ISSN: 2477-5126
e-ISSN: 2548-9356

Perbandingan Metode Klasifikasi serta Analisis Faktor Akademis Pola Kelulusan Mahasiswa di Perguruan Tinggi

Rani Aprillya Putri[1], Nenden Siti Fatonah[2]
[1,2]Jurusan Teknik Informatika, Fakultas Ilmu Komputer, Universitas Mercu Buana, Jakarta
[1,2]Jl. Raya, Meruya Sel., Kec. Kembangan, Jakarta, Daerah Khusus Ibukota Jakarta, 11650, Indonesia
email: [1]raniaprillya89@gmail.com, [2]nendenfatonah@gmail.com

***Abstrak*—Perkembangan Teknologi informasi di berbagai bidang diikuti dengan berkembangnya data. Pangkalan data yang menyimpan data pengelolaan pelaksanaan pendidikan tinggi dari seluruh perguruan tinggi yang terintegrasi secara nasional yaitu PDDIKTI. Data Mining merupakan proses penggalian data dari kumpulan database yang berjumlah besar yang digunakan untuk mendapatkan pengetahuan berupa informasi penting dan bermanfaat. Dalam penelitian ini penulis menerapkan penggunaan data mining untuk mengetahui klasifikasi pola akademik kelulusan mahasiswa. Pada penelitian ini penulis akan membandingkan metode klasifikasi *Decision Tree, Ensemble Learning (Bagging & Boosting)*, mengetahui faktor yang paling berpengaruh terhadap kelulusan serta Analisis data pola akademik mahasiswa. Hasil pengujian klasifikasi data kelulusan mahasiswa dengan hasil akurasi terbaik adalah metode algoritma *Ensemble Learning* Bagging atau *Random Forest* dengan melakukan *cross validation* dan *hyperparameter tuning (Grid Search CV)* dengan akurasi 96.1%. Penggunaan *cross validation* dan *hyperparameter tuning* terbukti dapat mempengaruhi dan mengoptimalkan akurasi learning. Faktor yang paling mempengaruhi pola kelulusan mahasiswa adalah jumlah SKS semester 5, Angkatan, Indeks Prestasi, Program Kelas dan Total Cuti.**

Keywords : Machine Learning, Ensemble Learning, Klasifikasi, PDDikti.

Abstract— The development of information technology in various fields is followed by the development of data. The database that stores data on the management of the implementation of higher education from all nationally integrated universities is PDDIKTI. Data mining is the process of extracting data from large databases that are used to find knowledge in the form of important and useful information. In this study applying data mining to determine the classification of student graduation academic patterns. In this study, the authors will compare the Decision Tree classification method, Ensemble Learning (Bagging & Boosting), find out the most influential factors on graduation and analyze student academic pattern data. The results of testing the classification of student graduation data with the best accuracy results are the Ensemble Learning Bagging or Random Forest algorithm method by doing cross validation and hyperparameter tuning (Grid Search CV) with an accuracy of 96.1%. The use of cross validation and hyperparameter tuning is proven to influence and optimize learning accuracy. The factors that most influence the student's graduation pattern are the number of credits for semester 5, Class, Achievement Index, Class Program and Total Leave.

Keywords: Machine Learning, Ensemble Learning, Classification, PDDikti.

I. PENDAHULUAN

Pangkalan data yang menyimpan data pengelolaan pelaksanaan pendidikan tinggi dari seluruh perguruan tinggi yang terintegrasi secara nasional yaitu PDDikti (Pangkalan Data Pendidikan Tinggi). Data yang disimpan di PDDikti adalah data yang akurat dan nyata, karena proses pelaporan data akademik dilakukan secara rutin setiap periodik[1]. Merugikan jika data yang berlimpah tersimpan di PDDikti tidak digunakan untuk mendapatkan informasi yang bermanfaat, misalnya untuk mengetahui kinerja dosen, penilaian akreditasi, membaca pola kelulusan mahasiswa, dan sebagainya. Perguruan Tinggi bersaing untuk menjadi yang terbaik serta dapat memberikan kualitas pengajaran yang efektif, persaingan yang kompetitif antara perguruan tinggi saat ini salah satu nya dilihat dari akreditasi. Akreditasi menjadi salah satu pengukuran kualitas dari perguruan tinggi di Indonesia yang dilakukan oleh Badan Akreditasi Nasional Perguruan Tinggi atau BAN PT (Instrument BAN PT, 2011), kualitas perguruan tinggi ini memiliki 7 standar pengukuran utama, diantaranya adalah ada Mahasiswa dan Lulusan[2].

