**LAPORAN PENELITIAN MANDIRI**



**Penerapan Teknik Random Oversampling Untuk Memprediksi Ketepatan Waktu Lulus Menggunakan Algoritma Random Forest**

**PENELITI**

**Sri Diantika, M.Kom (0310119501)**

**Hiya Nalatissifa, M.Kom (0307059701)**

**Riki Supriyadi, M.Kom (0326079601)**

**Nurlaelatul Maulidah, M.Kom (0303089601)**

**Ahmad Fauzi, M.Kom (0326119401)**

**FAKULTAS TEKNIK INFORMATIKA**

**UNIVERSITAS BINA SARANA INFORMATIKA**

**FEBRUARI 2024**

**HALAMAN PENGESAHAN**

**Judul Penelitian** : Penerapan Teknik *Random Oversampling* Untuk Memprediksi Ketepatan Waktu Lulus Menggunakan Algoritma *Random Forest*

**Peneliti**

Nama Lengkap : Sri Diantika, M.Kom

NIDN : 0310119501

Jabatan Fungsional : Asisten Ahli

Program Studi : Sistem Informasi (S1)

Nomor HP : 0857-0071-8802

Alamat surel (e-mail) : sri.szd@bsi.ac.id

**Anggota Peneliti (1)**

Nama Lengkap : Hiya Nalatissifa, M.Kom

NIDN : 0307059701

Jabatan Fungsional : Asisten Ahli

Program Studi : Sistem Informasi Akuntansi Kampus Tegal (D3)

**Anggota Peneliti (2)**

Nama Lengkap : Riki Supriyadi, M.Kom

NIDN : 202103282

Jabatan Fungsional : Asisten Ahli

Program Studi : Sains Data

**Anggota Peneliti (3)**

Nama Lengkap : Nurlaelatul Maulidah M.Kom

NIDN : 0303089601

Jabatan Fungsional : Asisten Ahli

Program Studi : Sistem Informasi Akuntansi Kampus Tegal (D3)

**Anggota Peneliti (4)**

Nama Lengkap : Ahmad Fauzi, M.Kom

NIDN : 0326119401

Jabatan Fungsional : Lektor

Program Studi : Sistem Informasi

Biaya yang diusulkan : Rp 4.000.000;

 Jakarta, 5 Februari 2024

Menyetujui, Peneliti

Ketua LPPM

*Cap LPPM dan tanda tangan*

(Dr. Taufik Baidawi, M.Kom) (Sri Diantika, M.Kom)

NIP. 200304891 NIP. 202111296

Mengetahui,

Rektor

*Cap UBSI dan tanda tangan*

 (Prof. Dr. Ir. Mochamad Wahyudi, M.Kom, MM, M.Pd, IPU, ASEAN Eng)

 NIP. 199810339

**DAFTAR ISI**

Halaman

Halaman Sampul i

Halaman Pengesahan ii

Daftar Isi iii

Ringkasan iv

**BAB I PENDAHULUAN** 1

1.1. Latar Belakang 1

1.2. Rumusan Masalah 2

1.3. Ruang Lingkup 3

1.4. Tujuan Penelitian 3

**BAB II TINJAUAN PUSTAKA** 4

2.1. Tinjauan Pustaka 4

 1. Data Mining 4

 2. Prediksi 5

 3. *Random Forest* 5

 4. *Split Validation* 5

 5. *Confussion Matrix* 6

**BAB III TUJUAN DAN MANFAAT PENELITIAN**  8

3.1. Tujuan Penelitian 8

3.2. Manfaat Penelitian 8

**BAB IV METODE PENELITIAN**  9

4.1. *Dataset* 9

4.2. *Preprocessing* 10

4.3. *Split Validation* 11

4.4. *Modelling* 12

4.5. *Evaluation* 13

**BAB V HASIL DAN PEMBAHASAN**  14

5.1. Hasil dan Pembahasan 14

**BAB VI KESIMPULAN DAN SARAN**  17

**DAFTAR PUSTAKA** 18

**LAMPIRAN** 19

Lampiran 1. Justifikasi Anggaran Penelitian 19

Lampiran 2. Biodata Peneliti 21

**RINGKASAN**

Ketepatan waktu lulus merupakan suatu hal yang didambakan mahasiswa, Selain penting bagi mahasiswa, ketepatan waktu lulus ini juga sangat penting bagi perguruan tinggi, hal ini dikarenakan aspek kelulusan mahasiswa menjadi salah satu aspek penilaian dalam sebuah proses akreditasi institusi suatu perguruan tinggi untuk menunjukan kualitasnya. Salah satu kendala yang dihadapi untuk mengetahui apakah seorang mahasiswa dapat lulus tepat waktu atau tidak adalah masa studi yang belum dapat dideteksi secara dini, hal ini akan berdampak pada terlambatnya kelulusan mahasiswa. Untuk menanggulangi tantangan tersebut, penelitian ini bertujuan untuk mengidentifikasi ketepatan waktu kelulusan mahasiswa sejak dini dengan memanfaatkan nilai Indeks Prestasi Kumulatif (IPK) yang dicapai mahasiswa selama masa study mereka. Penelitian perihal prediksi ketepatan waktu lulus mahasiswa ini menggunakan model random forest. Algoritma random forest menggabungkan beberapa algoritma dengan tipe yang sama, seperti multiple decision tree. Data yang digunakan sebagai objek riset ini, memiliki jumlah kelas data yang tidak seimbang, untuk mengatasi hal ini maka diterapkan teknik resampling random oversampling (ROS) dan juga menetapkan Split validation dimana akan dilakukan pembagian antara data pembelajaran sebesar 50% untuk data pengujian 50%. Untuk mengkaji model yang dibentuk, Penulis memakai metrics evaluation semacam akurasi, precision, serta recall. Hasil dari riset menampilkan jikalau model yang diusulkan bisa dengan baik melaksanakan prediksi dibanding dengan model lainnya, yaitu dengan hasil nilai precision sebesar 87,05%, uji akurasi sebesar 90,04%, recall 90,04%.Dari hasil ini dapat diartikan bahwa algoritma random forest dinilai baik dalam memprediksi ketepatan waktu lulus seorang mahasiswa

