JIKO (JURNAL INFORMATIKA DAN KOMPUTER)
Juni 2025, Volume: **9**, No. **2** | Pages 413–426
doi: 10.26798/jiko.v9i2.1886
e-ISSN : 2477-3964 – p-ISSN : 2477-4413

ARTICLE

Pemanfaatan Random Forest untuk Prediksi Ketepatan Waktu Kelulusan Mahasiswa

Studi Kasus: Institut Desain dan Bisnis Bali

Prediction of The Student Graduation Time Using Random Forest

Case Study: Bali Institute Design and Business

Gede Puspa,[1] Muhammad Febrian Rachmadhan Amri,[*,2] dan Made Prastha Nugraha[2]

[1]Sistem dan Teknologi Informasi, Institut Desain dan Bisnis Bali, Denpasar, Indonesia
[2]Sistem dan Teknologi Informasi, Institut Desain dan Bisnis Bali, Denpasar, Indonesia
*Penulis Korespondensi: febrian.rachmadhan@gmail.com

(Disubmit 13-02-25; Diterima 28-03-25; Dipublikasikan online pada 20-06-25)

Abstrak

Institut Desain dan Bisnis Bali, setiap tahun menghasilkan lulusan mahasiswa sesuai bidang yang ditempuhnya, dalam kurun waktu penyelesaian studi tepat waktu yaitu 4 (empat) tahun. Kelulusan tepat waktu mahasiswa merupakan indikator utama keberhasilan institusi pendidikan tinggi dalam menghasilkan lulusan yang kompeten. Namun, banyak faktor yang memengaruhi kelulusan mahasiswa, seperti prestasi akademik, tingkat kehadiran, keterlibatan dalam kegiatan akademik, serta faktor sosial dan ekonomi. Oleh karena itu, sistem prediksi kelulusan mahasiswa menjadi sebuah kebutuhan yang mendesak untuk membantu institusi dalam mengidentifikasi mahasiswa yang berisiko tidak lulus tepat waktu serta memberikan intervensi yang diperlukan. Dengan memanfaatkan teknologi kecerdasan buatan dan analisis data, sistem prediksi ini dapat memberikan wawasan berbasis data untuk mendukung pengambilan keputusan akademik. Sistem ini tidak hanya membantu dosen dan pihak administrasi dalam merancang strategi pembelajaran yang lebih efektif, tetapi juga memberikan manfaat bagi mahasiswa dengan menyediakan rekomendasi yang dapat meningkatkan peluang mereka untuk menyelesaikan studi tepat waktu. Penelitian ini bertujuan untuk memprediksi tingkat kelulusan mahasiswa tepat waktu dengan metode random forest untuk mengetahui metode yang lebih unggul dalam kasus tersebut. Dari penelitian ini yaitu sistem dapat memprediksi kelulusan mahasiswa tepat waktu dengan algoritma terbaik. Dengan mengetahui prediksi kelulusan mahasiswa, manajemen atau pihak kampus dapat melakukan pengambilan keputusan dan tindak lanjut terhadap mahasiswa diprediksi akan lulus tidak tepat waktu.

Kata kunci: Prediksi, Data Mining, Random Forest, Kelulusan Mahasiswa

Abstract

Bali Design and Business Institute, every year produces graduates according to their fields of study, within a timely completion period of 4 (four) years. On-time student graduation is the main indicator of the success of higher education institutions in producing competent graduates. However, many factors influence student graduation, such as academic achievement, attendance rate, involvement in academic activities, and social and economic factors. Therefore, a student graduation prediction system is an urgent need to help institutions identify students at risk of not graduating on time and provide the necessary interventions. By utilizing artificial intelligence technology and data, this prediction system can provide data-based insights to support academic decision making. This system not only helps lecturers and administrators in designing more effective learning strategies, but also benefits students by providing recommendations that can increase their chances of comple-

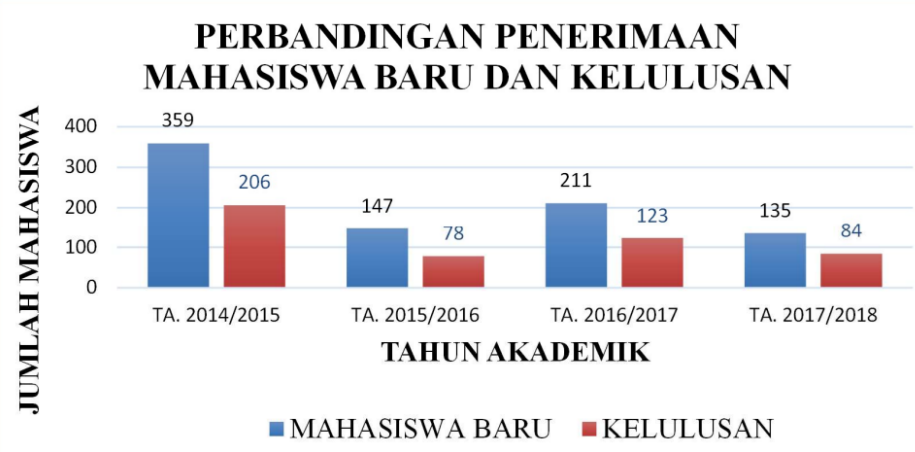
How to Cite: G. Puspa *et al.*, "Pemanfaatan Random Forest untuk Prediksi Ketepatan Waktu Kelulusan Mahasiswa Studi Kasus: Institut Desain dan Bisnis Bali", *JIKO (JURNAL INFORMATIKA DAN KOMPUTER)*, Volume: **9**, No.**2**, Pages 413–426, Juni 2025, doi: 10.26798/jiko.v9i2.1886.

ting their studies on time. This study aims to predict the rate of student graduation on time using the random forest method to find out which method is superior in this case. From this study, the system can predict student admissions on time with the best algorithm. By knowing the predicted student graduates, management or campus parties can make decisions and follow up on students who are predicted not to graduate on time.