Untuk dapat mencetak lulusan mahasiswa yang berkualitas dapat dilakukan dengan melakukan monitoring nilai akademik mahasiswa, meningkatkan kualitas dosen dan tenaga pendidikan dan memastikan mahasiswa dapat lulus tepat waktu. Untuk penilaian lama studi mahasiswa, jika lama studi mahasiswa suatu program studi memiliki kecenderungan kelulusan yang tepat waktu (4 tahun), maka penilaian kriteria lama studi untuk status akreditasi program studi akan baik begitu juga sebaliknya, ini yang menjadi salah satu standar penilaian perguruan tinggi dalam mendidik mahasiswa dan untuk menunjang kegiatan. Pengambilan keputusan tidak cukup hanya mengandalkan tindakan kuratif saja[2], diperlukan Tindakan preventif dengan melakukan suatu analisis data dan klasifikasi atau prediksi untuk dapat mengetahui pola data atau proses pengubahan data menjadi informasi. Informasi yang ada akan menjadi prototipe dengan

*) **penulis korespondensi**: Rani Aprillya Putri
Email: raniaprillya@gmail.com

diambil pola data agar dapat memberikan pengetahuan. Hal ini memunculkan adanya cabang ilmu baru untuk menyelesaikan permasalahan informasi serta pola yang penting atau menarik dari sejumlah data yang besar, yaitu data mining. Data Mining merupakan proses penggalian data dari kumpulan database yang berjumlah besar yang digunakan untuk mendapatkan pengetahuan berupa informasi yang penting serta bermanfaat[3]. Jadi, dengan penerapan data mining dapat mengetahui klasifikasi pola akademik kelulusan mahasiswa yang menjadi pengaruh besar dalam proses akreditasi.

Pada penelitian ini penulis akan membandingkan metode klasifikasi Machine Learning dan Ensemble Learning, mengetahui faktor yang paling berpengaruh signifikan terhadap kelulusan serta Analisis data pola akademik mahasiswa. Adapun metode algoritma yang digunakan untuk pengujian adalah Decision Tree, Ensemble Learning Voting, Ensemble Learning Bagging (Random Forest) dan Ensemble Learning Boosting (SGB).

II. PENELITIAN YANG TERKAIT

Pemilihan penggunaan metode algoritma serta data yang digunakan dalam penelitian ini adalah didasarkan dari penelitian-penelitian sebelumnya yang membahas penelitian dengan topik yang sama, yaitu pada penelitian pertama yang membahas perbandingan tingkat akurasi dengan topik memprediksi lama studi mahasiswa menggunakan metode K-Nearest Neighbor dengan nilai akurasi yaitu 53,08% dan prediksi Decision Tree 60,38%. Dan saran untuk penelitian selanjutnya adalah untuk menambah jumlah data dan memberikan variabel atau kriteria eksternal[4]. Dari penelitian ini didapatkan SVM memiliki akurasi yang paling bagus diantara keempat metode algoritma yang digunakan yaitu sebesar 80%, KNN 64%, Decision Tree 65% dan Naïve Bayes 77%[5]. Pada penelitian selanjutnya Peneliti berhasil mendapatkan nilai akurasi single classifier (naïve bayes) yaitu 77,4% dan nilai ensemble learning 96,8%. Dan untuk penelitian selanjutnya pemilihan fitur penting diharapkan bisa lebih dikembangkan sehingga didapati hasil klasifikasi yang lebih maksimal[6]. Pada penelitian yang lainnya dengan topik mencari atribut yang paling memberikan faktor signifikan terhadap performa mahasiswa, dalam penelitian ini melakukan prediksi menggunakan 2 algoritma Machine Learning dan mengetahui atribut yang paling mempengaruhi performa akademik atau atribut yang paling memberikan pengaruh signifikan. Umur, Pekerjaan, Jenis Kelamin, Kelas, dan Status tidak memiliki efek yang signifikan terhadap kesuksesan akademik, sedangkan IPK, SKS, Catatan Penting, Pekerjaan Ayah, and Konsumsi Makanan memberikan pengaruh yang signifikan, dan untuk algoritma dengan performa terbaik yaitu J48[7]. Dari jurnal-jurnal referensi diatas penulis tertarik untuk menggunakan Metode pengujian Algoritma Machine Learning Decision Tree, serta melakukan penambahan dengan pengujian menggunakan Ensemble Learning (Bangging (Random Forest) dan Boosting (SGB)) serta Analisis data pola akademik mahasiswa untuk mengetahui faktor yang paling berpengaruh signifikan terhadap kelulusan.

