e-ISSN 2962-6129
Seminar Nasional Mahasiswa Ilmu Komputer dan Aplikasinya (SENAMIKA)
Jakarta-Indonesia, 07 Mei 2023

Klasifikasi Ketepatan Lama Studi Mahasiswa Dengan Algoritma Random Forest Dan Gradient Boosting (Studi Kasus Fakultas Ilmu Komputer Universitas Pembangunan Nasional Veteran Jakarta)

Muhammad Labib Mu'tashim [1], Ati Zaidiah [2], Bambang Saras Yulistiawan[3]
[1]Informatika UPNVJ , [2,3]S1 Sistem Informasi UPNVJ
Jl. RS. Fatmawati Raya, Pd. Labu, Kec. Cilandak, Kota Depok, Daerah Khusus Ibukota Jakarta
[1] muhammadlm@upnvj.ac.id, [2] atizaidiah@upnvj.ac.id, [3]bambangsarasyulistiawan@upnvj.ac.id

Abstrak. Universitas umumnya pada setiap tahun menerima mahasiswa baru dan memiliki kuota yang berbeda pada setiap jurusannya, begitu juga dengan Fakultas Ilmu Komputer (FIK) Universitas Pembangunan Nasional (UPN) Veteran Jakarta. Melimpahnya data akademik di FIK UPN Veteran Jakarta bisa diolah sesuai yang dibutuhkan dan berguna mencari informasi penting demi pengembangan fakultas menjadi lebih baik. Maka dari itu dilakukan penelitian untuk menganalisis mahasiswa yang lulus tepat waktu maupun tidak tepat waktu dengan data mining menggunakan metode *Random Forest* dan *Gradient Boosting* untuk mengetahui tingkat akurasi dan menentukan mana model klasifikasi yang terbaik pada ketepatan lulus mahasiswa. Analisis menggunakan data mahasiswa S1 FIK UPN Veteran Jakarta angkatan 2015 - 2017. Hasil uji coba sampel pada 590 data, algoritma random forest 10 k-fold mendapatkan akurasi 82,64% dan pada gradient boosting 3 k-fold mendapatkan akurasi 79,66%. Hasil penelitian ini digunakan sebagai salah satu dasar pengambilan keputusan untuk menentukan kebijakan oleh pihak fakultas.

Kata Kunci : Klasifikasi, *Random Forest*, *Gradient Boosting*, Kelulusan

1. Pendahuluan

Pada setiap tahunnya universitas selalu melepaskan mahasiswanya yang telah lulus dan bergelar. Jika dilihat dari informasi kelulusan mahasiswa, mahasiswa yang lulus maupun belum lulus terdapat peningkatan atau penyusutan setiap tahunnya. Banyak mahasiswa yang lulus tepat pada waktunya, namun tidak jarang pula mahasiswa lulus tidak tepat pada waktunya[14].

Fakultas Ilmu Komputer Universitas Pembangunan Nasional Veteran Jakarta (FIK UPN Veteran Jakarta) merangkum data mahasiswanya dan mengolahnya sesuai dengan apa yang dibutuhkan. Peningkatan jumlah data diakibatkan dari tidak seimbangnya mahasiswa yang masuk dan yang sudah lulus membuat penumpukan dan jumlah mahasiswa yang semakin banyak. Data tersebut haruslah diolah dengan baik dengan teknik yang tepat.

Maka dari itu, untuk menganalisis dan mengklasifikasikan mahasiswa yang lulus digunakanlah teknik data mining untuk mengklasifikasikan ketepatan lama studi mahasiswa. Terdapat beberapa metode klasifikasi yang popular dalam data mining, namun dalam penelitian ini penulis menggunakan algoritma klasifikasi *Random Forest* dan *Gradient Boosting*.

Berdasarkan referensi dari beberapa jurnal, terdapat persamaan klasifikasi menggunakan algoritma *Random Forest* bahwa algoritma ini memiliki akurasi yang lebih baik dibanding algoritma lainnya, seperti pada jurnal Syauqi Amri Yahya pada tahun 2018, perbandingan menggunakan *Random Forest* dan SVM menghasilkan akurasi sebesar 80% pada *Random Forest*, begitu juga pada penelitian Erwinsyah Rico Agusta pada 2021 menyimpulkan bahwa akurasi *Random Forest* lebih tinggi sebesar 88,53% dibanding algoritma pembandingnya yakni *Naïve Bayes* yang sebesar 78,53%[2][16].

Penulis juga menggunakan algoritma klasifikasi Gradient Boosting sebagai perbandingan karena kedua algoritma ini adalah *ensemble learning* yang merupakan tingkat lanjut dari algoritma klasifikasi Decision Tree[13].

Dari penjelasan diatas, dilakukanlah analisis klasifikasi ketepatan kelulusan mahasiswa dengan menggunakan algoritma *Random Forest* dan *Gradient Boosting* untuk mengetahui akurasi yang terbaik. Analisis ini akan diterapkan pada FIK UPN Veteran Jakarta untuk menjadi informasi yang berguna dan untuk mengetahui mahasiswa yang memiliki potensi untuk lulus tepat pada waktunya ataupun lulus terlambat.

2. Landasan Teori

2.1 Kelulusan Mahasiswa

Pembelajaran disuatu perguruan tinggi dinyatakan lulus jika mahasiswanya mampu mengasah kemampuan keilmuannya yang didapat selama belajar di tempat tersebut[6].

Institusi pendidikan dan perguruan tinggi memiliki data mahasiswa yang berlimpah, seperti data jumlah kelulusan tiap tahunnya, akademik mahasiswa yang memiliki informasi yang berguna bagi pihak universitas[15].

Selama ini Fakultas Ilmu Komputer UPN Veteran Jakarta belum memiliki pola klasifikasi kelulusan tepat waktu sebagai acuan untuk mengklasifikasikan jumlah kelulusan tepat waktu. Jumlah mahasiswa yang banyak perlu juga mendapatkan perhatian khusus dari sisi penerimaan mahasiswa sebagai input fakultas dan ketepatan lulus mahasiswa sebagai *output* mahasiswanya.