KeyWords: Prediction, Data Mining, Random Forest, Student Graduation

1. Pendahuluan

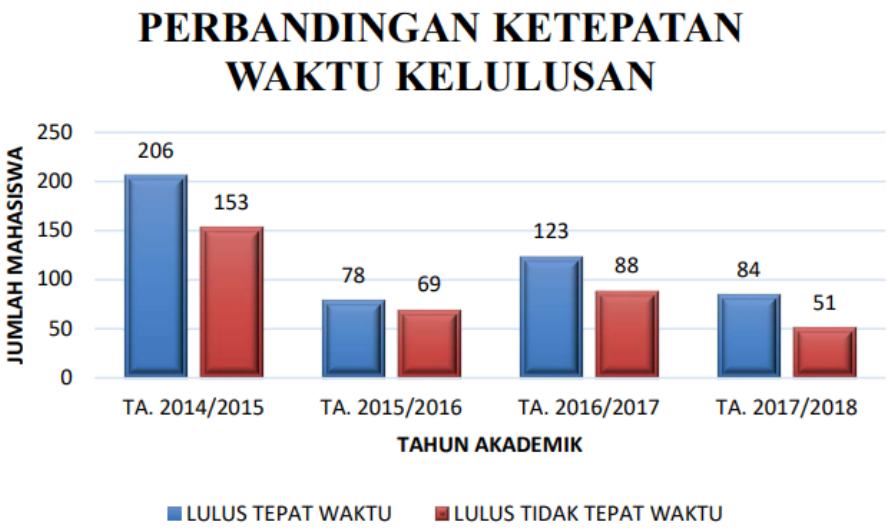
Perguruan tinggi merupakan lembaga yang menyelenggarakan proses pembelajaran, penelitian serta pengabdian kepada masyarakat atau lembaga penyelenggara Tri Dharma Perguruan Tinggi [1]. Setiap tahun perguruan tinggi menghasilkan lulusan mahasiswa sesuai bidang yang ditempuhnya, dalam kurun waktu penyelesaian studi tepat waktu yaitu 4 (empat) tahun. Penyelesaian studi mahasiswa merupakan salah satu indikator penilaian akreditasi institusi perguruan tinggi yaitu berdasarkan persentase kelulusan tepat waktu[2]. Kenyataannya tidak setiap mahasiswa dapat menyelesaikan masa belajarnya tepat waktu. Berdasarkan Peraturan Menteri Riset Teknologi dan Pendidikan Tinggi Nomor 44 Tahun 2015 tentang Standar Nasional Pendidikan Tinggi terkait dengan beban belajar dan masa belajar mahasiswa program sarjana paling lama 7 (tujuh) tahun[3]. Kondisi saat ini, jumlah pendaftaran mahasiswa baru pada perguruan tinggi Institut Desain dan Bisnis Bali tidak sebanding dengan jumlah mahasiswa yang lulus. Terdapat penurunan persentase kelulusan mahasiswa angkatan masuk 2014 hingga 2017. Pada mahasiswa angkatan masuk tahun akademik 2014/2015 terhitung hanya 57% mahasiswa yang lulus dari total 359 mahasiswa baru yang mendaftar pada tahun tersebut. Gambar 1.1 menujukkan lulus tidak tepat waktu masih cukup tinggi. Pada tahun akademik 2014/2015 mahasiswa baru berjumlah 356 sedangkan mahasiswa yang lulus berjumlah 206. Tahun akademik 2015/2016 jumlah mahasiswa baru sebanyak 147, mahasiswa yang lulus berjumlah 78. Tahun akademik 2016/2017 mahasiswa baru berjumlah 211, mahasiswa yang lulus sejumlah 123 demikian juga tahun akademik 2017/2018 mahasiswa baru sebanyak 135 dan mahasiswa yang lulus berjumlah 84.



Gambar 1. Perbandingan Penerimaan Mahasiswa dan Kelulusan

Berdasarkan data pada Gambar 1 ditelusuri lebih lanjut perbandingan jumlah mahasiswa yang lulus tepat waktu dan mahasiswa yang lulus tidak tepat waktu. Jika dibandingkan dengan jumlah data mahasiswa antara mahasiswa yang lulus tepat waktu dengan yang lulus tidak tepat waktu, komposisi jumlahnya tidak seimbang. Data menunjukkan mayoritas jumlah mahasiswa lulus tepat waktu, jumlah data yang tidak seimbang dapat mempengaruhi performa dari suatu algoritma klasifikasi[4]. Hal tersebut dapat menimbulkan terjadinya kesalahan klasifikasi terhadap dataset[5]. Sehingga perlu penanganan lebih lanjut terkait dengan data kelas tidak seimbang pada suatu dataset. Pada Gambar 2 merupakan grafik perbandingan kelulusan mahasiswa tahun akademik 2014/2015 hingga tahun akademik 2017/2018. Pada jumlah mahasiswa yang lulus tepat waktu dari tahun ke tahun rata mencapai 60% dan mengalami penurunan yaitu di tahun akademik 2014/2015 sejumlah 206 mahasiswa, tahun akademik 2015/2016 sejumlah 78 mahasiswa, tahun akademik 2016/2017 sejumlah 123 mahasiswa dan tahun akademik 2017/2018 sejumlah 84. Berbe-

da dengan jumlah mahasiswa yang lulus tidak tepat waktu rata rata per tahun mencapai 40%, di tahun akademik 2014/2015 sejumlah 153 mahasiswa, tahun akademik 2015/2016 sejumlah 69 mahasiswa, tahun akademik 2016/2017 sejumlah 88 mahasiswa dan tahun akademik 2017/2018 sejumlah 51 mahasiswa. Jika dibandingkan antara mahasiswa yang lulus tepat waktu dengan mahasiswa yang lulus tidak tepat waktu, perbandingannya sangat jauh sehingga data termasuk *imbalance class*.



Gambar 2. Perbandingan Ketepatan Waktu Lulusan

Hal ini menunjukkan terdapat permasalahan serius yang perlu ditindak lanjuti karena dapat berdampak pada nilai akreditasi perguruan tinggi. Belum diketahui penyebab pasti keterlambatan studi mahasiswa yang tidak lulus tepat waktu. Perlu adanya penggalian data yang masih tersembunyi serta pengolahan data sehingga menjadi pengetahuan dan informasi baru yang dapat dimanfaatkan untuk menindak lanjuti mahasiswa yang bermasalah pada tahun akademik berjalan[6]. Data Mining adalah proses penemuan pengetahuan dari volume data yang sangat besar[7]. Volume data yang besar akan menumpuk dan tidak berguna apabila tidak diolah dengan baik. Data Mining dapat mengekstrasi data menjadi pengetahuan yang belum diketahui, menjadi data yang bermanfaat untuk pengambilan keputusan. Data Mining banyak diimplementasikan pada berbagai bidang, seperti marketing dalam penerapan teknik Data Mining untuk Customer Relationship Management[8], pendidikan untuk memprediksi prestasi siswa berdasarkan motivasi[9], industri yaitu Data Mining dalam pengambilan keputusan customer service pada ISP[10], medis yaitu Data Mining bermanfaat salah satunya dalam memprediksi penyakit jantung[11], politik yaitu pada penelitian analisis sentimen tentang opini pilkada[12] dan lain sebagainya. Klasifikasi adalah salah satu teknik utama dalam penambangan data dan banyak digunakan di berbagai bidang[13].

Saat ini, proses pemantauan dan evaluasi terhadap mahasiswa yang berisiko mengalami keterlambatan kelulusan masih dilakukan secara manual atau berdasarkan penilaian subjektif dari dosen dan pihak akademik. Pendekatan ini memiliki keterbatasan dalam mengidentifikasi pola dan tren dari data mahasiswa secara komprehensif. Oleh karena itu, diperlukan sebuah sistem prediksi yang mampu secara akurat mengidentifikasi mahasiswa yang berpotensi tidak lulus tepat waktu, sehingga dapat dilakukan intervensi lebih awal untuk meningkatkan peluang kelulusan.