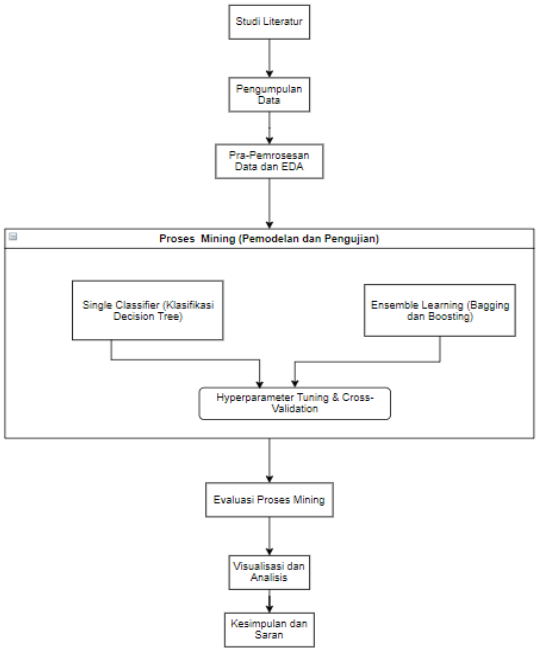
III. METODE PENELITIAN

Pada penelitian ini menggunakan data yang tersedia di PDDikti dimana PDDikti menjadi salah satu instrumen untuk melaksanakan penjaminan mutu (Kementrian Riset, 2017). Dalam pasal 56 ayat 2 UU No. 12 Tahun 2012 tentang data Pendidikan Tinggi sebagai halnya dimaksud pada ayat (1) berfungsi sebagai sumber informasi bagi Lembaga akreditasi, Pemerintah dan Masyarakat[1].

Penelitian ini merupakan penelitian kuantitatif karena penulis akan melakukan penelitian eksperimen dengan melakukan pengujian observasi antar variabel dan penelitian akan melakukan analisis statistik dan klasifikasi kelulusan dengan memperhatikan tingkat akurasi agar mendapat hasil klasifikasi dengan akurasi terbaik untuk menganalisa pola dan faktor akademis kelulusan mahasiswa di perguruan tinggi. Adapun tahapan alur dalam penelitian ini adalah sebagai berikut:

A. Alur Penelitian dengan Diagram Alir

Alur penelitian ini dilakukan dimulai dari studi literatur kemudian pengumpulan data, pra-pemrosesan dan EDA, Proses Mining (Pengujian dan Pemodelan), Evaluasi proses mining, Visualisasi dan Analisis dan terakhir kesimpulan dan saran. Adapun diagram alir penelitian ini digambarkan pada gambar 1.



Gbr. 1 diagram alir alur penelitian.

B. Rincian Alur Penelitian

Alur kerja Penelitian ini ditunjukkan pada Gambar 1, penjelasan setiap tahapannya sebagai berikut:

1. Studi Literatur

   Pada tahap ini penulis melakukan studi literatur dari berbagai referensi yang berkaitan dengan penelitian yang dilakukan. Studi literatur dalam penelitian ini penulis mempelajari tahapan pengumpulan atau penarikan data, konsep data mining, implementasi algoritma klasifikasi dan konsep penilaian akreditasi perguruan tinggi.
2. Pengumpulan Data

   Berdasarkan studi literatur data yang diperlukan untuk melakukan penelitian ini adalah data akademik

mahasiswa angkatan 2016-2018, dimana data mahasiswa Angkatan 2016-2017 digunakan untuk melakukan pelatihan dan validasi akurasi pemodelan serta data mahasiswa Angkatan 2018 digunakan sebagai data latih pengklasifikasian. Adapun attribute/fitur atau variabel independen yang digunakan dalam penelitian ini adalah Jurusan, Program Kelas, Jenis Kelamin, Umur, IP semester 3, IP semester 4, IP semester 5, Total SKS Sem 5, Total Cuti, Kota Lahir dan untuk label/target atau variabel dependen adalah lama studi.

3. Pra-Pemrosesan dan *Exploratory Data Analysis(EDA)*

   Tahap pengolahan data awal (pre-processing) adalah tahap pertama dalam melakukan proses data mining yang terdiri dari pembersihan dan transformasi data. Pada pra pemrosesan ini bertujuan untuk transformasi data sehingga data yang dipakai untuk pengaplikasian data mining lebih mudah diinterpretasikan untuk dianalisis. Selain itu, data yang digunakan juga dapat sesuai dengan penggunaan atau pengaplikasian yang dibangun sehingga hasil yang dikeluarkan juga sesuai dan optimal[5]. Setelah pra pemrosesan dilakukan EDA untuk mengetahui karakteristik data latih yang akan digunakan.
4. Proses Mining (Pemodelan dan Pengujian)
   a. Pemodelan, *Cross Validation & Hyperparameter Tuning.*

      Setelah pra pemrosesan data dan EDA dilakukan pemodelan machine learning dan ensemble learning untuk melakukan klasifikasi pola kelulusan mahasiswa. Kemudian selanjutnya dilakukan cross validation serta hyperparameter tuning untuk mendapatkan hasil pengujian yang lebih optimal.

      i. *Machine Learning*

         Machine learning merupakan cabang atau turunan pengaplikasian dari kecerdasan buatan. Cabang ilmu ini memfokuskan pada pembuatan sistem atau algoritma yang terus-menerus belajar dari data dan meningkatkan akurasinya. Dalam aplikasi *machine learning*, algoritma atau urutan proses statistik dilatih untuk dapat menemukan pola dan fitur tertentu dalam sejumlah data yang besar[8].