**BAB I**

**PENDAHULUAN**

* 1. **Latar Belakang**

Pendidikan saat ini menjadi aspek yang dibutuhkan dalam suatu kehidupan. Dalam dunia pendidikan akan banyak ditemui ilmu-ilmu pengetahuan yang nantinya akan bermanfaat untuk masa depan kita. Dalam proses menempuh pendidikan akan banyak ditemukan kendala, dari kendala-kendala ini diharapkan pelajar atau mahasiswa dapat menerapkan ilmu yang didapatkannya dan menemukan solusi untuk menyelesaikannya. Selain ilmu pengetahuan, hal lain yang didambakan pelajar atau mahasiswa adalah dapat lulus dengan tepat waktu (1)

Ketepatan waktu lulus ini merupakan impian dari mahasiswa, sebagai wujud atau bentuk dari perjuangan mereka selama beberapa tahun belajar di bangku kuliah. Selain ketepatan waktu kuliah ini penting bagi mahasiswa, penting juga bagi perguruan tinggi, hal ini disebabkan karena aspek kelulusan mahasiswa merupakan salah satu aspek penilaian dalam sebuah proses akreditasi institusi suatu perguruan tinggi untuk menunjukan kualitasnya (2) Kualitas dari perguruan tinggi ini dapat dilihat dari banyak aspek, selain dilihat dari rata-rata lama waktu alumninya mendapatkan pekerjaan, aspek penilaian lainnya yang tak kalah penting adalah presentase ketepatan waktu lulus dari mahasiswanya (3)

Jumlah besar persentase mahasiswa yang bisa menuntaskan masa *study* tepat waktu akan membuat tingkat evaluasi akreditasi naik. Pada lain sisi, mahasiswa yang bisa menuntaskan masa *study* tepat waktu tidak akan terbebani dengan Uang Kuliah Tunggal ataupun UKT yang wajib senantiasa dibayarkan sepanjang mahasiswa belum dinyatakan lulus tes skripsi. Tetapi masa *study* mahasiswa belum bisa dideteksi secara dini, sehingga bisa berakibat pada terlambatnya kelulusan mahasiswa (4) Untuk menaganalisa hal tersebut, maka banyak dilakukan penelitian mengenai ketepatan kelulusan mahasiswa, salah satunya pada penelitian Klasifikasi Ketepatan Waktu Kelulusan Mahasiswa yang berstudi kasus di suatu Universitas, membandingkan antara 2 metode yaitu metode *Naïve Bayes* dan *K-Nearest Neighbor.* Hasilnya menunjukan bahwa menggunakan Algoritma *K-Nearest Neighbor* dapat lebih baik mengklasifikasikan ketepatan waktu lulus mahasiswa dibuktikan dengan perolehan akurasi sebesar 97,68% dengan tingkat *recall* sebesar 86,11% dan presisi 100% (5) adapula penelitian dengan fokus yang sama yaitu mengenai ketepatan waktu lulus mahasiswa pada program studi Teknik Elektro dan Teknik Sipil. Penelitian ini memilsiki 3 *cluster* atau kelas data yaitu lulus tidak tepat waktu, tepat waktu dan melebihi masa *study*. Hasil penelitian menunjukan bahwa algoritma K-Means dan Fuzzy C- Means dapat mengidentifikasi ketepatan lulusan dengan baik (6)

Fiqih Satria sebelumnya melaksanakan riset serupa, di mana ia memproyeksikan ketepatan kelulusan mahasiswa di Fakultas Dakwah dan Ilmu Komunikasi UIN Raden Intan Lampung dengan memanfaatkan algoritma C4.5. Temuan dari penelitian tersebut menunjukkan bahwa algoritma C4.5 dapat digunakan untuk memprediksi waktu kelulusan mahasiswa, dengan tingkat *recall* sebesar 60.4%, presisi 70.70%, dan akurasi 58.2% (4) kemudian pada penelitian lain yang juga ingin memprediksi ketepatan waktu lulus dengan membandingkan 2 algoritma, yaitu algoritma C4.5 dan Id3, Temuan dari riset ini menunjukkan bahwa informasi yang diperoleh menunjukkan bahwa nilai akurasi yang dicapai oleh algoritma C4.5 cenderung lebih tinggi secara keseluruhan jika dibandingkan dengan nilai akurasi algoritma ID3 (7) selain itu ada pada penelitian lain menggunakan Algoritma Naïve Bayes menyatakan bahwa ia mampu menciptakan model prediksi untuk ketepatan waktu lulus mahasiswa, dengan tingkat akurasi mencapai 93,75%, presisi sebesar 93,75%, recall 100%, dan nilai AUC sebesar 1,000 (*excellent classification*) sehingga dapat dikatakan algoritma ini mampu memprediksi secara akurat (8) dilain penelitian dengan tema yang sama yaitu untuk memprediksi ketepatan waktu lulus menggunakan metode *Decision Tree* C4.5. Penelitian ini menggunakan SMOTE (*Synthetic Minority Oversampling Technique*) pada aplikasi WEKA untuk menyeimbangkan data di kelas minor, Hasil penelitian menunjukkan bahwa atribut IPK, merupakan akar tertinggi yang memberikan pengaruh tertinggi. Akurasi pada validasi silang 10 Lipatan dengan SMOTE dan tanpa SMOTE hasilnya menunjukkan bahwa: dengan SMOTE 83,055% dan tanpa SMOTE 82,644% (9)

Dari beberapa studi kasus yang sudah ada, maka akan dilakukan penelitian dengan tujuan yang sama yaitu untuk memprediksi ketepatan waktu lulus mahasiswa, akan tetapi, dalam penelitian ini akan dicoba beberapa hal baru guna menghasilkan suatu prediksi yang lebih akurat, yaitu dengan menerapkan teknik menyeimbangkan kelas data, hal ini dilakukan karena *dataset* yang digunakan dalam penelitian ini memiliki kelas data yang tidak seimbang. Penelitian ini mengunakan algoritma *random forest.*

* 1. **Rumusan Masalah**

Berdasarkanlatar belakang diatas, penulis dapat menarik beberapa permasalahan yang timbul, yaitu:

1. Perlu diketahui factor yang mempengaruhi ketepatan waktu lulus mahasiswa, agar dapat diatasi dan mahasiswa bisa lulus tepat pada waktunya
2. Perlu adanya alat untuk memprediksi ketepatan waktu lulus seorang Mahasiswa
	1. **Ruang Lingkup**

Agar lingkup yang akan dibahas tidak meluas atau menyimpang dari pokok pembahasan ke hal lain, serta untuk membatasi pembahasan masalah dan penyelesaiannya, maka perlu diberikan suatu batasan agar tidak menyimpang dari judul yang telah dibuat.