2.2 Esemble Learning

Ensemble Learning adalah metode dimana algoritma mempelajari data dengan cara menggabungkan beberapa algoritma secara bersamaan untuk mendapatkan hasil pemodelan yang kuat daripada hanya menggunakan satu algoritma saja.

Ada beberapa tipe dalam ensemble learning yakni *bagging*, *boosting*, dan *stacking*. Pada penelitian ini penulis menggunakan dua tipe ensemble learning yakni *bagging* dan *boosting*[13].

1) Bagging

Teknik *bagging* atau singkatan dari *boostrap aggregating* adalah metode ensemble yang dimana prosesnya menggunakan beberapa model dari algoritma yang sama, dan melatih setiap model pada sample yang berbeda dengan dataset yang sama. *Bagging* ini melakukan traning data secara terpisah (*pararel*), yang salah satu contohnya adalah algoritma *Random Forest* yang dipakai pada penelitian ini.

2) Boosting

Teknik boosting sendiri melakukan training data secara *sequential*, dimana model dibangun secara bertahap, yakni dengan melatih model baru untuk memperbaiki kesalahan pada model sebelumnya. Algoritma yang dipakai pada penelitian ini yakni *Gradient Boosting* merupakan algoritma *ensemble learning* dengan tipe *boosting*.

2.3 Algoritma Random Forest

Algoritma berbasis *tree* adalah metode pembelajaran dalam machine learning yang digunakan untuk memecahkan masalah yang membutuhkan data yang cukup besar. Sekian dari banyaknya Algoritma berbasis *tree* ini adalah Algoritma *Random Forest*. Algoritma ini diusulkan oleh Tin Kam Ho pada tahun 1995, yang merupakan kombinasi dari x *tree* yang digabung dan dijadikan satu model.

Random forest merupakan gabungan dari masing-masing tree yang membentuk hutan (*forest*) dengan melakukan tahapan training data yang dimiliki dan mendapatkan hasil akhir yang berbentuk voting[2].

2.4 Algoritma Gradient Boosting

Gradient Boosting adalah grup dalam algoritma machine learning yang merupakan metode peningkatan, yang secara *iterative* belajar dan mengkombinasikan banyak *weak learner* untuk membuat model yang kuat. Ide dibalik gradient *boosting* adalah mengambil hipotesis yang lemah/ *weak learner* dan membuat serangkaian penyesuaian sehingga meningkatkan kualitas hipotesis/ *learner*[7].

Gradient Boosting merupakan algoritma *ensemble learning* tipe *boosting* yang menggunakan model *decision tree* untuk memprediksi nilai. *Gradient boosting* bisa menyelesaikan kasus klasifikasi maupun regresi. Struktur data dari *gradient boosting* adalah *decision tree*[11].

2.4 Confussion Matrix

Ketika melakukan perhitungan akurasi pada algoritma klasifikasi, dilakukan metode *Confussion Matrix*. Metode ini menghasilkan nilai akurasi, presisi dan *recall*. Akurasi adalah prosentase ketepatan klasifikasi data yang diklasifikasikan secara benar setelah dilakukan pengujian[2]. Penelitian ini mengukur akurasi dengan metode *confusion matrix* sebagai berikut.

Tabel 1. Confusion Matrix

Kategori		**Nilai Sebenarnya**	
		Benar	Salah
Nilai Prediksi	Benar	TP	FP
	Salah	FN	TN

Keterangan :
TP : Klasifikasi benar pada prediksi dan benar pada nilai sebenarnya
FP : Klasifikasi benar pada prediksi dan salah pada nilai sebenarnya
FN : Klasifikasi salah pada prediksi dan benar pada nilai sebenarnya
TN : Klasifikasi salah pada prediksi dan salah pada nilai sebenarnya

3. Metode Penelitian

3.1 Pengumpulan Data

Penelitian ini mengambil dan mengumpulkan data yang berasal dari dari Biro Akademik dan Kemahasiswaan UPNVJ. Data ini adalah data mahasiswa yang berada di Sistem Informasi Akademik (SIAKAD). Variabel yang dipakai yakni data akademik mahasiswa S1 Informatika dan Sistem Informasi FIK Angkatan 2015, 2016, 2017.

Tabel 2. Variabel Penelitian

No	Variabel	Keterangan	Value	Jenis Data	Tipe Data
1	Program Studi	Program Studi Mahasiswa	Informatika S1 Sistem Informasi S1	Kategorikal	String
2	Jenis Kelamin	Jenis Kelamin mahasiswa	Laki-laki Perempuan	Kategorikal	String
3	Jenis Sekolah	Jenis pendidikan terakhir yang ditempuh sebelumnya.	MA SMA SMK SMTA	Kategorikal	String
4	Provinsi Sekolah	Provinsi asal sekolah mahasiswa yang bersangkutan	Jakarta Luar Jakarta	Kategorikal	String
5	Jalur Masuk	Jalur masuk mahasiswa ketika mendaftar ke universitas	SBMPTN SEMA-UPNVJ SNMPTN	Kategorikal	String
6	Beasiswa/ Non	Potongan biaya kuliah per semester	Non Beasiswa Beasiswa	Kategorikal	String
7	Indeks Prestasi Kumulatif (IPK)	Nilai prestasi belajar kumulatif yang memiliki rentang nilai 0 – 4.	0 – 4	Numerikal	Integer
8	Indeks Prestasi Semester (IPS)	Nilai prestasi belajar per semester yang memiliki rentang nilai 0 – 4.	0 – 4	Numerikal	Integer
9	UKT Kuliah	Keterangan biaya kuliah mahasiswa tiap semester	Rp500.000 – Rp9.100.000	Numerikal	Integer
10	Status Mahasiswa	Class target mahasiswa yang berpotensi lulus tepat waktu atau lulus terlambat	Tepat Waktu Terlambat	Kategorikal	String