Dalam penelitian ini teknik klasifikasi digunakan untuk memprediksi tingkat kelulusan mahasiswa tepat waktu. Klasifikasi umumnya menggunakan set data training dimana semua objeknya sudah terkait dengan label kelas yang diketahui. Teknik klasifikasi akan digunakan untuk menganalisis dan mengambil sifat dari data training untuk membangun suatu model. Kemudian model yang didapat akan digunakan untuk mengklasifikasikan objek baru[14]. Terdapat berbagai metode dalam klasifikasi yaitu decision tree, artifical neural network, support vector machine, nearest neighbour rule, klasifikasi berbasis fuzzy logic, ensemble learning dan deep learning dengan berbagai teknik dalam setiap metode tersebut. Penelitian terdahulu terkait klasifikasi untuk memprediksi tingkat kelulusan mahasiswa tepat waktu dengan mem-

bandingkan metode algoritma C4.5, Naïve Bayes, k-Nearest Neighbour (kNN) dan Support Vector Machine (SVM) dengan variabel target yaitu klasifikasi lulusan yang lulus tepat waktu yaitu 4 (empat) tahun atau kurang dan memiliki nilai IPK minimal 3,00. Sedangkan variabel-variabel prediktor yaitu jenis kelamin, indeks prestasi semester 3 (tiga), 4 (empat),5 (lima) dan 6 (enam). Pengujian model menggunakan confusion matrix, hasil perbandingan terlihat bahwa algoritma Naïve Bayes memiliki nilai yang paling baik untuk semua kategori performansi dibandingkan dengan algoritma lainnya. Untuk nilai akurasi dan AUC nilai terbesar adalah yang terbaik, sedangkan untuk error adalah nilai yang terkecil. Nilai AUC untuk Naïve Bayes dan C4.5 termasuk kedalam kategori "baik", sedangkan untuk algoritma SVM dan kNN termasuk kedalam kategori "cukup". Penelitian dilakukan untuk membandingkan model prediksi untuk tingkat skor rata-rata poin akhir (IPK) siswa lulus menggunakan data dari Fakultas Pendidikan selama tahun 2010 hingga 2012 dengan menggunakan algoritma dua pohon keputusan (C4.5 dan ID3), dan teknik Naïve Bayes dan K-NN[15]. Faktor-faktor yang diusulkan untuk mempengaruhi IPK kelulusan termasuk jenis kelamin siswa, beasiswa yang diberikan, sebelumnya latar belakang pendidikan, jenis penerimaan, bakat dan provinsi SMA. Analisis mengungkapkan bahwa algoritma Naïve Bayes memberikan akurasi keseluruhan terbaik 43,18%. Ini bisa membantu memprediksi skor kelulusan siswa di masa depan dan dukungan guru untuk memberikan saran pendidikan bagi siswa mereka dan untuk mengembangkan kualitas siswa di masa depan. Penelitian mengenai perbandingan teknik klasifikasi untuk klasifikasi lama masa studi dengan tujuan utama dari penelitian ini adalah untuk menentukan faktor-faktor yang mungkin berpengaruh pada semua wisudawan/ti yang lulus pada tahun 2016 menggunakan algoritma C4.5 dan algoritma CART dan juga untuk mengetahui akurasi perbandingan hasil klasifikasi dengan algoritma C4.5 dan algoritma CART. Hasil penelitian menunjukkan bahwa faktor yang mempengaruhi durasi semua kelulusan menggunakan algoritma C4.5 adalah jurusan (X4), sekolah wilayah (X5) dan daerah asal (X3) dan faktor-faktor yang mempengaruhi durasi semua kelulusan menggunakan algoritma CART adalah utama (X4) dan Kumulatif Indeks Prestasi (X1). Klasifikasi presisi dalam algoritma CART lebih baik daripada algoritma C4.5. Algoritma C4.5 tadinya mampu memprediksi dengan akurasi 40% sedangkan algoritma CART memiliki akurasi prediksi 60%. Penelitian algoritma C4.5 dan CART dalam memprediksi kategori Indeks Prestasi Mahasiswa dengan membandingkan akurasi prediksi kategori Indeks Prestasi (IP) semester pertama mahasiswa Fakultas Teknologi Informasi (FTI) Universitas Kristen Duta Wacana (UKDW) menggunakan algoritma C4.5 dan CART. Akurasi kedua algoritma dalam memprediksi tersebut diukur dengan menggunakan tabel crosstab. Pada jalur prestasi, akurasi kedua algoritma mampu mencapai 86,86%. Pada jalur nonprestasi, akurasi algoritma C4.5 sebesar 61,54% dan algoritma CART sebesar 63,16%. Dilihat dari segi akurasinya, algoritma C4.5 dan CART lebih baik digunakan untuk memprediksi jalur prestasi daripada jalur nonprestasi. Metode Random Forest merupakan salah satu teknik machine learning yang dapat digunakan untuk membangun model prediksi kelulusan mahasiswa dengan tingkat akurasi tinggi. Metode ini mampu menangani dataset dengan variabel yang kompleks dan memberikan hasil yang interpretatif dalam mengidentifikasi faktor utama yang mempengaruhi kelulusan mahasiswa.

Berdasarkan beberapa penelitian diatas penulis mencoba membandingkan metode algoritma *Random Forest* karena metode tersebut unggul dalam hasil tingkat akurasi yang didapatkan. Penelitian dilakukan di Institut Desain dan Bisnis Bali.

2. Metode

Peneltian ini dilakukan untuk memprediksi ketepatan waktu kelulusan mahasiswa yang mana hasil prediksinya nanti dapat digunakan oleh pihak perguruan tinggi untuk menentukan keputusan ataupun kebijakan yang akan diberikan ke mahasiswa yang berpotensi akan lulus tidak tepat waktu. Metode yang digunakan penelitian ini adalah *Cross-Industry Standard Process for Data Mining* (CRISP-DM) dengan Algoritma *Random Forest* pada fase pengolahan data. Berikut ini merupakan penjelasan detail dari masing-masing fase.

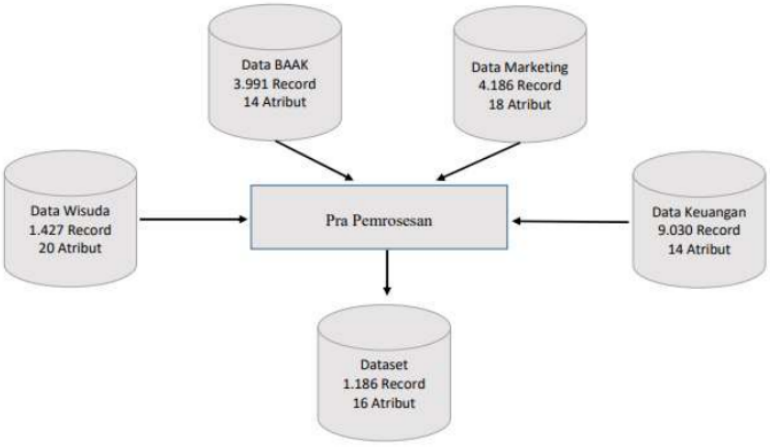
2.1 Fase 1 (*Business Understanding*)

Tahapan ini merupakan proses pemahaman terkait dengan latar belakang permasalahan yang di hadapi pada Institut Desain dan Bisnis Bali, sehingga menghasilkan pertanyaan penelitian dan tujuan dilakukan penelitian *Data Mining*. Pada tahap ini penelitian dilakukan dengan observasi, wawancara dan studi pus-

taka terkait permasalahan yang ada untuk menggali dan menganalisa langkah penyelesaian permasalahan yang akan dilakukan. Wawancara awal dilakukan dengan menggali informasi terkait kondisi permasalahan kelulusan mahasiswa. Kegiatan ini menghasilkan pertanyaan penelitian dan tujuan dilakukan penelitian *Data Mining*. Selanjutnya dilakukan studi pustaka untuk mengetahui penelitian terkait yang telah dilakukan sebelumnya. Studi pustaka yang terkumpul digunakan untuk landasan acuan sebagai referensi dalam wawancara dengan pakar yaitu ketua program studi dan penasehat akademik, untuk mengetahui kesinambungan antara atribut penelitian yang telah ada dengan permasalahan yang pada Institut Desain dan Bisnis Bali, sehingga dapat memutuskan atribut yang digunakan dalam penelitian ini. Penelitian ini tidak menggunakan atribut Indeks Prestasi karena pada Institut Desain dan Bisnis Bali menerapkan sistem paket SKS, sehingga setiap mahasiswa tetap dapat melanjutkan perkuliahan berdasarkan paket SKS yang telah ditentukan.