Ruang lingkup pembahasan dalam penelitian ini adalah:

1. Penelitian ini menggunakan *Random Forest* Untuk memprediksi ketepatan waktu lulus mahasiswa
2. *Dataset* yang digunakan merupakan dataset publik yang kemudian datanya ditambahkan dengan data pribadi yang dimiliki peneliti. Dataset yang digunakan memiliki 2 c*lass*
	1. **Tujuan Penelitian**

Tujuan dari penelitian ini adalah mendapatkan metode terbaik yang dapat melakukan prediksi ketepatan waktu lulus mahasiswa berdasarkan kriteria yang dimiliki Mahasiswa dengan baik. Metode yang diusulkan adalah metode *Random Forest* dan *tool* yang digunakan dalam penelitian ini adalah python

**BAB II**

**TINJAUAN PUSTAKA**

* 1. **Tinjauan Pustaka**

Tinjauan pustaka yang dijabarkan oleh peneliti dalam penelitian ini menggunakan beberapa referensi, seperti dari buku, prosiding dan jurnal. Berikut tinjauan pustaka yang mendukung teoritis dari penelitian ini

1. **Data Mining**

Proses analitis untuk meninjau kumpulan data dengan tujuan untuk menemukan adanya hubungan atau pola yang tidak terduga kemudian meringkas data tersebut dengan cara berbeda dengan sebelumnya agar dapat dipahami dan berguna bagi pemilik dari data, hal ini dapat dikatakan sebagai data mining Data mining (10) melakukan eksplorasi pada basis data guna menemukan pola yang tersembunyi, mencari informasi pemprediksi yang dimungkinkan terlewatkan atau terlupakan oleh para pelaku bisnis karena terletak diluar perkiraan mereka (11)

Teknik atau metode yang bisa digunakan dalam proses data mining jumlahnya sangatlah bervariasi, sehingga pemilihan teknik, algoritma atau metode yang pas dan sesuai sangat bertumpu pada proses dan tujuan *Knowledge discovery in database* (KDD) secara menyeluruh. *Data mining* dilakukan dengan melewati 3 tahap (12)

### Eksplorasi

Sebelum melakuakn penelitian Hal pertama yang dilakukan yaitumenyiapkan data dalam jumlah yang banyak, kemudian dilakukan proses *preprocessing* terhadap dataagarsinkron dengan kebutuhan penelitian, hal ini juga dilakukan guna membuang data yang berjumlah ganda. Data yang tersisa setelah melewati proses *preprocessing* iniadalah data yang benar-benar dapat digunakan sebagai objek penelitian.

### Pemodelan atau identifikasi pola

Pemodelan merupakan bagian penting dari proses data mining yang memilikitujuan untuk mengevaluasi sehingga akan menghasilkan mana prediksi yang akurat dan terbaik. Proses ini biasanya membutuhkan waktu yang lama karena menerapkan bisa jadi menerapkan beberapa model yang berbeda pada kumpulan data yang sama dan dilakukan berulang untuk nenatinya dibandingkan hasilnya dan dipilih mana yang lebih baik.

### Penerapan

Untuk mengetahui apakah model yang dibangun dapat dengan baik melakukan suatu prediksi maka dilakukan pengujian model terhadap data pembelajaran dan data pengujian.

1. **Prediksi**

Prediksi adalah upaya untuk memproyeksikan kondisi yang mungkin terjadi di masa depan dengan merujuk pada informasi dari periode sebelumnya. (13) Dalam data mining, fungsi merupakan sebagian kecil dari fungsi minor. Prediksi adalah suatu metode untuk menemukan potensi hasil di masa depan. Dalam teknologi data mining, dengan merujuk pada bukti dan temuan dari data, metode peramalan dapat mengantisipasi kemungkinan di masa yang akan datang. Perhitungan matematika dan statistik memainkan peran penting dalam proses prediksi. (14)

1. ***Random Forest***

*Random forest* ialah suatu teknik klasifikasi yang memperluas metode Decision Tree dengan memilih atribut secara acak pada setiap node guna menetapkan klasifikasi. Dalam proses klasifikasinya, keputusan didasarkan pada mayoritas suara dari pohon keputusan yang dihasilkan (15)

Algoritme *Random Forest* adalah suatu metode pembelajaran mesin yang mudah digunakan dan memiliki fleksibilitas. Keberagaman dan kesederhanaan membuatnya menjadi salah satu algoritme yang paling populer dalam penggunaannya (16)

1. ***Split Validation***

*Split validation* merupakan sebuah prosesmembagi (*split*) basis data secara acak menjadi beberapa bagian seperti bagian data untuk pembelajaran dan data pengujianuntuk mengevaluasi model yang diusulkan. *Split validation* juga dapat melakukan validasi untuk mengukur performa dari algoritma atau metode yang digunakan. Pengukuran performa pada umumnya dilakukan untukmenilai seberapa besarnilai akurasi sebuah model dalam memprediksi data sesuai dengan kelas data yang sebenarnya (17)



Gambar 1. Ilustrasi split validation

1. ***Confussion Matrix***

Untuk mengetahui performa sebuah model, peneliti dapat melihat pada *Confusion matrix,* karenadi dalam sebuah *Confusion matrix*berisi banyaknya baris data pengujian yang diprediksi benar dan tidak benar oleh model penelitan yang digunakan, tabel ini diperlukan untuk menentukan kinerja suatu model klasifikasi agar dapat ditentukan mana model yang memiliki performa terbaik(18) Dalam Confussion Matrix ada 4 istilah yang dapat merepresentasikan hasil dari klasifikasi, yaitu True Negatif (TN) merupakan jumlah data negatif yang dideteksi dengan benar, *False Negatif* (FN) merupakan jumlah data *positif* terdeteksi sebagai data *negative*, *True Positif* (TP) merupakan nilai positif yang dideteksi dengan benar dan *False Postif* (FP) merupakan data *negatif* namun terdeteksi secara *positif*(18)