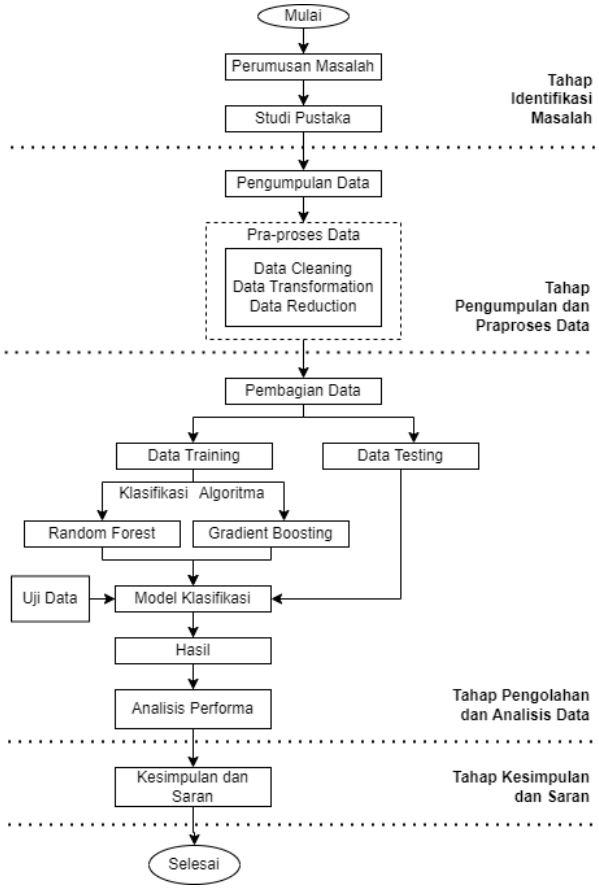
Pada proses selanjutnya yakni mengumpulkan data pendukung seperti literatur maupun jurnal yang berhubungan dengan penelitian kelulusan mahasiswa dan yang terkait dengan data mining.

Penelitian dengan judul Klasifikasi Ketepatan Lama Studi Mahasiswa Menggunakan SVM dan Random Forest yang dibuat oleh Syauqi Amri Yahya dari Universitas Islam Indonesia pada tahun 2018 melakukan prediksi mahasiswa lulus tepat waktu atau tidak tepat waktu berdasarkan input jenis kelamin, program studi, provinsi, pekerjaan ayah dan ibu, Pendidikan ayah dan ibu, SMA, IPK semester. Setelah dilakukan uji coba, dapat disimpulkan bahwa akurasi model pada metode *Random Forest* sebesar 80%, sedikit lebih tinggi dibandingkan dengan metode SVM yakni 68% pada kernel RBF dan 77% pada kernel sigmoid. Perbedaan mendasar pada penelitian ini adalah penulis tetap menggunakan algoritma *Random Forest* namun dengan pendekatan komparasi yang berbeda yakni *Gradient Boosting*[16].

Kemudan jurnal penelitian yang berjudul *Implementation of Data Mining for Drop-Out Prediction using Random Forest Method* dibuat oleh Meylani Utari, Budi Warsito, dan Retno Kusumaningrum dari Universitas Diponegoro yang dipublish pada tahun 2020. Atribut yang digunakan juga mirip dengan atribut pada penelitian penulis, namun dengan *class target* yakni *drop out* atau lulus. Penelitian ini menggunakan SMOTE untuk mengatasi kelas yang tidak seimbang, dan juga pemilihan k yang bervariasi mulai dari k = 2, k = 4, k = 6, k = 8 dengan akurasi tertinggi yakni sebesar 93,42% pada k=2[10].

3.2 Pengolahan Data

Data mahasiswa yang telah diambil dari AKPK yaitu berisi atribut Program Studi, Jenis Kelamin, Jenis Sekolah, Provinsi Sekolah, Jalur Masuk, Beasiswa/ Non, IPK, IPS, UKT Kuliah, dan Status Mahasiswa. Data yang sudah ada kemudian dilakukan program pre-processing data, dimana data yang kosong, tidak lengkap, maupun berisikan nilai NaN perlu dihapus agar tidak mengganggu perhitungan. Kemudian dilakukan proses *data transformation*, yaitu data yang tidak diperlukan dihapus juga dan hanya diambil data yang relevan dengan penelitian ini.



Gambar. 1. Alur Penelitian

Penelitian dimulai dengan pengumpulan jurnal dan referensi terlebih dulu seperti pada Gambar 1. Penulis mengambil referensi metode yang sesuai dengan penelitian ini, yakni terfokus pada klasifikasi ketepatan lama studi mahasiswa. Kemudian pada pengumpulan data, kami mencari data dari FIK UPNVJ dan AKPK Universitas, dan melakukan *pra-processing* data hingga data bersih dan siap dipakai. Data tersebut akan di lanjutkan ke proses perhitungan menggunakan *Random Forest* dan *Gradient Boosting*. Data yang sudah siap dibagi menjadi dua bagian yakni data testing dan juga data training. Setelah semua hal dilakukan, baru kemudian bisa diketahui klasifikasi ketepatan lama studi mahasiswa sesuai spesifikasi yang disiapkan.

4. Pembahasan dan Hasil

4.1 Pra-proses Data

Pra-processing data merupakan tahapan untuk menghilangkan noise agar menghasilkan data yang siap pakai untuk proses klasifikasi. Tahapan pre-processing yang ada pada penelitian ini adalah data cleaning, data transformation, dan data reduction.

Pada data cleaning, pembersihan data dilakukan untuk menghilangkan data yang tidak lengkap/ kosong. Pembersihan ini juga dilakukan untuk menghilangkan kolom yang tidak dipakai.

