan bahwa fitur Administrasi Sistem Komputer, Organisasi Komputer, dan Kewirausahaan memiliki nilai SHAP rata-rata yang tinggi pada beberapa model, terutama dalam *boosting gradient* dan XGBoost. Hal ini menunjukkan bahwa ketiga fitur ini secara konsisten berkontribusi signifikan terhadap proses prediksi dan dapat dianggap paling penting dalam menentukan status pekerjaan lulusan. Di sisi lain, fitur lain seperti Etika Profesional atau Pelaporan Kerja tampaknya memiliki nilai SHAP yang relatif rendah di sebagian besar model, yang mengindikasikan pengaruh yang lebih rendah. Dengan pendekatan ini, fitur-fitur penting tidak hanya dapat diidentifikasi, tetapi besarnya dampak juga dapat diukur secara kuantitatif, memberikan fondasi yang kuat untuk interpretasi model yang transparan dan dapat dijelaskan.

Dari keseluruhan fitur yang dianalisis, sebanyak 20 fitur utama diperoleh menggunakan nilai kepentingan rata-rata, yang mewakili nilai SHAP absolut rata-rata setiap fitur dan menunjukkan kontribusinya terhadap prediksi model. Fitur dengan nilai kepentingan SHAP rata-rata yang tinggi memiliki pengaruh yang lebih besar terhadap proses pengambilan keputusan model. Hasil pemeringkatan kepentingan fitur dapat dilihat pada Gambar 3. Gambar ini menunjukkan bahwa fitur-fitur dengan pengaruh paling signifikan terhadap prediksi model adalah Administrasi Sistem Komputer, Organisasi Komputer, Sistem Informasi, Kewirausahaan, Etika Profesional, Pelaporan Kerja, Pemrograman Web, Kecerdasan Buatan, IPK, dan Proyek Akhir I. Fitur-fitur ini memiliki nilai Shapley yang tinggi, yang menunjukkan bahwa perubahannya berdampak signifikan terhadap keputusan model.

Selain itu, fitur-fitur lain yang juga berdampak signifikan meliputi Pengantar Teknologi Informasi, Rekayasa Perangkat Lunak II, Pemrograman Dasar, Pancasila, Tugas Akhir II, Jaringan Komputer, Penambangan Data, Jaringan Komputer Lanjut, Durasi Magang, dan Pemrograman Perangkat Seluler. Meskipun kontribusinya lebih rendah dibandingkan fitur-fitur utama, fitur-fitur ini tetap berperan penting dalam menentukan prediksi model. Dengan demikian, hasil analisis SHAP ini memberikan wawasan tentang faktor-faktor yang paling berpengaruh dalam model, serta kontribusi setiap fitur terhadap keputusan prediksi.

V. KESIMPULAN

Penelitian ini telah mengevaluasi kinerja berbagai model pemelajaran mesin untuk klasifikasi berdasarkan akurasi, presisi, *recall*, dan *F*1-*score*. Hasilnya menunjukkan bahwa berbagai model pemelajaran mesin, termasuk *decision tree*, *random forest*, XGBoost, LightGBM, CatBoost, dan GBM, memberikan kinerja prediktif yang konsisten dan tinggi, dengan mencapai skor 92% untuk presisi, 92% untuk *recall*, dan 92% untuk *F*1-*score*, sedangkan akurasinya tetap sebesar 85%. Di sisi lain, regresi logistik menunjukkan *recall* yang lebih tinggi (100%), menghasilkan *F*1-*score* 96% dan akurasi keseluruhan 92%. Hasil ini menegaskan efektivitas regresi logistik dalam mengidentifikasi seluruh data relevan tanpa melewatkan satu pun data relevan.

Selain meningkatkan interpretabilitas model, analisis SHAP berperan penting dalam mengidentifikasi fitur-fitur yang paling berpengaruh dalam proses prediksi. Melalui analisis SHAP global, fitur-fitur diperingkat berdasarkan nilai rata-rata kepentingan SHAP. Hasilnya menunjukkan bahwa Administrasi Sistem Komputer, Organisasi Komputer, Sistem Informasi, Kewirausahaan, Etika Profesional, Pelaporan Kerja, Pemrograman Web, Kecerdasan Buatan, IPK, dan Tugas Akhir I merupakan fitur-fitur yang paling memengaruhi prediksi model. Kontributor yang rendah tetapi signifikan meliputi Pengantar Teknologi Informasi, Rekayasa Perangkat Lunak II, Pemrograman Dasar, Pancasila, Tugas Akhir II, Jaringan Komputer, Penambangan Data, Jaringan Komputer Lanjutan, Durasi Magang, dan Pemrograman Perangkat Seluler. Hasil ini memberikan pemahaman yang lebih mendalam tentang kontribusi atribut akademik dan profesional tertentu terhadap hasil prediktif.