2.2 Fase 2 (*Data Understanding*)

Tahap ini merupakan tahapan yang dimulai dengan pengumpulan data awal, berdasarkan hasil dari observasi, wawancara dan studi pustaka yang menghasilkan keputusan pakar berupa atribut yang diasumsi sebagai penyebab permasalahan ketepatan waktu kelulusan mahasiswa. Data yang digunakan merupakan data kelulusan mahasiswa tahun masuk angkatan 2014-2017 yang didapat dari beberapa divisi pada Institut Desain dan Bisnis Bali, yaitu divisi BAAK dengan atribut awal pada data kelulusan mahasiswa yaitu nomor, NIM, prodi, nama mahasiswa, IPK, status, asal, angkatan masuk, tahun lulus serta data mahasiswa pada data BAAK dengan atribut awal yaitu nomor, NIM, nama, jenis kelamin, nomor telepon, tempat lahir, tanggal lahir, prodi, kelas, status, alamat, NIK orang tua, nama orang tua, tahun akademik. Data dari divisi marketing dengan atribut awal yaitu nomor form, NIM, nama lengkap, alamat, daerah, asal sekolah, jurusan, tahun lulus, nomor telepon, tempat lahir, tanggal lahir, prodi, kelas, sumber informasi, tanggal datang, tanggal daftar, tanggal regist dan keterangan. Serta data dari divisi keuangan dengan atribut awal yaitu NIM, nama, tanggal transaksi, jumlah SKS, keterangan cuti, mata kuliah, kelas, tahun semester, keterangan tunggakan biaya. Data mentah yang didapat tidak langsung dapat digunakan, karena data awal yang didapat merupakan data keseluruhan mahasiswa yang berisi banyak atribut seperti atribut yang telah disebutkan diatas sehingga diperlukan pra pemrosesan data dengan penyeleksian atribut dan pra pemrosesan lain agar data dapat digunakan. Komposisi data di awal terlihat pada Gambar berikut.



Gambar 3. Komposisi Data Awal

2.3 Fase 3 (*Data Preparation*)

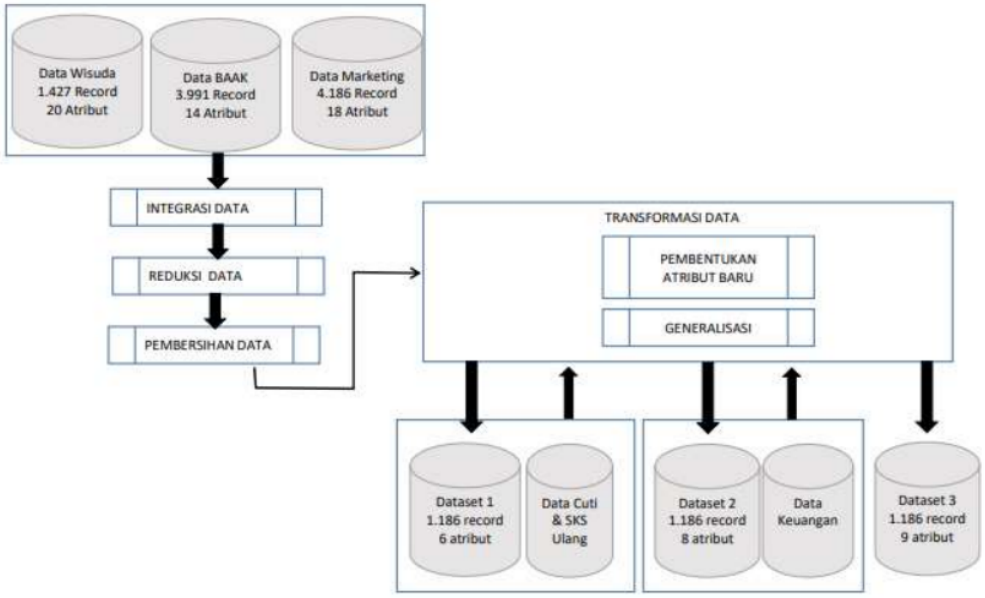
Tahap ini merupakan tahapan pengolahan data awal yang masih mentah dari data yang didapatkan, data yang didapat dalam bentuk file .xlsx. Tahap ini dilakukan secara manual menggunakan program microsoft excel. Gambar di atas merupakan tahapan pra pemrosesan yang dilakukan. Tahap pertama data awal yang menjadi rujukan utama adalah data wisudawan, dalam data wisudawan perlu dilakukan proses penghilangan atribut dan data yang tidak diperlukan. Atribut awal pada data ini adalah NIM, prodi, nama mahasiswa, IPK, SKS, tanggal lahir, tempat lahir, angkatan masuk, tahun lulus, atribut yang digunakan hanyalah atribut NIM, Nama, Program Studi dan tahun lulus mahasiswa angkatan masuk tahun 2014 hingga

tahun 2017. Selanjutnya dilakukan penambahan atribut dari data BAAK yang merupakan data mahasiswa, dengan atribut awal yaitu nomor, NIM, nama, jenis kelamin, nomor telepon, tempat lahir, tanggal lahir, prodi, kelas, status, alamat, orang tua, nama orang tuang, tahun akademik, atribut yang digunakan adalah jenis kelamin, Tanggal Lahir dan Kelas. Pada data BAAK berisi data mahasiswa dari seluruh angkatan, sehingga perlu melakukan seleksi data berdasarkan NIM dan nama mahasiswa yang sudah diseleksi sebelumnya. Dilanjutkan kembali dengan menambahkan data dari data marketing dengan atribut awal nomor form, NIM, nama lengkap, alamat, daerah, nama perusahaan, asal sekolah, jurusan, tahun lulus, nomor telepon, tempat lahir, tanggal lahir, prodi, kelas, sumber informasi, tanggal datang, tanggal daftar, tanggal regist dan keterangan kemudian dilakukan seleksi atribut, untuk atribut yang digunakan adalah asal sekolah. Pra-pemrosesan dilakukan dengan menyeleksi data berdasarkan data NIM dan nama mahasiswa. Pada tahap atribut yang sudah terkumpul adalah atribut NIM, Nama, Jenis Kelamin, Tanggal Lahir, Kelas dan Asal Sekolah. Data per tahun angkatan dibuat terpisah untuk memudahkan dalam pencarian data, tahun angkatan dibagi berdasarkan NIM yang berisi keterangan tahun angkatan pada digit ketiga dan keempat. Tahap selanjutnya mengatasi data kosong (*missing value*), karena masih terdapat atribut yang memiliki data kosong yaitu atribut tanggal lahir dan asal sekolah. Pada atribut tanggal lahir terdapat kurang lebih 25 data mahasiswa yang kosong dan atribut asal sekolah terdapat kurang lebih 84 data mahasiswa yang kosong. Untuk mengisi data tersebut, dilakukan pencarian data secara manual dengan mengecek kembali data tanggal lahir dan asal sekolah pada dataset BAAK dan dataset marketing berdasarkan NIM dan nama. Namun masih terdapat data yang kosong, maka dilakukan pencarian dalam dokumen mahasiswa yang terdapat pada hardcopy file dalam divisi BAAK. Selain itu cara lain juga ditempuh untuk mengisi data yaitu dengan menghubungi mahasiswa yang bersangkutan. Tahap selanjutnya melakukan transformasi data untuk atribut NIM yang di transformasikan menjadi tahun masuk dengan melihat digit ketiga dan keempat NIM, untuk selanjutnya dapat membentuk atribut baru yaitu lama studi sehingga dapat menentukan atribut label yaitu lulus tepat waktu dan lulus tidak tepat waktu, yang didapat dari pengurangan antara atribut tahun lulus dan tahun masuk. Jika lama studi kurang dari atau sama dengan 4 (Empat) tahun maka label yang dihasilkan lulus tepat waktu, sedangkan jika lama studi lebih dari 4 (Empat) tahun maka label yang dihasilkan adalah lulus tidak tepat waktu. Pada tahap ini diketahui bahwa data label termasuk data tidak seimbang, karena data lulus tepat waktu termasuk data mayoritas dengan perbandingan persentase 90:10. Kemudian transformasi atribut tanggal lahir menjadi usia mahasiswa saat masuk kuliah yang didapat dari mengurangi atribut tahun masuk dengan tahun lahir. Selanjutnya melakukan transformasi atribut kelas yang semula berisi keterangan kelas dan konsentrasi maka atribut dibagi menjadi dua yaitu kelas dan konsentrasi. Pada atribut asal sekolah dilakukan generalisasi dengan mengubahnya menjadi 5 (Lima) kategori yaitu SMK Negeri, SMK Swasta, SMA Negeri, SMA Swasta dan Paket C. Setelah dilakukan transformasi maka dilakukan penghapusan atribut NIM, Nama, Prodi, Kelas (awal), tahun masuk, tahun lulus dan tanggal lahir.