Tabel 3. Data yang sudah di Clean

prodi	tahun_angk	nama	jenis_kelamin	jenis_sekolah	prov_sekolah	jalur_masuk	bea_non	kel_ukt	tarif_ukt	status_mhs	ipk	ket_lulus	tgl_lulus	sem_1	sem_2	sem_3	sem_4	sem_5	sem_6	sem_7
Informatika	2015	RAHMAT FURQON	L	MAN	Sumatera Barat	SBMPTN	1	0	2400000	LULUS	3,05	TEPAT WAKTU	30-Jul-19	3,04	3,24	3,02	3,28	2,95	3,25	3,03
Informatika	2015	MOEHAMMAD ALDIN	L	SMAN	Jawa Barat	SBMPTN	0	7	9100000	LULUS	3,11	TERLAMBAT	28-Jul-20	2,92	3,1	2,99	3,38	2,82	2,5	2,75
Informatika	2015	BUDI PRASETYO	L	SMAN	DKI Jakarta	SBMPTN	0	7	9100000	LULUS	3,25	TEPAT WAKTU	30-Jul-19	3,08	3,35	3,28	3,56	3,11	3,2	3,46
Informatika	2015	ABDUL RAHIM	L	SMAS	DKI Jakarta	SBMPTN	0	3	5100000	LULUS	3,64	TEPAT WAKTU	30-Jul-19	3,53	3,55	3,73	3,81	3,44	3,88	3,82
Informatika	2015	ARRUM SEKAR MELATI	P	SMAS	DKI Jakarta	SBMPTN	0	7	9100000	LULUS	3,82	TEPAT WAKTU	30-Jul-19	3,95	3,89	3,94	3,84	3,51	3,75	3,89
Informatika	2015	SIGIT ISPRAMONO HADI	L	SMKN	DKI Jakarta	SBMPTN	0	7	9100000	LULUS	3,71	TERLAMBAT	27-Jan-22	3,76	3,75	3,78	3,84	3,54	3,38	3,72
Informatika	2015	KRISNA MALIK SUKARNO	L	SMAN	Jawa Barat	SBMPTN	0	7	9100000	LULUS	3,68	TEPAT WAKTU	30-Jul-19	3,76	3,78	3,79	3,84	3,47	3,62	3,72
Informatika	2015	AHMAD ZAKY ARROZY	L	SMAS	DKI Jakarta	SBMPTN	0	4	6100000	LULUS	3,47	TEPAT WAKTU	30-Jul-19	3,78	3,67	3,38	3,73	3,08	3,45	3,62
Informatika	2015	WILDAN MAHAD TAHTADI	L	SMAN	Jawa Barat	SBMPTN	0	7	9100000	LULUS	3,63	TEPAT WAKTU	30-Jul-19	3,86	3,73	3,8	3,8	3,22	3,25	3,79
Informatika	2015	DEA SAVIRA RAHMAWATI	P	SMAN	Kepulauan Riau	SBMPTN	0	3	5100000	LULUS	3,77	TEPAT WAKTU	30-Jul-19	3,97	3,94	3,97	3,67	3,57	3,81	3,86
Informatika	2015	MUHAMMAD ARIFUDIN HA	L	SMAS	Banten	SBMPTN	0	2	1000000	LULUS	3,45	TEPAT WAKTU	30-Jul-19	3,39	3,55	3,36	3,72	3,49	3,39	3,54
Informatika	2015	MOHAMAD RIZKI ALIF RAM	L	SMAN	DKI Jakarta	SBMPTN	0	3	5100000	LULUS	3,68	TEPAT WAKTU	30-Jul-19	3,69	3,72	3,77	3,72	3,82	3,57	3,6
Informatika	2015	YUGO BAYU PRASTYO	L	SMAN	Kepulauan Bangka	SBMPTN	0	7	9100000	LULUS	3,28	TEPAT WAKTU	30-Jul-19	3,3	3,23	3,43	3,53	2,89	3,12	3,5
Informatika	2015	GANDA HASIHOLAN TAMBI	L	SMAN	DKI Jakarta	SBMPTN	0	6	8100000	LULUS	3	TERLAMBAT	16-Feb-21	3,08	3,34	2,9	3,12	2,21	1,38	2,9
Informatika	2015	INDAH MUTIA HAPSARI	P	SMAN	DKI Jakarta	SBMPTN	0	6	8100000	LULUS	3,5	TEPAT WAKTU	30-Jul-19	3,75	3,74	3,36	3,39	3,28	3,71	3,41
Informatika	2015	JASON ERRYANTO TJHAND	L	SMAN	DKI Jakarta	SBMPTN	0	4	6100000	LULUS	3,68	TEPAT WAKTU	30-Jul-19	3,75	3,84	3,84	3,75	3,38	3,75	3,67
Informatika	2015	MUHAMMAD FAIZ FATAH	L	SMAN	DKI Jakarta	SBMPTN	0	2	1000000	LULUS	3,68	TEPAT WAKTU	30-Jul-19	3,59	3,85	3,86	3,75	3,59	3,57	3,57
Informatika	2015	RESHA TIWA ALUI SARAGIH	P	MAN	DKI Jakarta	SBMPTN	0	2	1000000	LULUS	3,44	TEPAT WAKTU	30-Jul-19	3,5	3,5	3,57	3,71	2,9	3,55	3,65
Informatika	2015	DEWI HAJAR	P	MAN	DKI Jakarta	SBMPTN	0	2	1000000	LULUS	3,56	TEPAT WAKTU	30-Jul-19	3,69	3,75	3,58	3,72	3,36	3	3,86
Informatika	2015	ATIKA DWI PUTRI	P	SMKN	Jawa Barat	SBMPTN	0	3	5100000	LULUS	3,69	TEPAT WAKTU	30-Jul-19	3,61	3,66	3,84	3,84	3,62	3,62	3,63
Informatika	2015	YUSUF FADHILAH	L	SMAN	DKI Jakarta	SBMPTN	0	2	1000000	LULUS	3,48	TEPAT WAKTU	30-Jul-19	3,42	3,34	3,67	3,75	3,1	3,62	3,62
Informatika	2015	WILLIAM ALEXZANDER	L	SMAN	Jawa Barat	SBMPTN	0	1	500000	LULUS	3,33	TEPAT WAKTU	30-Jul-19	3,8	3,3	3,45	3,42	3	3,04	3,56
Informatika	2015	CATYA INDRA PRATIWI	P	SMAN	Jawa Barat	SBMPTN	0	4	6100000	LULUS	3,57	TEPAT WAKTU	30-Jul-19	3,62	3,7	3,84	3,76	3,44	3,12	3,53
Informatika	2015	DWI WAHYU PERMATA	L	SMKN	Jawa Barat	SBMPTN	0	2	1000000	LULUS	3,51	TEPAT WAKTU	30-Jul-19	3,61	3,55	3,64	3,74	3,19	3,25	3,54
Informatika	2015	TYARA PUNDHI OVIANA	P	SMAN	DKI Jakarta	SBMPTN	0	7	9100000	LULUS	3,79	TEPAT WAKTU	30-Jul-19	3,91	3,81	3,84	3,85	3,64	3,8	3,86
Informatika	2015	MARSHA NADIRA	P	SMAN	Banten	SBMPTN	0	7	9100000	LULUS	3,87	TEPAT WAKTU	30-Jul-19	4	3,97	3,95	3,94	3,67	3,62	3,86
Informatika	2015	HANIF ALFI HAMMAMI	L	SMAN	Jawa Barat	SBMPTN	0	6	8100000	LULUS	3,7	TEPAT WAKTU	30-Jul-19	4	3,77	3,78	3,71	3,54	3,4	3,72
Informatika	2015	RADHITYA GILANG DWI PR	L	SMKS	Jawa Barat	SEMA UPNVJ	0	5	7100000	LULUS	3,3	TEPAT WAKTU	30-Jul-19	3,18	3,27	3,33	3,64	3,07	3,35	3,56
Informatika	2015	MUHAMAD ADJIE BAGASKO	L	SMKS	Jawa Barat	SEMA UPNVJ	0	3	5100000	LULUS	3,16	TEPAT WAKTU	30-Jul-19	3,33	2,94	2,99	3,28	3	3,21	3,25