Temuan ini menekankan pentingnya integrasi teknik XAI, seperti SHAP, ke dalam pemodelan prediktif guna menjamin transparansi dan interpretabilitas dalam pengambilan

keputusan. Penelitian di masa mendatang dapat mengeksplorasi teknik pemilihan fitur, *dataset* yang lebih besar, dan optimasi model untuk meningkatkan akurasi prediktif. Selain itu, penelitian perkembangan temporal variabel akademik dan profesional dapat memberikan wawasan yang lebih mendalam dalam memprediksi kesuksesan karier jangka panjang. Peningkatan berkelanjutan pada model pemelajaran mesin dan metode *explanation* akan sangat penting dalam meningkatkan keandalan dan penerapan analitik prediktif dalam pendidikan dan ketenagakerjaan.

KONFLIK KEPENTINGAN

Penulis menyatakan bahwa selama penelitian dan penulisan artikel ilmiah berjudul "*Interpretable Machine Learning* untuk Prediksi Penempatan Kerja: Analisis Fitur Berbasis SHAP," tidak terdapat konflik kepentingan dengan pihak mana pun.

REFERENSI

[1] World Economic Forum, "The Future of Jobs Report," 2023. [Online]. Tersedia: https://www.weforum.org/reports/the-future-of-jobs-report-2023

[2] M.H. Baffa, M.A. Miyim, dan A.S. Dauda, "Machine learning for predicting students' employability," *UMYU Sci*., vol. 2, no. 1, hal. 001–009, Mar. 2023, doi: 10.56919/usci.2123_001.

[3] A.A. binti Kahlik dan A.Y.S. Al-Hababi, "Predicting post-internship employability using ensemble machine learning approach," *J. Cogn. Sci. Hum. Dev.*, vol. 10, no. 2, hal. 87–101, Sep. 2024, doi: 10.33736/jcshd.7518.2024.

[4] S. Ramos-Pulido, N. Hernández-Gress, dan G. Torres-Delgado, "Analysis of soft skills dan job level with data science: A case for graduates of a private university," *Informatics*, vol. 10, no. 1, hal. 1–13, Mar. 2023, doi: 10.3390/informatics10010023.

[5] H. Sahlaoui dkk., "Predicting and interpreting student performance using ensemble models and Shapley additive explanations," *IEEE Access*, vol. 9, hal. 152688–152703, Okt. 2021, doi: 10.1109/ACCESS.2021.3124270.

[6] S. Ramos-Pulido, N. Hernández-Gress, dan G. Torres-Delgado, "Exploring the relationship between career satisfaction and university learning using data science models," *Informatics*, vol. 11, no. 1, hal. 1–18, Mar. 2024, doi: 10.3390/informatics11010006.

[7] Y. Aswini, J. Jersha, S.B. Chakravarthi, dan S. Aditiya B., "Predicting student placement outcomes using machine learning techniques," *Int. J. Nov. Trends Innov. (IJNTI)*, vol. 2, no. 10, hal. 63–69, Okt. 2024.

[8] M.K. Shukla dkk., "Students placement prediction model using logistic regression," dalam *Int. Conf. Innov. Adv. Technol. Eng.*, 2017, hal. 1–4.

[9] M. Kumar dkk., "Predicting college students' placements based on academic performance using machine learning approaches," *Int. J. Mod. Educ. Comput. Sci. (IJMECS)*, vol. 15, no. 6, hal. 1–13, Des. 2023, doi: 10.5815/ijmecs.2023.06.01.

[10] C. Patro dan I. Pan, "Decision tree-based classification model to predict student employability," dalam *Proc. Res. Appl. Artif. Intell.*, 2021, hal. 327–333, doi: 10.1007/978-981-16-1543-6_32.

[11] H.Q. Nguyen dkk., "Career path prediction using XGBoost model and students' academic results," *CTU J. Innov. Sustain. Dev.*, vol. 15, no. Special issue: ISDS, hal. 62–75, Okt. 2023, doi: 10.22144/ctujoisd.2023.036

[12] D. Mhamdi dkk., "Job recommendation based on recurrent neural network approach," *Procedia Comput. Sci.*, vol. 220, hal. 1039–1043, Mar. 2023, doi: 10.1016/j.procs.2023.03.145.