         - *Decision Tree*

           Decision Tree (DT) merupakan model prediksi atau klasifikasi yang menerapkan struktur pohon atau struktur berhirarki. Konsep dari *decision tree* ini adalah dengan mengubah data menjadi decision tree (pohon keputusan) dan cabang aturan-aturan keputusan[9]. Setiap pohon memiliki cabang yang terdapat node dan setiap node menggambarkan fitur dalam kategori yang akan diklasifikasikan serta setiap subset mendefinisikan nilai yang dapat diambil oleh node[10].

      ii. *Ensemble Learning*

         Ensemble learning merupakan metode yang menggabungkan beberapa algoritma dengan aturan tertentu sedemikian rupa sehingga dapat mengkombinasikan kekuatan beberapa algoritma sehingga dapat memiliki performa generalisasi atau secara penyamarataan yang lebih baik daripada klasifikasi tunggal[11].

         - *Bagging*

           Bagging atau *bootstrap aggregating,* adalah metode ensemble yang melibatkan pelatihan algoritma yang sama berkali-kali dengan menggunakan subset berbeda yang diambil sampelnya dari data pelatihan. Prediksi keluaran akhir didapatkan dengan dirata-ratakan di seluruh prediksi semua sub-model. Bagging umumnya meningkatkan akurasi klasifikasi dengan mengurangi varians dari kesalahan klasifikasi[9]. Random Forest (RF) adalah pengklasifikasi ensemble, yang digunakan untuk klasifikasi dan analisis regresi. RF bekerja dengan membuat berbagai Decision Tree dalam fase pelatihan dan mengeluarkan label kelas yang memiliki suara mayoritas. RF mencapai akurasi klasifikasi yang tinggi dan dapat menangani outlier dan noise dalam data[12].

         - *Boosting*

           Dalam boosting, beberapa model dilatih secara berurutan, dan setiap model belajar dari kesalahan model sebelumnya. Boosting akan memberikan bobot ke instance pelatihan, dan nilai bobot ini diubah tergantung pada seberapa baik instance pelatihan terkait dipelajari oleh classifier; bobot untuk instance yang salah diklasifikasikan akan ditingkatkan[9]. Gradient boosting adalah algoritma pembelajaran ensemble yang dikombinasikan dengan boosting dan decision tree. Inti dari algoritma SGB adalah untuk meminimalkan fungsi loss antara fungsi klasifikasi dan fungsi nyata[13].

      iii. *Cross Validation*

         Cross Validation digunakan untuk mengevaluasi validitas prediktif dari persamaan regresi linier yang digunakan untuk meramalkan kriteria kinerja dari skor pada serangkaian tes[14].

      iv. *Hyperparameter Tuning*

         Hyperparameter adalah variabel pengoptimal yang dijalankan selama fase pelatihan untuk mendapatkan nilai rata-rata yang dioptimalkan setelah beberapa proses coba-coba. Untuk mengatasi kendala

overfitting dengan pencarian Grid Search. Model GridSearchCV yang diambil dari Scikit-learn digunakan untuk mendapatkan parameter terbaik. Fokusnya adalah untuk mendapatkan parameter yang paling optimal[15].

b. Pengujian

Pengujian dilakukan dengan membagi data latih, data validasi dan data uji dimana data latih dan data validasi menggunakan data akademik mahasiswa Angkatan 2016 dan 2017 yang merupakan alumni yang sudah menyelesaikan studinya sehingga terdapat label lama studi sehingga melatih pemodelan dengan melihat pola data latih dan validasi dan data latih menggunakan data Angkatan 2018 yang belum memiliki label lama studi untuk diklasifikasikan menggunakan pemodelan yang paling efektif.

5. Evaluasi Proses Mining

Tahap ini merupakan Evaluasi algoritma yang digunakan, pada tahap ini menghasilkan hasil Evaluasi berupa nilai akurasi yang digunakan perbandingan efektifitas antara algoritma yang digunakan dan menampilkan *Feature Importance*.

6. Visualisasi dan Analisis

Pada tahap ini dilakukan Analisa dan visualisasi hasil implementasi data mining untuk mengetahui perbandingan algoritma pengklasifikasian pola kelulusan mahasiswa serta hasil klasifikasi prediksi pola kelulusan mahasiswa.

7. Kesimpulan dan Saran

Setelah proses analisis dan visualisasi tahap terakhir dalam penelitian ini adalah menarik kesimpulan dari apa yang sudah penulis lakukan dalam penelitian ini berupa algoritma yang paling efektif dan masukan untuk perguruan tinggi serta saran untuk penelitian selanjutnya.

IV. HASIL DAN PEMBAHASAN

Dari tahapan penelitian diatas didapatkan hasil penelitian ini adalah sebagai berikut :

A. Alur Penelitian dengan Diagram Alir

Penelitian ini menggunakan data PDDIKTI Universitas X tahun akademik 2016-2018 yang ditambahkan dengan data dari PDDIKTI internal dari universitas. Pengumpulan data awal diperoleh menggunakan query SQL pada aplikasi *Navicat* untuk mengambil data yang dibutuhkan yang disimpan dengan format Excel. Adapun data yang diperoleh dari hasil pengkuierian SQL disimpan dalam file 031019Akm.xlsx, 031019Mhs.xlsx dan stts_kelas.xlsx.