Dari seluruh data yang ada yakni 410 baris data terdapat 52 data missing value yang tersebar di beberapa atribut penelitian, yakni di kolom atribut IPK dan IP Semester. Proses selanjutnya adalah data transformation. Proses ini adalah proses perubahan data ke bentuk yang sesuai dengan data mining. Perubahan data ini bisa ke dalam kategori atau nilai tertentu, namun tidak mengubah isi data.

Tabel 4. Data yang sudah di transformation

prodi	tahun_angk	nama	jenis_kelamin	jenis_sekolah	prov_sekolah	jalur_masuk	bea_non	kel_ukt	tarif_ukt	status_mhs
0	2015	MUHAMMAD IRSAL LUBIS	0	1	1	0	0	2	1000000	NA
0	2015	RAHMAT FURQON	0	0	1	0	1	0	2400000	LULUS
0	2015	MOEHAMMAD ALDIN	0	1	1	0	0	7	9100000	LULUS
0	2015	BUDI PRASETYO	0	1	0	0	0	7	9100000	LULUS
0	2015	ABDUL RAHIM	0	1	0	0	0	3	5100000	LULUS
0	2015	ANDI RAHMADI	0	3	1	0	0	2	1000000	NA
0	2015	ARRUM SEKAR MELATI	1	1	0	0	0	7	9100000	LULUS
0	2015	SIGIT ISPRAMONO HADI	0	2	0	0	0	7	9100000	LULUS
0	2015	KRISNA MALIK SUKARNO	0	1	1	0	0	7	9100000	LULUS
0	2015	AHMAD ZAKY ARROZY	0	1	0	0	0	4	6100000	LULUS
0	2015	WILDAN MAHAD TAHTADI	0	1	1	0	0	7	9100000	LULUS
0	2015	DEA SAVIRA RAHMAWATI	1	1	1	0	0	3	5100000	LULUS
0	2015	MUHAMMAD ARIFUDIN HANAFIA	0	1	1	0	0	2	1000000	LULUS
0	2015	Jabbar Nugratullah	0	1	1	0	0	7	9100000	NA
0	2015	MOHAMAD RIZKI ALIF RAMDHANI	0	1	0	0	0	3	5100000	LULUS
0	2015	YUGO BAYU PRASTYO	0	1	1	0	0	7	9100000	LULUS
0	2015	GANDA HASIHOLAN TAMBUNAN	0	1	0	0	0	6	8100000	LULUS
0	2015	INDAH MUTIA HAPSARI	1	1	0	0	0	6	8100000	LULUS

Selanjutnya tahap reduction, merupakan pengurangan volume dataset menjadi lebih kecil dibanding sebelumnya. Penelitian ini melakukan pengurangan volume dengan mengurangi atribut yang tidak diperlukan dalam penelitian. Dari total 26 atribut, dikurangi hingga menjadi 15 atribut yang dipakai pada penelitian ini.

4.2 Persiapan Data Training dan Data Testing

Data yang akan dianalisis terdapat 2 kelas, yaitu tepat waktu dan terlambat, dengan jumlah tiap kelas yakni 296 data dan 34 data. Perbedaan data pada dua kelas tersebut sangatlah jauh, sehingga data dikatakan *imbalance* atau tak seimbang. Maka dari itu diperlukan *balancing data* dengan metode *over sampling* untuk menghasilkan masing masing kelas yang jumlah yang seimbang.

Metode *over sampling* ini bekerja dengan cara membuat sampel data baru secara acak sebanyak data pada kelas mayoritas. Penelitian ini memiliki jumlah kelas yang sedikit terdapat pada kelas terlambat dan jumlah kelas mayoritas adalah kelas tepat waktu, sehingga *over sampling* bekerja dengan cara membuat data baru pada data kelas terlambat sebanyak kelas tepat waktu.

Tabel 5. Balancing Data

Jenis Data	Tepat Waktu	Terlambat	Total
Data Asli	296	34	330
Data setelah oversampling	295	295	590

Dalam melakukan proses klasifikasi, data dibagi terlebih dulu yakni *data training* dan *data testing*. *Data training* atau data latih dipakai untuk membuat model klasifikasi, sedangkan *data testing* dipakai untuk menguji apakah model data training sebelumnya baik dalam menganalisis *data testing*. Pembagian data ini berguna untuk mengetahui ketepatan metode dalam melakukan klasifikasi data.

Tabel 6. Pembagian data training dan data testing.

Keterangan	Data Training	Data Testing	Total
Jumlah	413	177	590
Prosentase	70%	30%	100%

Dari Tabel 6, total data dalam penelitian ini adalah 590 data, dengan komposisi 70% data training sebanyak 413 data, dan 30% data testing sebanyak 177 data. Pembagian data dilakukan secara acak dengan bantuan software Spyder Python.