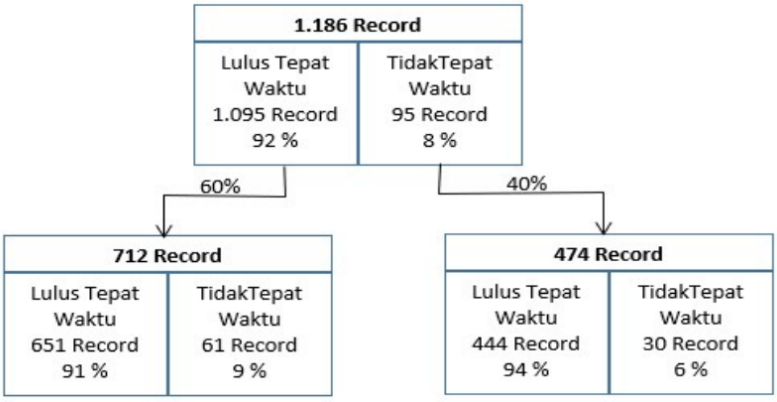
Kegiatan pra pemrosesan menghasilkan dataset yang terdiri dari 11 atribut dengan 1.186 record data. Pembagian dataset dilakukan menggunakan metode stratified random sampling. Tahap pertama data diurutkan berdasarkan label Lulus Tepat Waktu dan Lulus Tidak Tepat Waktu, kemudian dihitung persentase setiap label. Setelah itu data dibagi menjadi dua yang terdiri dari data latih 60% dan data uji 40%. Data dipilih secara acak dengan tetap memperhatikan komposisi label yang telah dihitung persentasenya.

2.4 Fase 4 (*Modeling*)

Tahap ini merupakan tahapan membuat model pembelajaran dengan metode Random Forest berdasarkan penelitian awal yang telah dilakukan sebelumnya. Metode terbaik yang digunakan pada penelitian ini adalah metode RF (*Random Forest*). Model pembelajaran dibuat dengan menggunakan aplikasi Jupyter Notebook dengan bahasa pemrograman Python. Modeling dimulai dengan melakukan *import library* yang akan digunakan, selanjutnya dilakukan pra pemrosesan kembali dengan menghapus data yang kosong. Kemudian dilanjutkan dengan melakukan transformasi data, karena mayoritas data yang dimiliki adalah data bertipe kategorikal. Pada Jupyter Notebook data kategorikal tidak dapat diproses. Data di transformasi menggunakan fungsi *One-Hot Encoding* dengan mengubah data kategorikal menjadi numerik, karena data yang ada bukan data ordinal maka menggunakan *One Hot Encoding*. Untuk data ordinal seperti atribut kemampuan finansial dan label menggunakan fungsi map untuk mengubah data menjadi numerik. Selanjutnya dilakukan *scaling* untuk normalisasi data atribut usia, jumlah SKS ulang dan jumlah cuti.dengan ra-



Gambar 4. Pra-Pemrosesan Data



Gambar 5. Pembagian Data Latih dan Data Uji

nge 0-1. Kemudian dilanjutkan dengan menerapkan *Random Forest* untuk mengatasi data tidak seimbang. Tahap berikutnya membagi data menjadi data latih dan data uji yaitu 60:40 dengan juga menerapkan metode *stratified random sampling*. Dilanjutkan dengan melakukan pencarian kernel terbaik dengan fungsi *params_grid*. Pada Jupyter Notebook, fungsi *params_grid* akan terus mencari kernel terbaik dengan juga melakukan 10 *Cross Validation*. Untuk mengetahui kernel terbaik maka dipanggil dengan fungsi *best_params_grid*. Setelah didapat kernel terbaik maka selanjutnya dilakukan evaluasi prediksi dengan data uji.

2.5 Fase 5 (*Evaluation*)

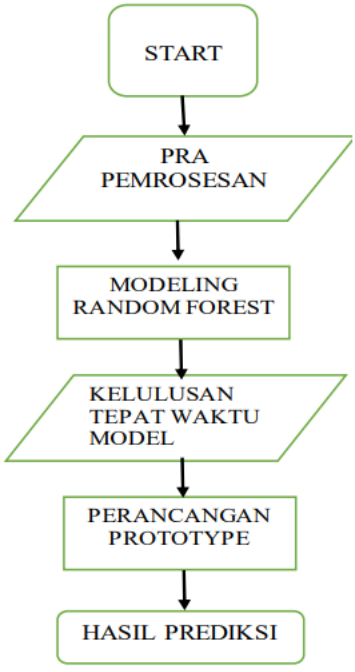
Tahap ini dilakukan untuk mengevaluasi model algoritme terbaik yang didapat dari hasil perbandingan algoritme pada tahap sebelumnya. Pada tahap ini dilakukan evaluasi model dengan menggunakan *confusion matrix* yang memperhatikan hasil akurasi, presisi dan recall serta AUC karena penelitian ini menggunakan data tidak seimbang. Jika memang tidak sesuai maka perlu dilakukan proses ulang dengan penambahan data atau pengurangan data sesuai dengan tujuan utama *Data Mining* dilakukan. Apabila model dirasa sudah sesuai dengan masalah bisnis maka dapat dilanjutkan dengan tahap selanjutnya. Evaluasi dilakukan dengan perhitungan ketepatan hasil prediksi dengan data aktual yang digunakan untuk mengetahui tingkat kesalahan yang terjadi.