Tabel 1. Pengujian confussion matrix

|  |  |
| --- | --- |
| **Klasifikasi** | **Hasil Prediksi** |
| **Benar** | **Salah** |
| Data Actual | Benar | *True Positif* (TP) | *False Negative* (FN) |
| Salah | *False Positif* (FP) | *True Negatif* (TN) |

Dengan data *confussion matrix*, maka akan dapatkan sebuah data yang lain yang pastinya akan sangat berguna untuk mengukur performa sebuah algoritma atau model yang digunakan, adapun data tersebut antara lain:

### Akurasi

Sebuah Parameter yang menunjukan seberapa tepat algoritma mengklasifikasikan data yang ada, persamaan yang digunakanuntuk menghitung nilai akurasi (18)

Akurasi$ =\frac{TP+TN}{TP+TN+FP+FN}$

### Presisi

Tingkat keakuratan hasil dari klasifikasi dengan jumlah total pengenalan yang dilakukan sistem. Untuk Formula menghitung nilai presisi dapat dituliskan menggunakan rumus persamaan berikut:

Presisi$ =\frac{True positif (TP)}{True positif\left(TP\right)+False positif (FP)}$

### *Recall*

*Recall* menunjukan total data yang betul diklasifikasi dalam sebuah kelas dibagi dengan jumlah total dalam kelas tersebut. Untuk Formula menghitung nilai *recall* dapat dituliskan menggunakan rumus persamaan berikut:

$$Recall=\frac{True positif (TP)}{True positif\left(TP\right)+False negatif(FN)}$$

### F1-*Score*

F1 *score* digunakan untuk menilai rata-rata *precision* dan *recall* hasil klasifikasi. Perhitungan F1 *Score* dapat dinyatakan dalam bentuk formula sebagai berikut:

$$F1 score =2\*\frac{(Recall\*Precision)}{(Recall+Precision)}$$

**BAB III**

**TUJUAN DAN MANFAAT PENELITIAN**

* 1. **Tujuan Penelitian**

Tujuan dari penelitian ini adalah mendapatkan metode terbaik yang dapat melakukan prediksi ketepatan waktu lulus berdasarkan kriteria yang dimiliki mahasiswa dengan baik. Metode yang diusulkan adalah metode *random forest* dan *tool* yang digunakan dalam penelitian ini adalah python

* 1. **Manfaat Penelitian**

Adapun manfaat dalam penelitian ini adalah untuk mempermudah melakukan prediksi ketepatan waktu lulus mahasiswa, sehingga jika ditemukan mahasiswa yang memiliki kendala mengenai ketepatan waktu lulus, maka dapat dengan mudah di prediksi.

**BAB IV**

**METODE PENELITIAN**

Pada bagian metode penelitian ini akan diuraikan secara detail, mulai dari dataset yang digunakan pada penelitian hingga bagaimana alur atau tahapan dari penelitian klasifikasi jenis hewan pada kebun binatang. Untuk digram alur penelitian dapat dilihat pada gambar 2 berikut ini:



Gambar 2. Diagram alur penelitian

* 1. ***Dataset***

Kumpulan data yang dijadikan sebagai objek atau bahan dari riset ini adalah data ketepatan waktu lulus pada salah satu universitas, akan tetapi untuk nama universitas tersebut dirahasiakan. Data ini diperoleh dari data terbuka atau umum dengan url https://www.kaggle.com/ yvirgantara/on-time-graduation-classification/. *Dataset* tersebut berisi kumpulan data Indeks Prestasi Kumulatif (IPK) mahasiswa. Jumlah data yang digunakan sebanyak 1687 data. Setiap sampel memiliki 4 *parameter* dan label kelas yang mengidentifikasinya ada sebanyak 2 kelas data, yaitu tepat dan tidak tepat.

## ***Preprocessing***

Salah satu proses yang tidak boleh dilewatkan sebelum masuk ke tahap prediksi adalah preprocessing. Tahap ini dianggap penting karena tidak semua data yang digunakan sebagai objek dalam penelitian sudah siap (19), terkadang terdapat data yang kosong hingga data yang *double*. Proses ini pula dicoba supaya informasi yang hendak digunakan cocok dengan kebutuhan riset. Beberapa langkah *preprocessing* juga diuji, seperti mengisi informasi yang kosong, menghapus data ganda atau *duplicate*, memeriksa konsistensi informasi, membersihkan data, dan memperbaiki kesalahan yang mungkin terdapat dalam *dataset* (20)

Proses *preprocessing* awal yang dilakukan pada penelitian ini adalah dengan melakukan proses pembersihan data, proses ini dilakukan guna menghilangkan data yang memiliki *missing value* dan *noise*. kemudian juga dilakukan pembersihan data ganda atau *double.*

Alat yang dioperasikan dalam penelitian ini untuk penulisan kode program melibatkan penggunaan Google *Colaboratory* dengan penerapan bahasa pemrograman *Python* dan model *Random Forest*.

## ***Split Validation***

Setelah menyelesaikan persiapan data melalui proses *preprocessing*, langkah selanjutnya adalah memisahkan data menjadi dua bagian. Bagian pertama berfungsi sebagai data pembelajaran, yang akan digunakan untuk melatih model, sementara bagian kedua akan dijadikan sebagai data pengujian. Data pembelajaran ini digunakan untuk menginstruksikan model, sedangkan data pengujian berperan dalam mengevaluasi sejauh mana model yang diusulkan mampu melakukan prediksi dengan akurat (21)

Setelah selesai pada tahap *preprocessing* ditemukan data sebanyak 1687 data siap digunakan dan 5 parameter dengan 2 kelas data, yaitu kelas tepat dan kelas tidak tepat. Data yang sudah siap digunakan ini kemudian di bagi menjadi 2 bagian, yaitu 50% data digunakan untuk mengajarkan model penelitian untuk mengenali data dan 50% lagi data untuk menguji model penelitian, agar dapat menilai apakah suatu model penelitian sudah dengan akurat mengenali suatu data, dalam penelitian ini didapatkan data pembelajaran sebanyak 1349 data dan data pengujian sebanyak 338 data.