4.3 Pemodelan dan Klasifikasi Data

Pada tahap ini dilakukan pengujian klasifikasi kelulusan mahasiswa menggunakan software Spyder untuk membuat pemodelan dan akurasi dari algoritma random forest dan juga gradient boosting.

a. Algoritma Klasifikasi Random Forest

Training model memakai fungsi RandomForestClassifier dari library sklearn.ensemble. Training data ini berguna untuk pemodelan algoritma agar hasil klasifikasi lebih akurat.

```
# Train the model using the training sets X_train dan y_train
clf = RandomForestClassifier(n_estimators=100, criterion="gini",
                             max_depth=3, min_samples_split=2, bootstrap=True,
                             min_samples_leaf=1, min_impurity_decrease=0,
                             random_state=42)
clf.fit(X_train,y_train)
```

Gambar 2. Kode program klasifikasi random forest

Pada fungsi RandomForestClassifier di program menggunakan python, ada kriteria yang umumnya digunakan, yakni :

- n_estimators : jumlah tree yang dibentuk, diset 100 buah.
- criteration : kriteria split node, diset gini.
- max_depth : kedalaman tree, diset 3 tingkatan node.
- min_samples_split : percabangan tree, minimal 2.
- boostrap : data boostrap setiap tree, diset True.
- min_samples_leaf : jumlah leaf minimal, diset 1 buah.
- min_impurity_decrease : pengurangan value gini, diset 0.

b. Algoritma Klasifikasi Gradient Boosting

Pada tahap ini, training model memakai fungsi GradientBoostingClassifier dari library sklearn.ensemble. Training data ini berguna untuk pemodelan algoritma agar hasil klasifikasi lebih akurat.

```
# (((((((((((((((((((( PEMODELAN GRADIENT BOOSTING CLASSIFIER )))))))))))))))))))

# import Gradient Boosting Classification Model
from sklearn.ensemble import GradientBoostingClassifier

# Train the model using the training sets X_train dan y_train
gb_clf = GradientBoostingClassifier(loss="log_loss", learning_rate=0.01,
                                    max_depth=3, min_samples_split=2,
                                    n_estimators=100, min_samples_leaf=1,
                                    min_impurity_decrease=0, random_state=42)
gb_clf.fit(X_train, y_train)
```

Gambar 3. Kode program klasifikasi gradient boosting.

Pada fungsi GradientBoostingClassifier di program menggunakan python, ada kriteria yang umumnya digunakan, yakni :

- loss = “log loss” : fungsi loss pada gradient boosting.
- learning_rate : hyper parameter, diset 0.01
- n_estimators : jumlah tree yang dibentuk, diset 100 buah.
- criteration : kriteria split node, diset gini.
- max_depth : kedalaman tree, diset 3 tingkatan node.
- min_samples_split : percabangan tree, minimal 2.
- min_samples_leaf : jumlah leaf minimal, diset 1 buah.
- min_impurity_decrease : pengurangan value gini, diset 0.

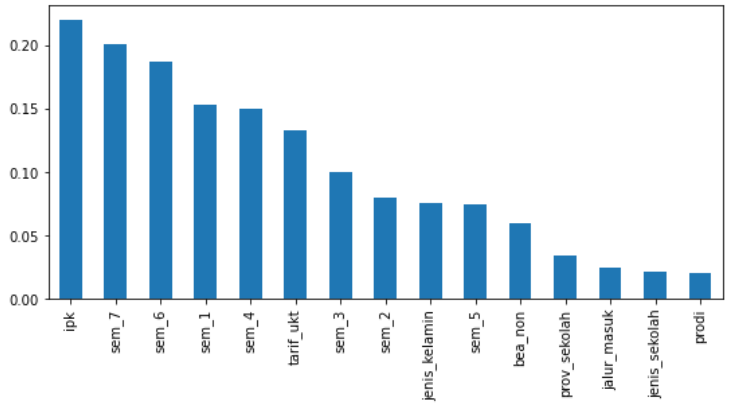
Model yang sudah terbentuk dari perhitungan algoritma menggunakan data mining kemudian dites dengan data testing.

```
# Model sudah terbentuk kemudian digunakan untuk menguji data testing X_test
predictions = gb_clf.predict(X_test)
y_prediksi = pd.Series(predictions)

print('\nPREDIKSI dari X_test')
print(y_prediksi.value_counts())

print('\nDATA ASLI dari y_test')
print(y_test.value_counts())
```

Gambar 4. Kode program jumlah data testing random forest dan gradient boosting

4.3 Pemeringkatan Atribut

Pemodelan Random Forest dan Gradient Boosting dan pembagian data yang dibentuk sebelumnya berguna untuk mengetahui atribut/ hubungan variabel dan tingkatannya. Dilakukan pemeringkatan atribut guna mengetahui atribut mana saja yang paling berpengaruh dalam proses klasifikasi data.



Gambar 5. Grafik pemeringkatan atribut

Pemeringkatan atribut untuk lebih jelasnya ada pada Tabel 7 dibawah ini.

Tabel 7. Pemeringkatan atribut

Nama Atribut	Value	Ranking	Prosentase
IPK	0,219982	1	14%
IP Semester 7	0,200745	2	13%
IP Semester 6	0,186850	3	12%
IP Semester 1	0,152998	4	10%
IP Semester 4	0,149607	5	10%
Tarif UKT	0,132990	6	9%
IP Semester 3	0,099506	7	6%
IP Semester 2	0,079724	8	5%
Jenis Kelamin	0,075670	9	5%
IP Semester 5	0,074368	10	5%
Beasiswa/ Non	0,059329	11	4%
Provinsi Sekolah	0,034249	12	2%
Jalur Masuk	0,024570	13	2%
Jenis Sekolah	0,021203	14	1%
Program Studi	0,019898	15	1%

Perhitungan akurasi model ini menggunakan fungsi metrics.accuracy_score dengan library sklearn, untuk mencari prosentasi keberhasilkan model dalam mengkalkulasi klasifikasi data testing. Setelahnya, Confussion Matrix ditampilkan dengan fungsi confusion_matrix. Algoritma Random Forest dan Gradient Boosting menggunakan kode program confussion matrix yang sama.

```
# ------------------ PERHITUNGAN AKURASI DAN CONFUSION MATRIX ----------------

# Tampilkan akurasi dari perbedaan antara target class data asli dan prediksi
from sklearn.metrics import accuracy_score
print("\nAkurasinya adalah : ", accuracy_score(y_test, y_prediksi))

# import fungsi confussion matrix
from sklearn.metrics import confusion_matrix
conf_mat = confusion_matrix(y_test, y_prediksi)
print("\nConfusion Matrix : ")
print(conf_mat)

# import fungsi classification report
from sklearn.metrics import classification_report
print(classification_report(y_test, y_prediksi, digits=4))
```

Gambar 6. Kode program confussion matrix.