Tabel 1. Perhitungan Confusion Matrix

	f_{ij}	Kelas Hasil Prediksi (j)	
		Kelas = Lulus Tepat Waktu	Kelas = Lulus Tidak Tepat Waktu
Kelas Asli	Kelas = Lulus Tepat Waktu	f11	F10
	Kelas = Lulus Tidak Tepat Waktu	f01	F00

2.6 Fase 6 (*Deployment*)

Tahap ini merupakan tahap perancangan prototipe untuk mengimplementasikan data dengan algoritme terbaik yang sudah dipilih sehingga memudahkan pengguna dalam melakukan prediksi ketepatan lulusan mahasiswa. Perancangan prototipe dalam penelitian ini diimplementasikan dalam platform Google Collab dengan input dan output library yang dimiliki oleh python, dengan bahasa pemrograman Python. Sehingga model dapat dijalankan pada prototipe yang telah dibuat.



Gambar 6. Tahapan Perancangan Prototype

2.7 Metode Pengumpulan Data

Pengumpulan data dimulai dengan melakukan observasi langsung berdasarkan informasi yang didapatkan bahwa jumlah mahasiswa yang lulus lebih sedikit dibanding jumlah mahasiswa yang mendaftar di awal tahun akademik. Selanjutnya dilakukan wawancara dengan pakar yaitu ketua program studi dan pembimbing akademik terkait dengan permasalahan ketepatan kelulusan mahasiswa, serta atribut yang mempengaruhi permasalahan tersebut yang juga didasarkan pada atribut-atribut penelitian yang telah dilakukan sebelumnya. Berdasarkan atribut yang disarankan oleh pakar, data di kumpulkan dari berbagai sumber divisi diantaranya divisi BAAK untuk data mahasiswa yang telah lulus, divisi marketing untuk mendapatkan data mahasiswa yang mendaftar dan divisi keuangan untuk mendapatkan data jumlah cuti dan SKS mengulang penelitian dilakukan di Institut Desain dan Bisnis Bali. Data yang digunakan merupakan data mahasiswa angkatan masuk 2014 hingga tahun 2017 jenjang Strata 1 Program Studi Desain Komunikasi Visual dan Desain Interior, program kelas reguler dan non regular. Data mahasiswa yang telah lulus, divisi marketing untuk mendapatkan data mahasiswa yang mendaftar dan divisi keuangan untuk mendapatkan data jumlah cuti dan SKS mengulang penelitian dilakukan di Institut Desain dan Bisnis Bali. Data yang digunakan merupakan data mahasiswa angkatan masuk 2014 hingga tahun 2017 jenjang Strata 1 Program Studi Desain Komunikasi Visual dan Desain Interior, program kelas reguler dan non regular.

2.7.1 Pembersihan Data

Menyatukan seluruh data dilakukan dengan cara manual menggunakan aplikasi microsoft excel. Data awal yang menjadi patokan adalah daftar data mahasiswa yang telah lulus dari angkatan tahun masuk 2014 hingga tahun 2017, namun data tersebut belum memiliki atribut lengkap sesuai kebutuhan. Atribut hanya terdiri dari NIM dan tahun lulus. Sehingga perlu data lain untuk melengkapinya, maka dilakukanlah penyeleksian data mahasiswa. Data mahasiswa merupakan data yang di ekspor dari aplikasi sistem informasi Akademik (BAA), yang berisi biodata mahasiswa dari seluruh angkatan. Sehingga data tidak dapat langsung digunakan dan perlu penyeleksian. Dari data mahasiswa didapat atribut jenis kelamin, tanggal lahir dan kelas yang akan membentuk dua atribut jenis kelas dan konsentrasi. Dikarenakan atribut yang diperlukan belum lengkap maka, dilakukan penyeleksian data kembali yaitu data yang didapat dari divisi marketing. Data dari divisi marketing merupakan data aplikan yang mendaftar dari tahun masuk 2014 hingga 2017, atribut yang didapat dari data ini adalah asal sekolah. Atribut yang diperlukan masih belum lengkap sehingga perlu dilakukan peryortiran data kembali yaitu data yang didapat dari divisi keuangan. Namun data dari divisi keuangan tidak langsung didapatkan, sehingga perlu waktu tunggu. Untuk mengisi waktu tunggu, penelitian tetap dilanjutkan dengan melakukan uji coba data yang ada.

NIM	JK	TANGGAL LAHIR	KELAS	ASAL SEKOLAH	TAHUN LULUS
1410010002002	P	2-Jan-96	REGULAR	SMAN 1 BANGLI	2018
1410010003004	L	24-Jun-96	REGULAR	SMAN 1 NEGARA	2018
1410010003011	L	23-Jul-94	REGULAR	SMK Harapan	2018
1410010003012	L	12-Jan-97	REGULAR	SMKN 1 KLUNGKUNG	2018
1410010003015	L	10-Jan-91	NON REGULAR	SMAK STELLA MARIS	2018
1410010003016	L	25-Sep-94	REGULAR	SMA Katolik Blitar	2018
1410010003018	L	25-Jan-96	REGULAR	SMA Niti Mandala Club	2018
1410010003019	L	23-Jun-93	REGULAR	SMKN 1 Ngelegok	2018
1410010003020	L	1-Jul-94	NON REGULAR	SMK Katolik Blitar	2018
1410010003021	L	15-Sep-96	REGULAR	SMAN 1 SINGARAJA	2018
1410010003022	L	19-Aug-95	NON REGULAR	SMAN 4 SINGARAJA	2018
1410010003024	P	17-Sep-96	REGULAR	SMK SARASWATI 2 DPS	2018
1410010003025	P	17-Jun-91	REGULAR	SMK Muhammadiyah 2 SMG	2018
1410010003026	L	13-Sep-96	REGULAR	SMAN 8 DENPASAR	2018
1410010003029	L	20-Feb-95	NON REGULAR	SMAN 1 Pasaman	2018
1410010004001	L	30-Oct-97	REGULAR	SMA PGRI ROGO JAMPI	2018
1410010004002	L	5-Feb-83	NON REGULAR	SMAN 7 DPS	2018
1420020001002	P	5-Jun-91	NON REGULAR	SMKN 2 DPS	2018
1420020001004	P	19-Jan-96	NON REGULAR	SMKN 1 Tegalalang	2018

Gambar 7. Contoh Dataset Pra Pemrosesan Awal

2.7.2 Transformasi Data

1. Pembentukan atribut baru
   Terdapat 5 (Lima) atribut yang di transformasi untuk membentuk atribut baru yaitu atribut NIM yang di transformasikan menjadi tahun masuk untuk selanjutnya dapat membentuk atribut baru yaitu lama studi sehingga dapat menentukan atribut label yaitu lulus tepat waktu dan lulus tidak tepat waktu, yang didapat dari pengurangan antara atribut tahun lulus dan tahun masuk. Jika lama studi kurang dari atau sama dengan 4 (Empat) tahun maka label yang dihasilkan lulus tepat waktu, sedangkan jika lama studi lebih dari 4 (Empat) tahun maka label yang dihasilkan adalah lulus tidak tepat waktu. Kemudian transformasi atribut tanggal lahir menjadi usia mahasiswa saat masuk kuliah yang didapat dari mengurangi atribut tahun masuk dengan tahun lahir. Selanjutnya melakukan transformasi atribut kelas menjadi atribut program kelas dan konsentrasi. Pada tahapan ini dilakukan pembentukan atribut baru, yaitu atribut Masa Studi dan Label. Atribut Masa Studi diperoleh dengan cara menghitung Tahun Lulus dikurangi dengan Tahun Masuk. Kemudian atribut Label diperoleh dengan cara melihat lama masa studi. Jika masa studi <= 4 tahun, maka Label berisi Tepat Waktu, selain itu Label berisi Tidak Tepat Waktu. Hasil dari tahapan ini adalah dataset dengan 17 atribut dan 1186 record.
2. *Data Reduction*
   Pada tahapan ini dilakukan dimensionality reduction, yaitu dengan cara: a. Menghapus atribut NIM, Nama, Tanggal Lahir karena menurut Manajemen tidak berpengaruh pada saat penentuan lulus tepat waktu atau tidak. b. Menghapus atribut Jumlah SKS Lulus karena atribut bernilai tunggal yaitu 148 SKS c. Menghapus atribut Tahun Masuk dan Tahun Lulus karena sudah dibentuk atribut Masa Studi Sehingga Dataset berjumlah 11 atribut dan 1186 record.
3. Generalisasi
   Perubahan nilai atribut dilakukan pada atribut Prodi dan Asal Sekolah. Nilai Atribut Prodi diubah dari 90241 menjadi Desain Interior dan nilai 90421 menjadi Desain Komunikasi Visual. Kemudian Asal sekolah digeneralisasi menjadi SMA Negeri, SMA Swasta, SMK Negeri, SMK Swasta, MAN, dan Paket C.