Data Lulus Tepat waktu (1687 data)

*Data Training* 50% (*843 data*)

*Data Testing* 50%(844 data)

*Random Oversampling*

*Gambar 5 Proses Split Validation*

## ***Random Oversampling***

Setelah pembagian data selesai didapatkan data pembelajaran sebanyak 843 data dan data pengujian sebanyak 844 data. Pada data pembelajaran terdapat kelas data tepat sebanyak 773 data dan kelas data tidak tepat sebanyak 70 data. Dapat dilihat bahwa data ini berjumlah tidak seimbang antara kelas data tepat dan tidak tepat. Untuk mengatasi hal ini maka perlu dilakukannya *resampling* data. Teknik *resampling* dapat dikatakan bagian dari proses *preprocessing* untuk membuat distribusi data seimbang dalam proses pembelajaran (22) Ketidakseimbangan kelas data akan menghasilkan efek yang kurang baik bagi hasil prediksi seperti, kelas data minoritas sering disalah prediksikan sebagai kelas mayoritas (23) Oleh karena itu perlu diterapkannya teknik *resampling* yang diterapkan adalah teknik *random oversampling* (ROS)

Proses menyeimbangkan data dilakukan dengan metode menaikkan jumlah data dari kelas minoritas ke dalam data pembelajaran secara acak. Proses pemberian data ini dilakukan secara berulang sampai didapatkan jumlah data kelas minoritas *balance* dengan jumlah kelas mayoritas. Penggambaran dari proses *random oversampling* (ROS) bisa dilihat pada gambar 6

*Gambar 6 Proses Split Validation*

Setelah didapatkan kelas data yang seimbang maka, langkah selanjutnya adalah proses pengujian model *random forest* menggunakan data yang telah melalui tahap *resampling* sebelumnya dengan jumlah data masing-masing kelas sebanyak 773 data.

## ***Modelling***

Setelah kelas data seimbang, langkah selanjutnya adalah menerapkan model *random forest*. Random forest adalah metode data mining dan merupakan hasil pengembangan dari metode *decision tree* (24) Cara kerjanya, *random forest* memilih atribut secara acak untuk membuat jumlah pohon K dengan atribut yang berbeda tanpa pemangkasan. *Pada random forest* data pengujian akan diuji pada semua pohon yang dibangun dan kemudian output yang paling sering akan ditugaskan berulang hingga model yang dihasilkan memenuhi kriteria tertentu, contohnya nilai *loss function* yang cukup kecil. Penelitian ini menggunakan *Parameter* *random forest* seperti dapat dilihat pada Tabel 1

Tabel 1. *Parameter Random Forest*

| ***Parameter*** | **Nilai** |
| --- | --- |
| *N Estimator* | 10 |
| *Min sample split* | 2 |
| *Num trees* | 10 |
| *Boostrape* | *True* |
| *Random state* | 10 |
| *Seed* | 10 |
| *Criterion* | Gini |
| *Min sample leaf* | 1 |
| *Max Features* | *None* |

## ***Evaluation***

Langkah terakhir dalam serangkaian tahapan penelitian ini adalah evaluasi. Pada tahap evaluasi ini, hasil pengujian dibandingkan untuk menentukan kinerja model yang lebih baik. Evaluasi kinerja model yang dibangun melibatkan pemeriksaan *confusion matrix*, yang memberikan informasi tentang kebenaran prediksi data terhadap data sebenarnya. Dengan menggunakan *confusion matrix*, kita dapat memperoleh nilai F1-*Score*, Akurasi, dan *Recall*.

**BAB V**

**HASIL DAN LUARAN YANG DICAPAI**

* 1. **Hasil dan Luaran**

*Dataset* yang digunakan sebagai objek atau materi dalam penelitian ini adalah data mengenai ketepatan waktu lulus di salah satu universitas yang tidak akan diungkapkan namanya. Pada Tabel 2, dijabarkan deskripsi mengenai *dataset* yang digunakan dalam penelitian ini.

|  |
| --- |
| Tabel 2. Data Penelitian |

|  |  |  |
| --- | --- | --- |
| **No** | **Nama**  | **Keterangan** |
| 1 | Nama Variabel  | IP1, IP2, IP3 dan IP4 yaitu Nilai Indeks Prestasi Kumulatif (IPK) mahasiswa  |
| 2 | Kelas Data | Lulus dan Tidak Lulus |
| 3 | Jumlah Data  | 1687 |

Model yang diterapkan dalam penelitian ini adalah model algoritma *random forest*, dan pengujian dilaksanakan dengan menggunakan Python dengan desain model sebagai berikut:



*Gambar 7 Model Penelitian Random Forest*

Dengan menerapkan teknik *resampling* yaitu *random Oversampling* (ROS) untuk mengatasi ketidakseimbangan kelas data pada *dataset* yang digunakan dan juga model *random forest* untuk melakukan prediksi ketepatan waktu lulus, peneliti mendapatkan hasil yang diuraikan melalui *matris evaluation, yang* menampilkan bahwa model *random forest* lebih baik dalam melakukan prediksi ketepatan waktu lulus jika dibandingkan dengan model lain.

* + - 1. Akurasi *random forest*

Untuk mengetahui bahwa model melakukan prediksi dengan benar, yaitu meprediksi ketepatan waktu lulus sesuai dengan kelas datanya, maka dapat dilihat dari besaran akurasi yang dihasilkan oleh model yang dipakai. Pada Tabel 3 diuraikan hasil pengujian model terhadap data tepatan waktu lulus menggunakan teknik *resampling* *random oversampling* (ROS) dan mencoba beberapa teknik *split data*. Nilai akurasi tertinggi didapatkan dari model algoritma *random forest* dengan teknik *split data* sebesar 50% untuk *data pengujian* dan 50% untuk data pembelajaran yaitu sebesar 90,04%.