TABEL I
DATA YANG TERSIMPAN DALAM FILE AWAL

No	Nama Kolom	Tipe Data	Ket.
1	KDPT	int	Kode Perguruan Tinggi
2	NMPT	text	Nama Perguruan Tinggi
3	KDPST	int	Kode Progran Studi
4	NMPS	text	Nama Program Studi
5	JEN	text	Jenjang
6	SMAW	int	Semester Awal Masuk
7	NIM	int	Nomor Induk Mahasiswa
8	NMMHS	text	Nama Mahasiswa
9	JNSDAFTAR	text	Jenis Daftar
10	SMT	int	Semester
11	STATMHS	text	Status Mahasiswa (Aktif, Cuti, Double Degree, Keluar, Nonaktif)
12	IPS	int	Indeks Prestasi Semester
13	IPK	int	Indeks Prestasi Kumulatif
14	SKSMST	int	SKS Semester
15	SKSTOTAL	int	SKS Total Semester
16	TMPTLHR	text	Tempat Lahir
17	TGLLHR	date	Tanggal Lahir
18	NIK	int	Nomor Induk Keluarga
19	KEL	text	Jenis Kelamin
20	TGLMSK	date	Tanggal Masuk
21	STPIDMSMHS	text	Status Peserta Didik Mahasiswa Baru
22	SKSDIAKUI	int	SKS Diakui
23	KET	int	Keterangan (Lulus, Mengundurkan Diri, Aktif, Dikeluarkan, Wafat)
24	TGLKLR	date	Tanggal Keluar
25	NOIJAZAH	text	Nomor Ijazah
26	SMTLULUS	int	Semester Lulus
27	Program	text	Jenis Kelas (Reguler I dan Reguler II)

B. Pra-pemrosesan dan EDA

Pada tahap ini melakukan pembersihan data pada data 031019Akm.xlsx atau master_akm, 031019Mhs.xlsx atau master_mhs, dan stts_kelas.xlsx atau program_kls. Pada data master_akm mengambil data jenjang S1 dengan jenis daftar peserta didik baru dan status mahasiswa aktif, cuti atau nonaktif, dengan menggunakan data master_akm ini juga dapat menghasilkan total cuti, ip semester 3, ip semester 4, ip semester 5, IPK Semester 5 dan Total SKS tempuh semester 5. Selanjutnya langsung menggabungkan dua dataset yaitu data 031019Mhs.xlsx atau master_mhs dengan stts_kelas.xlsx atau program_kelas, pada proses pembersihan dataset ini dimulai dari hanya mengambil data mahasiswa yang aktif dan lulus, membuat kolom umur, membuat kolom lama studi (dihitung dari perhitungan tanggal keluar atau lama tahun lulus). Setelah pra pemrosesan kedua dataset dilakukan penggabungan antara data master_mhs dengan master_akm yang disimpan dalam dataframe all_data. Pada dataset all_data dilakukan pembersihan kembali sehingga tidak ada data yang kosong(null) dan menyesuaikan tipe data dari masing-masing kolom.

Hasil all_data hasil pra pemrosesan yang berisikan 14695 data dan 15 kolom yang ditampilkan pada tabel 2.

TABEL II
DATA HASIL PRA PEMROSESAN

	NMPS	SMAW	...	...	...	...	UMUR	PROGRA	LAMA STD

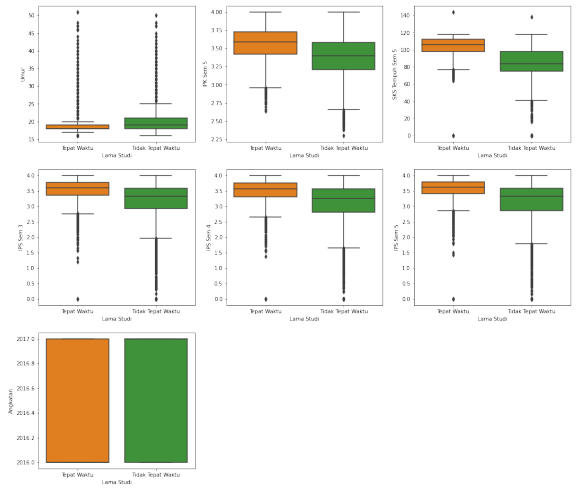
ISSN: 2477-5126
e-ISSN: 2548-9356

								M	
0	Teknik Elektro	2016	…	…	…	…	18	Reg 1	1
1	Teknik Elektro	2016	…	…	…	…	19	Reg 1	0
2	Teknik Elektro	2016	…	…	…	…	19	Reg 1	1
.	.	.	…	…	…	…	.	.	.
.	.	.	…	…	…	…	.	.	.
14694	Desain Komunikasi Visual	2018	…	…	…	…	22	Reg 2	2

Selanjutnya proses EDA, merupakan langkah awal yang penting untuk setiap proses penemuan pengetahuan, di mana peneliti data secara interaktif dapat mengeksplorasi kumpulan data yang tidak dikenal dengan mengeluarkan urutan operasi analisis[14]. EDA dilakukan untuk mengetahui persebaran, karakteristik atau menemukan pola, hubungan, dan anomali untuk menginformasikan analisis selanjutnya. Pada proses ini penulis melakukan pemrosesan EDA pada data latih (data Angkatan 2017 dan 2018).