Transformasi dilakukan dengan mengubah data sesuai kebutuhan dan mengelompokkannya dengan kategori baru. Berikut ini merupakan contoh hasil dari proses transformasi data.

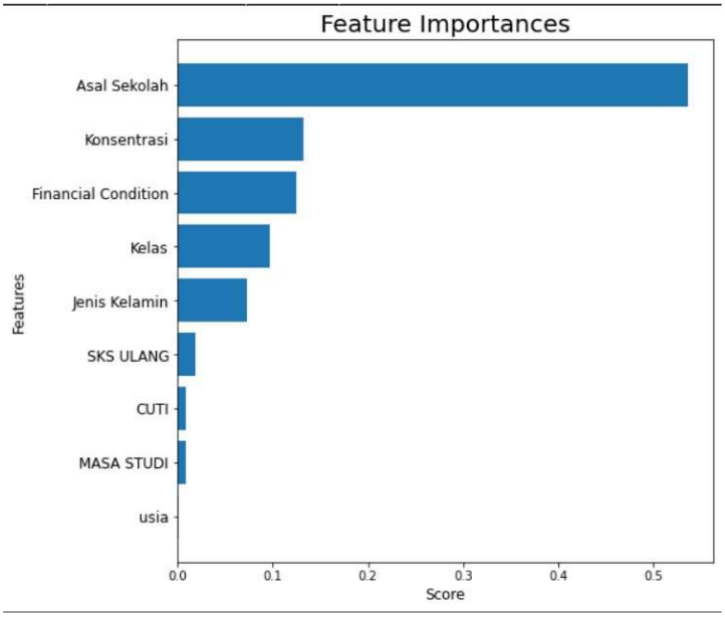
kd_pro	JK	usia	KELAS	ASAL SEKOLAH	MASA STUDI	CUTI	SKS ULANG	FINANCIAL	KONSENTRASI	LABEL
DKV	P	25	REGULAR	SMA Negeri	4	0	0	Sangat Baik	Interior Produk dan Eksibisi	Tepat Waktu
DKV	L	25	REGULAR	SMA Negeri	4	0	0	Sangat Baik	Interior Bisnis	Tepat Waktu
DKV	L	27	REGULAR	SMK Swasta	4	0	0	Sangat Baik	Binis Multi Media	Tepat Waktu
DKV	L	24	REGULAR	SMK Negeri	4	0	0	Sangat Baik	Periklanan	Tepat Waktu
DKV	L	30	NON REGULAR	SMA Swasta	4	0	0	Sangat Baik	Interior Produk dan Eksibisi	Tepat Waktu
DKV	L	27	REGULAR	SMA Swasta	4	0	0	Sangat Baik	Periklanan	Tepat Waktu
DKV	L	25	REGULAR	SMA Swasta	4	0	0	Sangat Baik	Binis Multi Media	Tepat Waktu
DKV	L	28	REGULAR	SMK Negeri	4	0	0	Sangat Baik	Periklanan	Tepat Waktu
DKV	L	27	NON REGULAR	SMK Swasta	4	0	0	Sangat Baik	Binis Multi Media	Tepat Waktu

Gambar 8. Contoh Data Hasil Transformasi

Dilanjutkan dengan menampilkan feature importance dari dataset setelah dilakukan transformasi data yang menampilkan bahwa asal sekolah, konsentrasi dan finansial merupakan 3 atribut terpenting yang digunakan dalam proses pembelajaran algoritma yang digunakan.

3. Hasil dan Pembahasan

Berdasarkan hasil prediksi menggunakan algoritma Random Forest, tabel 4.5 merupakan hasil confusion matrix memperoleh hasil akurasi, presisi, recall, AUC dan g-mean score. Maka selanjutnya dilakukan pengujian dengan confusion matrix untuk membuktikan ketepatan hasil yang didapatkan.



Gambar 9. *Feature Importance*

Tabel 2. Hasil Confusion Matrix

		Diprediksi Sebagai	
		Tidak Tepat Waktu	Tepat Waktu
Label	Tidak Tepat Waktu	36	0
	Tepat Waktu	0	439

Perhitungan hasil secara manual menggunakan *confusion matrix*:

$$Akurasi = \frac{439 + 36}{439 + 36 + 0 + 0} \times 100\% = 1,000$$

$$Presisi = \frac{439}{0 + 439} \times 100\% = 1$$

$$TP_{rate} = Recall = \frac{439}{0 + 439} \times 100\% = 1$$

$$TN_{rate} = Specify = \frac{36}{36 + 0} = 1$$

$$FP_{rate} = \frac{0}{0 + 3} = 0$$

$$AUC = \frac{1 + 1,000 - 0}{2} = 1,000$$

$$G - mean\ score = \sqrt{1,000 \times 1,000} = 1,000$$

Dalam penerapannya, penelitian ini diimplementasikan dalam platform Google Collab dengan input dan output library yang dimiliki oleh Python. Berikut adalah tampilan layar dari aplikasi yang dibuat:

3.1 Tampilan Input

Tampilan layar input yang digunakan untuk melakukan pengisian data yang akan diprediksi. Tampilan ini berisi beberapa field inputan, antara lain : Usia, masa studi, cuti, sks ulang, jenis 59 kelamin, kategori kelas, kondisi finansial, konsentrasi pendidikan, asal sekolah. Pada tampilan ini, user akan melakukan input data mahasiswa yang akan diprediksi.