Tabel 3. Nilai Akurasi Hasil Pengujian Model

|  |
| --- |
| **Akurasi** |
| **Data *training*** | **Data *Testing*** | **Algoritma** |
| ***Random Forest*** | **Gradient Boosting** | **LGBM** |
| **50** | 50 | 90.04% | 73.70% | 89.80% |
| **60** | 40 | 88.30% | 70.80% | 88.90% |
| **70** | 30 | 89.90% | 69.60% | 88.95% |
| **80** | 20 |  89.05% | 68.9%  | 88.8% |

* + - 1. *Recall* *Random Forest*

Untuk mengetahui bahwa model melakukan prediksi dengan *actual*, yaitu meprediksi ketepatan waktu lulus sesuai dengan kelas datanya dibandingkan dengan kelas ketepatan waktu lulusnya tersebut, maka dapat dilihat dari besaran *recall* yang dihasilkan oleh model yang dipakai. Pada Tabel 4 diuraikan hasil pengujian model terhadap data ketepatan waktu lulus menggunakan teknik *resampling* *random oversampling* (ROS) dan mencoba beberapa teknik *split data*. Nilai akurasi tertinggi didapatkan dari model algoritma *random forest* dengan teknik *split data* sebesar 50% untuk *data pengujian* dan 50% untuk data pembelajaran yaitu sebesar 90,04%.

Tabel 4. Nilai *Recall* Hasil Pengujian Model

|  |
| --- |
| ***Recall*** |
| **Data *training*** | **Data *Testing*** | **Algoritma** |
| ***Random Forest*** | ***Gradient Boosting*** | **LGBM** |
| **50** | 50 | 90.04% | 73.70% | 89.80% |
| **60** | 40 | 88.30% | 70.80% | 88.90% |
| **70** | 30 | 89.90% | 69.60% | 88.95% |
| **80** | 20 |  89.05% | 68.9%  | 88.8% |

## *Precision Random Forest*

Untuk memahami perbandingan antara prediksi *true positif* dan jumlah data yang diprediksi positif, kita dapat mengobservasi besaran nilai presisi yang dihasilkan oleh model yang digunakan. Pada Tabel 4, hasil pengujian model terhadap data ketepatan waktu lulus menggunakan teknik *resampling random oversampling* (ROS) dan beberapa teknik *split* *data*. Model algoritma *random forest* dengan teknik *split data* sebesar 50% untuk data pengujian dan 50% untuk data pembelajaran menghasilkan nilai akurasi tertinggi, yakni sebesar 90,04%.

Tabel 5. Nilai Presisi Hasil Pengujian Model

|  |
| --- |
| **Presisi** |
| **Data *training*** | **Data *Testing*** | **Algoritma** |
| ***Random Forest*** | ***Gradient Boosting*** | **LGBM** |
| **50** | 50 | 87.05% | 87.90% | 86.70% |
| **60** | 40 | 86.40% | 86.20% | 85.10% |
| **70** | 30 | 85.10% | 88.80% | 85.10% |
| **80** | 20 |  89.05% | 68.9%  | 88.8% |

**BAB VI**

**KESIMPULAN DAN SARAN**

* 1. **Kesimpulan**

Penelitian yang dilakukan untuk memprediksi ketepatan waktu lulus menggunakan model *random forest* memberikan hasil bahwa prediksi terbaik didapatkan dengan menerapkan parameter *criterion gini* dengan n\_*estimators* sebanyak 10, *boostrape true*, *num trees* 10, *seed* sebanyak 10, *min sample leaf* sebanyak2 dan *max features* adalah *none*. Serta dengan menerapkan *Split validation* atau pembagian antara data pembelajaran sebesar 50% untuk data *pengujian* dan 50%. Selain parameter dan pembagian data, hal yang perlu dilakukan lagi adalah menyeimbangkan data yaitu menggunakan teknik *random oversampling* (ROS). Untuk mengevaluasi model yang disusun, Penulis menggunakan *metrics evaluation* seperti akurasi, *recall*, dan *precision*.

Hasil dari penelitian menunjukkan bahwa model yang diusulkan dapat dengan baik melakukan prediksi dibandingkan dengan model lainnya, yaitu dengan hasil nilai uji akurasi sebesar 90,04%, *recall* 90,04%, dan *precision* sebesar 87,05%. Dengan nilai ini maka tingkat keberhasilan untuk memprediksi ketepatan waktu lulus mahasiswa akan semakin akurat.

* 1. **Saran**

Untuk revisi hasil yang lebih akurat, diharapkan pada riset selanjutnya terdapat hal- hal yang perlu dicoba khususnya yang belum ada pada riset ini, seperti akumulasi *dataset* yang dijadikan objek riset sehingga sebaran data bisa *balance* serta *representative*

**DAFTAR PUSTAKA**

1. Heryana N, Ghassani FZ, Usrahmawan MA, Juardi D, Solehudin A. Kajian Model Prediksi Ketepatan Kelulusan Mahasiswa: Sebuah Literatur Review. Syntax J Inform. 2020;9(1):54.

2. S.A Amosea. Prediksi Ketepatan Waktu Lulus Mahasiswa Teknik Informatika Menggunakan Algoritma Naive Bayes. Dr Diss Inst Teknol Sepuluh Nop. 2021;

3. Jananto A. Algoritma Naive Bayes Untuk Mencari Perkiraan Waktu Studi Mahasiswa. INTECOMS J Inf Technol Comput Sci. 2021;4(1):20–9.

4. Satria F, Zamhariri Z, Syaripudin MA. Prediksi Ketepatan Waktu Lulus Mahasiswa Menggunakan Algoritma C4.5 Pada Fakultas Dakwah Dan Ilmu Komunikasi UIN Raden Intan Lampung. J Ilm Matrik. 2020;22(1):28–35.

5. Anwarudin A, Andriyani W, DP BP, Kristomo D. The Prediction on the Students’ Graduation Timeliness Using Naive Bayes Classification and K-Nearest Neighbor. J Intell Softw Syst. 2022;1(1):75.