[13] X. Xue dkk., "Convolutional recurrent neural networks with a self-attention mechanism for personnel performance prediction," *Entropy*, vol. 21, no. 12, hal. 1–16, Des. 2019, doi: 10.3390/e21121227.

[14] M. Abdelaal, C. Hammacher, dan H. Schöning, "REIN: A comprehensive benchmark framework for data cleaning methods in ML pipelines," dalam *Proc. 26th Int. Conf. Extending Database Technol. (EDBT 2023)*, 2023, hal. 499–511.

[15] R. LaRose dan B. Coyle, "Robust data encodings for quantum classifiers," 2020, *arXiv:2003.01695*.

[16] K. Zhang dkk., "Description-enhanced label embedding contrastive learning for text classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 10, hal. 14889–14902, Okt. 2024, doi: 10.1109/TNNLS.2023.3282020.

[17] M.M. Suarez-Alvarez, D.-T. Pham, M.Y. Prostov, dan Y.I. Prostov, "Statistical approach to normalization of feature vectors and clustering of mixed datasets," dalam *Proc. R. Soc. A*, 2012, hal. 2630–2651, doi: 10.1098/rspa.2011.0704.

[18] Q.-M. Tan, "Normalization in mathematical simulations," dalam *Dimensional Analysis*, Heidelberg, Jerman: Springer, 2011, hal. 161–179.

[19] V.L. Miguéis, A. Freitas, P.J.V. Garcia, dan A. Silva, "Early segmentation of students according to their academic performance: A predictive modelling approach," *Decis. Support Syst.*, vol. 115, hal. 36-51, Nov. 2018, doi: 10.1016/j.dss.2018.09.001.

[20] L. Breiman, "Random forests," *Machine Learning*, vol. 45, hal. 5–32, Okt. 2001, doi: 10.1023/A:1010933404324.

[21] T. Chen dan C. Guestrin, "XGBoost: A scalable tree boosting system," dalam *KDD '16: Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, 2016, hal. 785–794, doi: 10.1145/2939672.2939785.

[22] G. Ke dkk., "LightGBM: A highly efficient gradient boosting decision tree," dalam *NIPS'17: Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, 2017, hal. 3149 – 3157.

[23] L. Prokhorenkova dkk., "CatBoost: Unbiased boosting with categorical features," dalam *NIPS'18: Proc. 32nd Int. Conf. Neural Inf. Process. Syst.*, 2017, hal. 6639 – 6649.

[24] A. Natekin dan A. Knoll, "Gradient boosting machines, a tutorial," *Front. Neurorobotics*, vol. 7, hal. 1–21, Des. 2013, doi: 10.3389/fnbot.2013.00021.

[25] S. Sperandei, "Understanding logistic regression analysis," *Biochem. Med.*, vol. 24, no. 1, hal. 12–18, Feb. 2014, doi: 10.11613/BM.2014.003.

[26] A.B. Parsa dkk., "Toward safer highways, application of XGBoost and SHAP for real-time accident detection and feature analysis," *Accid. Anal. Prev.*, vol. 136, hal. 1–8, Mar. 2020, doi: 10.1016/j.aap.2019.105405.

[27] K.K.P.M. Kannangara, W.-H. Zhou, Z. Ding, dan Z. Hong, "Investigation of feature contribution to shield tunneling-induced settlement using Shapley additive explanations method," *J. Rock Mech. Geotech. Eng.*, vol. 14, no. 4, hal. 1052–1063, Agu. 2022, doi: 10.1016/j.jrmge.2022.01.002.

[28] M.T. Syamkalla, S. Khomsah, dan Y.S.R. Nur, "Implementasi algoritma CatBoost dan Shapley additive explanations (SHAP) dalam memprediksi popularitas game indie pada platform Steam," *J. Teknol. Inf. Ilmu Komput.*, vol. 11, no. 4, hal. 777–786, Agu. 2024, doi: 10.25126/jtiik.1148503.

[29] H. Kamel dan M.Z. Abdullah, "Distributed denial of service attacks detection for software defined networks based on evolutionary decision tree model," *Bul. Tek. Elekt. Inform.*vol. 11, no. 4, hal. 2322–2330, Agu. 2022, doi: 10.11591/eei.v11i4.3835.

[30] M.K.M. Almansoori dan M. Telek, "Anomaly detection using combination of autoencoder and isolation forest," dalam *1st Workshop Intell. Infocommunication Netw. Syst. Serv. (WI2NS2)*, 2023, hal. 25–30, doi: 10.3311/wins2023-005.