3.2 Tampilan Output

Tampilan layar Output setelah sistem melakukan prediksi. Pada tampilan ini akan mengeluarkan hasil prediksi sistem yaitu Tepat Waktu atau Tidak tepat Waktu.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1186 entries, 0 to 1185
Data columns (total 11 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   kd_pro        1186 non-null   object
 1   JK            1186 non-null   object
 2   usia          1186 non-null   int64
 3   KELAS         1186 non-null   object
 4   ASAL SEKOLAH  1186 non-null   object
 5   MASA STUDI    1186 non-null   int64
 6   CUTI          1186 non-null   int64
 7   SKS ULANG     1186 non-null   int64
 8   FINANCIAL     1186 non-null   object
 9   KONSENTRASI   1186 non-null   object
 10  LABEL         1186 non-null   object
dtypes: int64(4), object(7)
memory usage: 102.0+ KB
[[439   0]
 [  0  36]]
Accuracy: 1.0
Masukan Usia :
```

Gambar 10. Tampilan Input

```
Masukan Konsentrasi Pendidikan :
 0. Animasi
 1. Bisnis Multi Media
 2. Interior Bisnis
 3. Interior Graphic dan Animasi
 4. Interior Proudk dan Eksibisi
 5. Periklanan
 6. Vidio Grafi dan Photo Grafi
0
Masukan Asal Sekolah :
 0. MAN
 1. Paket C
 2. SMA Negeri
 3. SMA Swasta
 4. SMK Negeri
 5. SMK Swasta
2
Hasil Prediksi Adalah :  ['Tepat Waktu']
```

Gambar 11. *Feature Importance*

4. Simpulan

Penelitian ini menggunakan pengujian dan evaluasi dari aplikasi yang dibuat menggunakan dataset dan algoritma yang diusulkan, maka dapat disimpulkan bahwa: Metode *Random Forest* dan *Naive Bayes* mengungguli klasifikasi prediksi lulus tepat waktu dibandingkan algoritma lainnya dengan performa akurasi 100%, KNN 97%, SVM 93% dan *Decision Tree* 99%. Performa akurasi *Random Forest* dan *Decision Tree* berada pada 99% namun indikator Presisi dan Recall algoritma Random Forest mengungguli KNN dengan performa Presisi 100% dan Recall 100% sementara KNN dengan performa Presisi 90% dan Recall 77%.

Perancangan model dengan algoritma *Random Forest* berhasil dilakukan dan mempunyai 4 versi berdasarkan teknik pembagian data. Pertama, model Random Forest dengan pembagian data menggunakan corss validation 10-fold menghasilkan performa akurasi 98%. Kedua, model Random Forest dengan pembagian data 60:40 dengan performa akurasi 100%. Ketiga, model Random Forest dengan pembagian data 70:30 dengan performa akurasi tertinggi 100% Kempat, model *Random Forest* dengan pembagian data 80:20 dengan performa akuransi tertinggi 100%.

Algoritma Random Forest dalam melakukan prediksi Lulus Tetap Waktu pada data mahasiswa wisudawan yang berjumlah 1.186 data dapat menangani imbalance dataset dengan menghasilkan performa akurasi 100%, *Area Under Curve* (AUC) 100%, Precision 100% dan Recall 100%. AUC sebesar 100% menunjukkan bahwa model mampu membedakan Label Lulus Tetap Waktu pada dataset wisudawan mahasiswa Institut Desain dan Bisnis Bali (IDB Bali).

Penelitian ini dapat dikembangkan dengan mengimplementasikan pada sistem informasi akademik pada Institut Desain dan Bisnis Bali, sehingga dapat langsung diakses oleh ketua program studi dan penasehat akademik. Untuk penelitian mendatang dapat mencoba menggunakan atribut lain serta algoritme lain serta dapat juga dengan menggunakan metode ensemble seperti bagging, boosting dan stacking yang selanjutnya dapat dibandingkan dengan penelitian ini serta pengembangan platform aplikasi berbasis web atau mobile, sehingga diperoleh pengetahuan baru yang lebih baik.

Pustaka

[1] M. A. D. S. Rahmi, *Manajemen Perguruan Tinggi*. Prenada Media, 2024, [Daring]. Tersedia pada: https://books.google.co.id/books?id=alY0EQAAQBAJ.

[2] M. M. P. D. H. A. Rusdiana and M. P. D. Nasihudin, *KESIAPAN MANAJEMEN AKREDITASI INSTITUSI PERGURUAN TINGGI: (Studi di PTKIS Wilayah II Jawa Barat dan Banten)*. Pusat Penelitian dan Penerbitan UIN SGD Bandung, 2021, [Daring]. Tersedia pada: https://books.google.co.id/books?id=dExVEAAAQBAJ.

[3] S. Januar, *Mutu Pendidikan: Implementasi Sistem Penjaminan Mutu Internal (SPMI) di Sekolah Binaan*. Gunawana Lestari, [Daring]. Tersedia pada: https://books.google.co.id/books?id=HsldEAAAQBAJ.

[4] Y. Sun, M. S. Kamel, A. K. C. Wong, and Y. Wang, "Cost-sensitive boosting for classification of imbalanced data," *Pattern Recognition*, vol. 40, no. 12, pp. 3358–3378, 2007.

[5] B. Machado, T. Rodrigues, Z. Lopes, R. Lopes, and M. Mesquita, "Paraparesis: A rare presentation of thrombosis of the abdominal aorta," *European Journal of Internal Medicine*, vol. 24, p. e256, Oct 2013.

[6] S. R. I. Made, H. Manurung, M. F. R. Amri, G. S. Mahendra, and I. P. I. Yoga, "Student perceptions of the implementation of big data in sinergy as learning optimization at the bali institute of design and business," *International Journal of Engineering, Science & Information Technology (IJESTY)*, vol. 4, no. 1, pp. 44–47, 2024.

[7] M. H. B. Roslan and C. J. Chen, "Educational data mining for student performance prediction: A systematic literature review (2015–2021)," *International Journal of Emerging Technologies in Learning (iJET)*, vol. 17, no. 5, pp. 147–179, Mar 2022.

[8] L. Zahrotun, "Implementation of data mining technique for customer relationship management (crm) on online shop tokodiapers.com with fuzzy c-means clustering," in *Proceedings of the International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*, 2017.

[9] J. N. Purwaningsih and Y. Suwarno, "Predicting students achievement based on motivation in vocational school using data mining approach," in *2016 4th International Conference on Information and Communication Technology (ICoICT)*, 2016, pp. 1–5, [Daring]. Tersedia pada: https://api.semanticscholar.org/CorpusID:16409601.

[10] M. F. R. Amri, M. H. Umam, A. Wibowo, and I. M. S. Ramayu, "Internet service provider user customer lifetime segmentation analysis using rfm and k-means algorithm," *Sinkron: Jurnal dan Penelitian Teknik Informatika*, vol. 8, no. 1, pp. 306–316, Jan 2024.

[11] S. R. Fernanda and P. D. Mardika, "Implemetasi data mining untuk mendiagnosa penyakit jantung dengan algoritma naïve bayes," in *Seminar Nasional Riset dan Inovasi Teknologi (SEMNAS RISTEK)*, 2025, pp. 114–118.

[12] A. Rossi, T. Lestari, R. S. Perdana, and M. A. Fauzi, "Analisis sentimen tentang opini pilkada dki 2017 pada dokumen twitter berbahasa indonesia menggunakan naïve bayes dan pembobotan emoji," vol. 1, no. 12, pp. 1718–1724, 2017, [Daring]. Tersedia pada: http://j-ptiik.ub.ac.id.

[13] K. Parmar, D. Vaghela, and P. Sharma, "Performance prediction of students using distributed data mining," in *2015 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)*, 2015, pp. 1–5.

[14] C. Chotimah, "Penerapan metode mutual information dan bayes network untuk klasifikasi penyelesaian studi," *MJRICT: Musamus Journal Of Research Information and Communication Technology*, vol. 2, no. 1, pp. 26–34, 2019, [Daring]. Tersedia pada: https://ejournal.unmus.ac.id/index.php/mjrict.

[15] A. Desiani, S. Yahdin, and D. Rodiah, "Prediksi tingkat indeks prestasi kumulatif akademik mahasiswa dengan menggunakan teknik data mining," *Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK)*, vol. 7, no. 6, pp. 1237–1244, 2020.