6. Setiaji GG, Khoirudin K, Vydia V. Komparasi Metode Clustering K-Means Dan Fuzzy C-Means Untuk Mempredeksi Ketepatan Waktu Lulus. J Pengemb Rekayasa dan Teknol. 2019;15(1):38.

7. Faizah T, Jananto A. Perbandingan Algoritma C4 . 5 Dan Id3 Untuk Prediksi. Jatisi. 2021;8(2):980–90.

8. Mulyadi C, Sugiarto L. Penggunaan algoritma naïve bayes untuk prediksi ketepatan waktu lulus mahasiswa diploma 3 STMIK Cipta Darma Surakarta. Teknomatika [Internet]. 2021;11(01):21–30. Available from: http://ojs.palcomtech.ac.id/index.php/teknomatika/article/view/512

9. Samuel YT, Hutapea JJ, Jonathan B. Predicting the Timeliness of Student Graduation Using Decision Tree C4.5 Algorithm in Universitas Advent Indonesia. Int Conf Information, Commun Technol Syst. 2019;

10. Utomo DP, Mesran M. Analisis Komparasi Metode Klasifikasi Data Mining dan Reduksi Atribut Pada Data Set Penyakit Jantung. J Media Inform Budidarma. 2020;4(2):437.

11. Siregar AM, Puspabhuana A. Pengolahan Data Menjadi Informasi dengan Rapid Miner. Surakarta: Kekata Group; 2017. 200 p.

12. Arhami M, Nasir M. Data Mining - Algoritma dan Implementasi. Yogyakarta: ANDI; 2020. 226 p.

13. Almumtazah N, Azizah N, Putri YL, Negeri I, Ampel S. Prediksi jumlah mahasiswa baru menggunakan metode regresi linier sederhana. 2021;18:31–40.

14. Suryanto AA, Muqtadir A. PENERAPAN METODE MEAN ABSOLUTE ERROR ( MEA ) DALAM ALGORITMA REGRESI LINEAR UNTUK PREDIKSI PRODUKSI PADI. 2019;(1):78–83.

15. Luthfiana Ratnawati, Dwi Ratna Sulistyaningrum. Penerapan Random Forest untuk Mengukur Tingkat Keparahan Penyakit pada Daun Apel. J Sains Dan Seni Its. 2019;8(2):71–7.

16. Azhari M, Situmorang Z, Rosnelly R. Perbandingan Akurasi , Recall , dan Presisi Klasifikasi pada Algoritma. 2021;5(April):640–51.

17. Suryono S, Utami E, Luthfi ET. Klasifikasi Sentimen Pada Twitter Dengan Naive Bayes Classifier. Angkasa J Ilm Bid Teknol. 2018;10(1):89.

18. Purnama JJ, Rahayu S, Nurdiani S, Haryanti T, Mayangky NA. Analisis Algoritma Klasifikasi Neural Network Untuk Diagnosis Penyakit Kanker Payudara. J Pilar Nusa Mandiri. 2019;15(2):149–56.

19. Hermawan L, Bellaniar Ismiati M. Pembelajaran Text Preprocessing berbasis Simulator Untuk Mata Kuliah Information Retrieval. J Transform. 2020;17(2):188.

20. Nasir HD, Saharuna Z. Prediksi Tingkat Polusi Udara Dengan Data Mining. Pros Semin Nas Tek Elektro dan Inform. 2020;90–5.

21. Diantika S, Nalatissifa H, Supriyadi R, Maulidah N, Fauzi A. IMPLEMENTASI MULTI-CLASS GRADIENT BOOSTING UNTUK MENGKLASIFIKASIKAN JENIS HEWAN PADA KEBUN BINATANG. 2023;17(1):32–40.

22. Bagui S, Li K. Resampling imbalanced data for network intrusion detection datasets. J Big Data [Internet]. 2021;8(1). Available from: https://doi.org/10.1186/s40537-020-00390-x

23. Chen W, Li K. Self-supervised Learning for Medical Image Classification Using Imbalanced Training Data. Commun Comput Inf Sci. 2022;1590 CCIS:242–52.

24. Mao Y, He Y, Liu L, Chen X. Disease Classification Based on Eye Movement Features With Decision Tree and Random Forest. Front Neurosci. 2020;14(August):1–11.

**LAMPIRAN 1**

**Realisasi Biaya Penelitian**

|  |  |  |  |
| --- | --- | --- | --- |
| **NO** | **KEBUTUHAN** | **JUMLAH BIAYA** | **SUB TOTAL** |
| 1  | **Peralatan Penunjang** |   |   |
|   | a.Buku | Rp. 500.000 |  |
|   | b.Modem | Rp. 300.000 |  |
|  | c.Harddisk eksternal | Rp. 700.000 |  |
|   |  Sub Total |  | Rp. 1.50**0.000** |
|  2 | **Bahan Habis Pakai** |   |   |
|   | a.Pulsa telepon | Rp. 200.000 |  |
|   | b. Voucher internet | Rp. 200.000 |  |
|  | c. Alat tulis | Rp. 250.000 |  |
|  | d. Kertas A4 | Rp. 150.000 |  |
|  | e.Tinta Printer | Rp. 450.000 |  |
|  | f.Biaya Fotocopy | Rp. 50.000 |  |
|   |  Sub Total |  | Rp.1.500.000 |
|  3 | **Perjalanan** |   |   |
|   | 1. Transportasi perjalanan
 | Rp.1.000.000 |  |
|  | Sub Total |  | Rp.1.000.000 |
| 4 | **Biaya Lain-lain** |  |  |
|  | a. Biaya Publikasi | Rp.200.000 |  |
|   |   Sub Total |  | Rp.200.000 |
|  | BIAYA DITERIMA | Rp. 4.000.000 |
|  | TOTAL PENGELUARAN | Rp. 4.000.000 |
|  | SALDO | Rp. 0; |

**LAMPIRAN 2**

**Biodata Ketua dan Anggota Tim Peneliti**

**A. Identitas Ketua Peneliti**

* 1. **Identitas Diri**

a. Nama Lengkap dan Gelar : Sri Diantika, M. Kom

b. NIDN : 0310119501

c. Jabatan Fungsional : Instruktur

d. Program Studi : Sistem Informasi

e. Perguruan Tinggi : Universitas Bina Sarana Informatika

f. Jangka Waktu Penelitian : 8 Bulan

**2. Riwayat Pendidikan**

|  |  |  |
| --- | --- | --- |
|  | **S-1** | **S-2** |
| Nama Perguruan Tinggi  | Universitas BSI Bandung |  Universitas Nusa Mandiri Jakairta |
| Tahun Lulus | 2019 | 2021 |

Semua data yang saya isikan dan tercantum dalam biodata ini adalah benar dan dapat dipertanggungjawabkan secara hukum. Apabila di kemudian hari ternyata dijumpai ketidaksesuaian dengan kenyataan, saya sanggup menerima sanksi.