Tabel 8. Pengujian hasil RF dengan Confussion Matrix

Lama Studi	True 0 (Lulus Tepat Waktu)	True 1 (Lulus Terlambat)
Pred 0 (Lulus Tepat Waktu)	74 (TP)	13 (FN)
Pred 1 (Lulus Terlambat)	14 (FP)	76 (TN)

Tabel 9. Pengujian hasil GB dengan Confussion Matrix

Lama Studi	True 0 (Lulus Tepat Waktu)	True 1 (Lulus Terlambat)
Pred 0 (Lulus Tepat Waktu)	74 (TP)	23 (FN)
Pred 1 (Lulus Terlambat)	9 (FP)	71 (TN)

4.4 Data Tambahan

Data tambahan pada perhitungan dan pemodelan sebelumnya adalah membuktikan bahwa pemodelan yang dibentuk tidak mengalami overfitting, dan juga menampilkan plot tree pada masing-masing algoritma.

Overfitting pada pemodelan sering terjadi dan membuat hasil tidak akurat, maka dari itu perlu pembuktian agar pemodelan yang dibentuk sebelumnya adalah Good Fit.

Tabel 10. Pengecekan overfitting dengan akurasi

Data	Pengecekan Overfitting	
	Random Forest	Gradient Boosting
Training Accuracy	87,40%	86,68%
Validation Accuracy	84,74%	81,92%

Overfitting terjadi ketika ada perbedaan akurasi yang sangat besar antara data testing dan data training. Pada Tabel 10, akurasi data training dan testing pada Random Forest adalah 87,40% dan 84,74%, yang tidak memiliki perbedaan jauh diantaranya. Kemudian pada Gradient Boosting, akurasi training berkisar 86,68% dan testing sebesar 81,92%. Bisa disimpulkan bahwa pemodelan yang dibentuk di sub-bab sebelumnya merupakan pemodelan yang bagus dan tidak mengalami overfitting.

4.5 Pengujian Data

Pengujian data hasil klasifikasi dari klasifikasi Random Forest dan Gradient Boosting dilakukan dengan K-Fold Cross Validation dan atribut yang digunakan. Dilakukan uji variasi atribut dengan 2 atribut hingga 15 atribut menggunakan K-Fold Cross Validation dengan menggunakan 40 tree dan 100 tree. Pemilihan atribut disesuaikan dengan pemeringkatan atribut pada Tabel 11.

Tabel 11. Uji Variasi Atribut

Jumlah Atribut	Jumlah Tree	Nama Atribut
2	40 dan 100	IPK, IPS 7
5	40 dan 100	IPK, IPS 7, IPS 6, IPS 1, IPS 4
10	40 dan 100	IPK, IPS 7, IPS 6, IPS 1, IPS 4, Tarif UKT, IPS 3, IPS 2, Jenis Kelamin, IPS 5
15	40 dan 100	IPK, IPS 7, IPS 6, IPS 1, IPS 4, Tarif UKT, IPS 3, IPS 2, Jenis Kelamin, IPS 5, Beasiswa/ Non, Provinsi Sekolah, Jalur Masuk, Jenis Sekolah, Program Studi

Tabel 12. Hasil Akurasi Pengujian 2 Atribut

Percobaan	Fold	Akurasi
Random Forest (40 Pohon)	3	74,68%
Random Forest (100 Pohon)	3	68,55%
Gradient Boosting (40 Pohon)	3	66,67%
Gradient Boosting (100 Pohon)	3	77,96%
Random Forest (40 Pohon)	5	67,15%
Random Forest (100 Pohon)	5	74,61%
Gradient Boosting (40 Pohon)	5	66,23%
Gradient Boosting (100 Pohon)	5	74,68%
Random Forest (40 Pohon)	10	70,65%
Random Forest (100 Pohon)	10	70,62%
Gradient Boosting (40 Pohon)	10	66,73%
Gradient Boosting (100 Pohon)	10	71,83%

Tabel 13. Hasil Pengujian Seluruh Atribut

Jumlah Atribut	Percobaan	Fold	Akurasi
2	Random Forest (40 Pohon)	3	74,68%
	Gradient Boosting (100 Pohon)	3	77,96%
5	Random Forest (40 Pohon)	3	77,40%
	Gradient Boosting (100 Pohon)	10	81,40%
10	Gradient Boosting (40 Pohon)	3	78,53%
	Random Forest (40 Pohon)	10	82,48%
15	Gradient Boosting (100 Pohon)	3	79,66%
	Random Forest (40 Pohon)	10	82,64%

Pada Tabel 13 di tiap jumlah atribut yang diuji, terdapat perbedaan pada fold, dikarenakan sudah dilakukan pemilihan akurasi yang tertinggi dari setiap foldnya yakni 3, 5 dan 10 Fold, seperti pada Tabel 12 sebagai contoh.

Dari seluruh percobaan yang sudah dilakukan pada sub bab sebelumnya, telah diketahui bawasannya jumlah k dalam fold mempengaruhi hasil akurasi, begitu juga dengan jumlah tree sebanyak n. Perbandingan akurasi terbaik Random Forest dan Gradient Boosting dari berbagai variasi pada tahap sebelumnya ditunjukkan pada Tabel 14.

Tabel 14. Hasil Optimasi Semua Variasi

Jumlah Atribut	Akurasi Random Forest	Akurasi Gradient Boosting
2	74,68%	77,96%
5	77,40%	81,40%
10	82,48%	78,53%
15	82,64%	79,66%

Berdasarkan Tabel 14 bisa dilihat bahwa metode algoritma klasifikasi menggunakan Random Forest dan Gradient Boosting, dimana metode Random Forest mendapatkan nilai akurasi lebih tinggi yakni 82,64% dibandingkan dengan metode Gradient Boosting yaitu sebesar 79,66% yang artinya bahwa klasifikasi ketepatan lulus mahasiswa dapat diklasifikasikan dengan baik.