- Boxplot Data Latih

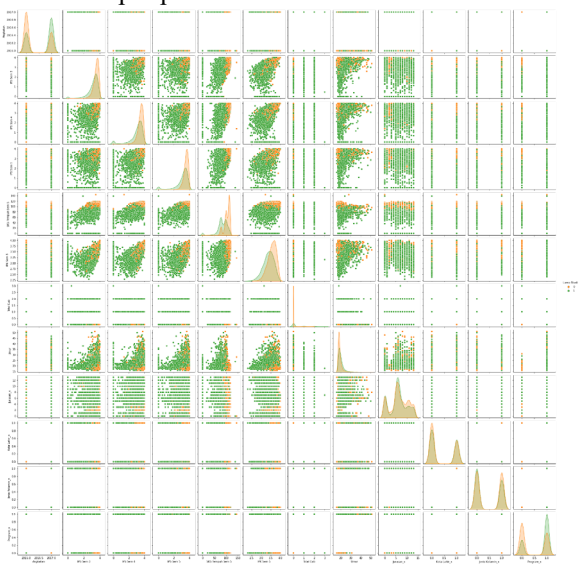
Boxplot adalah grafik yang dapat menampilkan distribusi data atau bagaimana nilai-nilai dalam data yang tersebar, berisikan ringkasan lima perhitungan yaitu minimum, kuartil pertama (Q1), median(Q2), kuartil ketiga (Q3), dan maksimum. Boxplot berguna untuk menampilkan banyak informasi secara ringkas[16]. Persebaran data menggunakan grafik boxplot ditampilkan pada gambar 2. Pada Boxplot gambar 2. Menunjukkan bahwa data yang tersebar sudah cukup baik untuk dilakukan pemrosesan pembelajaran, data ini memiliki nilai ekstrim atau outliers jka dilihat dari boxplot ini, namun dalam pengujian ini data yang outliers tidak dihilangkan. Boxplot ini menggunakan data Lama Studi sebagai parameter perbandingan untuk fitur atau variabel lainnya.



Gbr.2. Boxplot Data Latih

- Pairplot Data Latih

Pairplot memvisualisasikan data untuk menemukan hubungan di antara variabel. Pair Plot pasangan dibangun di atas dua grafik density dan scatter plot[17]. Contohnya dari plot ini kita melihat bahwa SKS Sem 5 dengan umur berkorelasi yang menunjukkan bahwa orang-orang yang mendaftar kuliah pada usia lebih muda <30 memiliki jumlah SKS sem 5 yang lebih optimal. Dari histogram, kita mengetahui bahwa variabel populasi dan gdp sangat condong ke kanan. Gambar 3 menggambarkan visualisasi pairplot data latih.

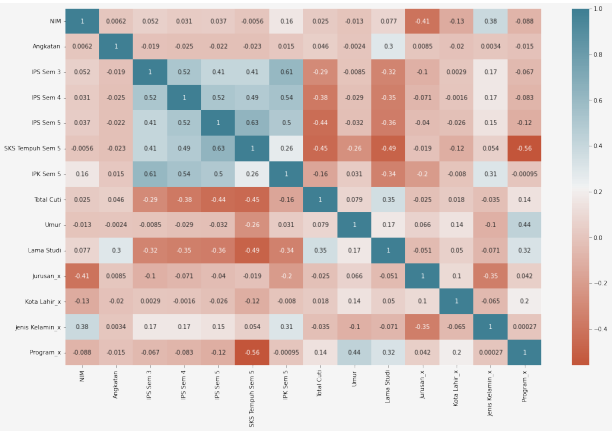


Gbr. 3. Pairplot Data Latih

- Correlation Matrix Data Latih

Correlation Matrix adalah matriks atau grafik yang menampilkan korelasi. Grafik ini baik digunakan dalam fitur yang menunjukkan hubungan linier antara satu sama lain. Pada visualisasi matriks korelasi ini menggunakan visualisasi heatmap dimana data yang berwarna pekat menunjukkan korelasi yang liner baik negatif ataupun positif sedangkan semakin pudar warna berarti semakin tidak berkorelasi antar fitur[18]. Pada dataset ini menunjukkan fitur yang paling berkorelasi dengan lama studi dan 5 fitur yang paling berkorelasi menurut matriks korelasi data latih adalah SKS Tempuh Sem 5, IPS Sem 5, Total Cuti, IPS Sem 4, IPK Sem 5, Program Kelas, dan IPS Sem 5. Gambar 4 menggambarkan Correlation Matrix data latih.