Demikian biodata ini saya buat dengan sebenarnya untuk memenuhi salah satu persyaratan dalam pengajuan Penelitian Mandiri.

Jakarta, 5 Februari 2024

Peneliti



 (Sri Diantika, M. Kom)

 NIP. 202111296

**B. Identitas Anggota Peneliti**

1. **Identitas Diri**

a. Nama Lengkap dan Gelar : Hiya Nalatissifa, M. Kom

b. NIDN : 0307059701

c. Jabatan Fungsional : Asisten Ahli

d. Program Studi : Sistem Informasi Akuntansi Kampus

 Tegal (D3)

e. Perguruan Tinggi : Universitas Bina Sarana Informatika

f. Jangka Waktu Penelitian : 8 Bulan

1. **Riwayat Pendidikan**

|  |  |  |
| --- | --- | --- |
|  | **S-1** | **S-2** |
| Nama Perguruan Tinggi  | Universitas BSI Bandung  | Sj Universitas Nusa Mandiri Jakarta |
| Tahun Lulus | 2019 | 2021 |

Semua data yang saya isikan dan tercantum dalam biodata ini adalah benar dan dapat dipertanggungjawabkan secara hukum. Apabila di kemudian hari ternyata dijumpai ketidak sesuaian dengan kenyataan, saya sanggup menerima sanksi.

Demikian biodata ini saya buat dengan sebenarnya untuk memenuhi salah satu persyaratan dalam pengajuan Penelitian Mandiri.

 Jakarta5 Februari 2024

 Anggota Peneliti



 (Hiya Nalatissifa, M. Kom)

 NIP. 202109235

**Biodata Anggota Peneliti 2**

1. **Identitas Diri**

a. Nama Lengkap dan Gelar : Riki Supriyadi, M.Kom

b. NIP : 202103282

c. Jabatan Fungsional : AsistenAhli

d. Program Studi : Sains Data

e Perguruan Tinggi : Universitas Nusa Mandiri

f. Jangka Waktu Penelitian : 8 Bulan

1. **Riwayat Pendidikan**

|  |  |  |
| --- | --- | --- |
|  | **S-1** | **S-2** |
| Nama Perguruan Tinggi  | Universitas BSI Bandung |  Universitas Nusa Mandiri jakarta |
| Tahun Lulus | 2019 | 2021 |

Semua data yang saya isikan dan tercantum dalam biodata ini adalah benar dan dapat dipertanggungjawabkan secara hukum. Apabila di kemudian hari ternyata dijumpai ketidak sesuaian dengan kenyataan, saya sanggup menerima sanksi.

Demikian biodata ini saya buat dengan sebenarnya untuk memenuhi salah satu persyaratan dalam pengajuan Penelitian Mandiri.

 Jakarta, 5 Februari 2024

 Anggota Peneliti



 (Riki Supriyadi, M.Kom)

 NIP. 202103282

**Biodata Anggota Peneliti 3**

* + - 1. **Identitas Diri**

a. Nama Lengkap dan Gelar : Nurlaelatul Maulidah, M. Kom

b. NIDN : 0303089601

c. Jabatan Fungsional : Asisten Ahli

d. Program Studi : Sistem Informasi Akuntansi Kampus

 Tegal (D3)

e Perguruan Tinggi : Universitas Bina Sarana Informatika

f. Jangka Waktu Penelitian : 8 Bulan

* + - 1. **Riwayat Pendidikan**

|  |  |  |
| --- | --- | --- |
|  | **S-1** | **S-2** |
| Nama Perguruan Tinggi  | Universitas BSI Bandung |  Universitas Nusa Mandiri |
| Tahun Lulus | 2019 | 2021 |

Semua data yang saya isikan dan tercantum dalam biodata ini adalah benar dan dapat dipertanggungjawabkan secara hukum. Apabila di kemudian hari ternyata dijumpai ketidak sesuaian dengan kenyataan, saya sanggup menerima sanksi.

Demikian biodata ini saya buat dengan sebenarnya untuk memenuhi salah satu persyaratan dalam pengajuan Penelitian Mandiri.

 Jakarta, 5 Februari 2024

 Anggota Peneliti



 (Nurlaelatul Maulidah, M. Kom)

 NIP. 202110280

**Biodata Anggota Peneliti 4**

**Identitas Diri**

a. Nama Lengkap dan Gelar : Ahmad Fauzi, M. Kom

1. NIDN : 0326119401
2. Jabatan Fungsional : Lektor
3. Program Studi : Sistem Informasi
4. Perguruan Tinggi : Universitas Bina Sarana Informatika
5. Jangka Waktu Penelitian : 8 Bulan

**2. Riwayat Pendidikan**

|  |  |  |
| --- | --- | --- |
|  | **S-1** | **S-2** |
| Nama Perguruan Tinggi  | Universitas BSI Bandung |  STMIK Nusa Mandiri Jakarta |
| Tahun Lulus | 2018 | 2020 |

Semua data yang saya isikan dan tercantum dalam biodata ini adalah benar dan dapat dipertanggungjawabkan secara hukum. Apabila di kemudian hari ternyata dijumpai ketidak sesuaian dengan kenyataan, saya sanggup menerima sanksi.

Demikian biodata ini saya buat dengan sebenarnya untuk memenuhi salah satu persyaratan dalam pengajuan Penelitian Mandiri.

Jakarta, 5 Februari 2024

Anggota Peneliti

 

 (Ahmad Fauzi, M.Kom)

 NIP. 201903058