5. Penutup

5.1 Kesimpulan

Berdasarkan hasil analisis yang telah dilakukan pada bab sebelumnya dapat ditarik kesimpulan sebagai berikut
.

1. Penelitian ini melakukan klasifikasi terhadap kelulusan mahasiswa S1 FIK UPN Veteran Jakarta menggunakan Random Forest dan Gradient Boosting. Pada metode Random Forest, tingkat akurasi tertinggi sebesar 82,64% menggunakan nilai k=10, dengan 40 pohon dan menggunakan 15 atribut. Sedangkan pada Gradient Boosting, mendapatkan akurasi sebesar 79,66% dari nilai k=3 dan juga menggunakan 100 pohon dengan 15 atribut yang sama.

2. Untuk kedua pemodelan dari seluruh percobaan yang sudah dilakukan, telah diketahui bawasannya jumlah k dalam fold mempengaruhi hasil akurasi, begitu juga dengan jumlah tree sebanyak n (jumlah n pada kasus ini antara 40 dan 100 tree). Kemudian algoritma Random Forest membuktikan bahwa akurasi algoritma ini sedikit lebih tinggi dibandingkan dengan Algoritma Gradient Boosting.

5.2 Saran

1. Untuk penelitian berikutnya, dapat dilakukan perbandingan antara Random Forest dengan algoritma klasifikasi lainnya untuk membuktikan akurasi terbaik diantara keduanya.

2. Mengembangkan hasil klasifikasi dengan membuat GUI Aplikasi yang berguna dan mempermudah dosen ataupun fakultas untuk melihat mahasiswa yang memiliki potensi lulus tidak tepat waktu.

3. Menambahkan lebih banyak atribut juga dapat mempengaruhi ketepatan lulus mahasiswa dan juga bisa dengan memperbanyak variasi dari atributnya.

6. Daftar Pustaka

[1] Achmad Bisri and Rinna Rachmatika (2019) "Integrasi Gradient Boosted Trees dengan SMOTE dan Bagging untuk Deteksi Kelulusan Mahasiswa", Jurnal Nasional Teknik Elektro dan Teknologi Informasi, 8(4), pp. 309-314. doi: 10.22146/jnteti.v8i4.2554.

[2] Agusta, Erwinsyah Rico. (2021). Komparasi Metode Naïve Bayes Dan Random Forest Untuk Memprediksi Ketepatan Waktu Lulus Mahasiswa (Studi Kasus: Mahasiswa Fakultas Sains Dan Teknologi Universitas Sanata Dharma). (Skripsi, Universitas Sanata Dharma).

[3] Agwil, W., Fransiska, H., & Hidayati, N. (2020). Analisis Ketepatan Waktu Lulus Mahasiswa Dengan Menggunakan Bagging CART. FIBONACCI: Jurnal Pendidikan Matematika dan Matematika, 6(2), 155-166.

[4] Brownlee, Jason. 2020. 4 Types of Classification Tasks in Machine Learning. https://machinelearningmastery.com/types-of-classification-in-machine-learning/ (diakses tanggal 26 Februari 2022).

[5] Hayati, I., Marzal, J., & Saputra, E. (2021). Klasifikasi Mahasiswa Berpotensi Drop Out Menggunakan Algoritma Decision Tree C4. 5 Dan Naive Bayes Di Universitas Jambi (Doctoral dissertation, Universitas Jambi).

[6] HIMMAWAN, M., & Agus Ulinuha, S. T. (2021). Perancangan Sistem Analisis Kelulusan Mahasiswa Dalam Implementasi Penerapan Data Mining Pada Bidang Pendidikan (Doctoral dissertation, Universitas Muhammadiyah Surakarta).

[7] Inoxoft. (2021, Jan). Gradient Boosting Classifier. Medium. https://medium.com/geekculture/gradient-boosting-classifier-f7a6834979d8

[8] Kurniawati, Galuh N. 2021. Algoritma Machine Learning yang Harus Kamu Pelajari di Tahun 2021. https://www.dqlab.id/algoritma-machine-learning-yang-perlu-dipelajari (diakses tanggal 20 Februari 2022).

[9] Mashlahah, S. (2013). Prediksi kelulusan mahasiswa menggunakan metode decision tree dengan penerapan algoritma C4. 5 (Doctoral dissertation, Universitas Islam Negeri Maulana Malik Ibrahim).

[10] M. Utari, B. Warsito and R. Kusumaningrum, "Implementation of Data Mining for Drop-Out Prediction using Random Forest Method," 2020 8th International Conference on Information and Communication Technology (ICoICT), 2020, pp. 1-5, doi: 10.1109/ICoICT49345.2020.9166276

[11] Nelson, Dan. (2022, Jul). Gradient Boosting Classifiers in Python with Scikit-Learn. StackAbuse. https://stackabuse.com/gradient-boosting-classifiers-in-python-with-scikit-learn/

[12] Plaosan, van Suprapto. 2022. Learning Box : Algoritma Random Forest. http://learningbox.coffeecup.com/05_2_randomforest.html (diakses tanggal 28 Februari 2022).

[13] Rachmi, Adhelia Nurfira. (2020). Implementasi Metode Random Forest Dan Xgboost Pada Klasifikasi Customer Churn. (Skripsi, Universitas Islam Indonesia).

[14] Sinaga, Artha Dian. (2020). Prediksi Kelulusan Mahasiswa Fakultas Sains Dan Teknologi Universitas Sanata Dharma Menggunakan Metode Klasifikasi Naïve Bayes. (Skripsi, Universitas Sanata Dharma).

[15] Tambunan, R. H. (2020). Analisis Prediksi Kelulusan Mahasiswa Tepat Waktu Berdasarkan Kinerja Akademis Mahasiswa Menggunakan Algoritma Naïve Bayes dengan Implementasi Data Mining Studi Kasus: Departemen Teknik Industri USU.

[16] Yahya, S. A. (2018). Klasifikasi Ketepatan Lama Studi Mahasiswa Menggunakan Metode Support Vector Machine Dan Random Forest (Studi Kasus: Data Lama Studi Alumni Universitas Islam Indonesia Tahun Kelulusan 2000-2017).