ISSN: 2477-5126
e-ISSN: 2548-9356



Gbr 4. Correlation Matriks Data Latih

C. Proses Mining (Pemodelan dan Pengujian)

Setelah data dilakukan pra pemrosesan, tahap selanjutnya untuk dapat menerapkan metode learning terhadap data adalah dengan *encode* data yang masih *string* atau *object* menjadi tipe data angka dan melakukan binning pada data tempat lahir.

Kemudian data latih dibagi kembali menjadi data latih dan data validasi untuk melatih model pembelajaran algoritma data mining serta evaluasi pemodelan dengan kolom feature_cols = ['Jurusan_x', 'Angkatan', 'IPS Sem 3', 'IPS Sem 4','IPS Sem 4','SKS Tempuh Sem 5','IPK Sem 5', 'Total Cuti','Kota Lahir_x','Jenis Kelamin_x','Umur','Program_x'] sebagai fitur X atau variabel independen dan ['Lama Studi'] sebagai label target Y atau variabel dependen dan membagi persentase data validasi 30% dan data latih 70% dari 9855 baris data

Kemudian setelah membagi data latih dan validasi dilakukan pemodelan dan memasukkan data kedalam model algoritma. Tabel 3 menunjukkan algoritma dan hyperparameter yang digunakan pada pemodelan ini.

TABEL III
PEMODELAN ALGORITMA

NO	NAMA	PARAMETER (TANPA HYPERPARAMETER TUNING)	CROSS VALIDATION	PARAMETER (HYPERPARAMETER TUNING)
Single Learning				
1	Decision Tree	class_weight=None, criterion='gini', max_depth=None, min_samples_leaf=1, min_samples_split=2, random_state=None, splitter='best'	10 Fold	class_weight='balanced', criterion='entropy', max_depth=15, min_samples_leaf=42, min_samples_split=2, random_state=10, splitter='best'
Ensemble Learning				
3	Bagging	criterion='gini', max_depth=None, max_features='auto', min_samples_leaf=1, min_samples_split=2, n_estimators=100, n_jobs=None, random_state=None,	25 Fold	criterion='gini', max_depth=50, max_features='auto', min_samples_leaf=2, min_samples_split=5, n_estimators=400, n_jobs=None, random_state=10,
4	Boosting	n_estimators=100, criterion='friedman_mse', max_depth = 3, min_samples_split=2, min_samples_leaf=1, max_features='auto', min_impurity_decrease=0.0, min_impurity_split=None, random_state=None	24 Fold	n_estimators=50, criterion='friedman_mse', max_depth = 5, min_samples_split=2, min_samples_leaf=1, max_features='None', min_impurity_decrease=0.0, min_impurity_split=None, random_state=10

Setelah dilakukan pemodelan dan optimasi algoritma dengan menggunakan cross validation dan hyperparameter tuning sehingga mendapatkan hyperparameter terbaik dengan menggunakan GridSearch selanjutnya melakukan pengujian menggunakan data validasi untuk mendapatkan akurasi hasil pengujian.

D. Evaluasi Proses Mining

Setelah pemodelan dan pengujian dilakukan evaluasi model algoritma yang digunakan yang menggunakan angka akurasi sebagai parameter pembanding antara model algoritma yang digunakan dalam pengujian. Hasil akurasi data validasi pengujian model algoritma ditunjukkan pada tabel 4.

TABEL 4

ISSN: 2477-5126
e-ISSN: 2548-9356

EVALUASI HASIL AKURASI PENGUJIAN DATA VALIDASI

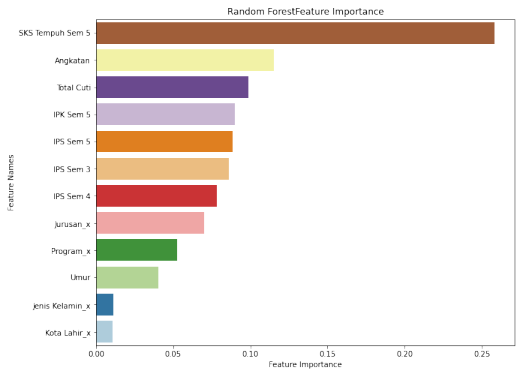
NO	NAMA	SEBELUM HYPERPARAMETER TUNING	CROSS VALIDATION (mean)	SESUDAH HYPERPARAMETER TUNING
Single Learning				
1	Decision Tree	78.4%	78.3%	84.6%
Ensemble Learning				
2	Bagging	83.6%	83.8%	96.1%
4	Boosting	83.5%	84.3%	85.5%

Setelah pengujian menggunakan data validasi dan mendapatkan hasil akurasi, dari hasil akurasi didapatkan model yang paling optimal dalam melakukan klasifikasi pola akademik kelulusan mahasiswa dan kemudian menggunakan model algoritma dengan akurasi terbaik atau bagging Random Forest untuk pengujian data latih atau data Angkatan 2018 yang belum memiliki label target. Hasil pelabelan disimpan dalam dataframe sehingga data pada all_data memiliki label yang sama antara 0 dan 1 atau lulus tepat waktu dan tidak tepat waktu.

- *Feature Importance*

  Setelah mendapatkan hasil akurasi dan evaluasi pengujian, menggunakan model algoritma dengan akurasi paling tinggi dan optimal digunakan untuk mengimplementasikan metode feature importance sebagai tolak ukur besaran pengaruh berbagai fitur data atau variabel yang dilatih kepada performa model. Dalam algoritma bagging atau RandomForest hal ini dilakukan dengan menghitung rata-rata pengurangan reduksi impurity yang dihasilkan oleh suatu fitur di seluruh pohon.

  Pada pengujian ini fitur yang paling mempengaruhi model dalam mengklasifikasikan lama studi adalah SKS Tempuh Semester 5, hal ini sangat memungkinkan karena jumlah sks tempuh mahasiswa pada semester 5 menjadi parameter penting dalam menentukan lama studi mahasiswa untuk dapat memenuhi syarat minimal sks untuk lulus. Selanjutnya diikuti dengan fitur Angkatan yang mungkin Angkatan 2016 dan 2017 memiliki pola akademik yang berbeda begitu juga dengan program studi yaitu regular 1 yang merupakan mahasiswa atau pelajar waktu penuh dengan regular 2 yang merupakah mahasiswa dan juga sambil bekerja atau paruh waktu dan fitur-fitur lainnya juga memberikan pengaruh untuk model mempelajari pola data latih. Plot fitur yang paling mempengaruhi model pengujian untuk dataset pola kelulusan mahasiswa digambarkan pada gambar Gambar 5.



Gbr 5. Feature importance bagging algorithm

E. Visualisasi dan Analisis

1. Korelasi matriks dan Feature Importance method

Pada grafik korelasi matrik menunjukkan variabel yang memiliki korelasi terkuat dengan variabel lama studi adalah SKS Tempuh Sem 5, IPS Sem 5, Total Cuti, IPS Sem 4, IPK Sem 5, Program Kelas, IPS Sem 3, Angkatan, Umur, Jenis Kelamin, Jurusan dan Kota Lahir. Sedangkan untuk grafik hasil feature importance urutan variabel atau fitur yang paling mempengaruhi model algoritma dalam memprediksi data adalah SKS Tempuh Sem 5, Angkatan, Total Cuti, IPK Sem 5, IPS Sem 5, IPS Sem 3, IPS Sem 4, Jurusan, Program Kelas, Umur, Jenis Kelamin dan Kola Lahir.

2. Persebaran mahasiswa per-angkatan dan jurusan

Dari persebaran data mahasiswa per jurusan dan per angkatan menunjukkan jurusan Manajemen, Ilmu Komunikasi dan Akuntansi merupakan jurusan yang memiliki peminat mahasiswa yang tiga terbanyak namun Manajemen dan Akuntansi mengalami penurunan peminat pada tahun 2018 sedangkan Teknik Sipil, Teknik informatika dan Psikologi mengalami kenaikan peminat pada tahun 2018.

3. Status mahasiswa per jurusan (mengundurkan diri, lulus, aktif, dikeluarkan, wafat)

Dari persebaran status mahasiswa per jurusan menunjukkan mahasiswa paling banyak berstatus lulus pada jurusan manajemen, akuntansi, dan ilmu komunikasi. Dari gambar grafik ini juga menunjukkan perbandingan grafik mahasiswa jurusan psikologi memiliki jumlah mahasiswa aktif jauh lebih banyak daripada yang lulus dan mahasiswa yang mengundurkan diri juga lebih banyak daripada yang lulus. Hal ini sama dengan jurusan DKV, Arsitektur, dan desain produk. Sebaliknya ada jurusan Teknik Sipil, Teknik Elektro dan Teknik Industri memiliki lulusan yang lebih banyak daripada mahasiswa yang aktif ataupun mengundurkan diri.

4. Grafik Jumlah Kelulusan Mahasiswa Per Semester Lulus

Pada grafik kelulusan mahasiswa menunjukan peningkatan jumlah mahasiswa yang lulus dari tahun 2019 tahun ajaran ganjil sampai 2020 tahun ajaran ganjil. Peningkatan drastis ditunjukkan pada tahun ajaran 2019 ganjil ke tahun 2019 genap. Hal ini bisa jadi dikarenakan adanya penerapan kuliah online

semenjak pandemi covid-19 sehingga terjadi peningkatan lulusan.

5. Persentase lama studi

Pada perbandingan persentase lama studi menunjukkan persentase mahasiswa yang masih aktif sebesar 32%, mahasiswa yang sudah lulus dengan tepat waktu sebesar 33% dan mahasiswa yang lulus tidak tepat waktu sebesar 35%. Pada Angkatan 2016 persentase mahasiswa tidak tepat waktu mencapai lebih dari 35%, dan dari 3 angkatan 2016-2017 angkatan dengan persentase tidak tepat waktu terkecil adalah Angkatan 2017 dengan persentase 36%.

6. Persentase lama studi per jurusan

Pada persentase lama studi per jurusan menunjukkan jurusan yang paling konsisten dalam memiliki persentase mahasiswa yang lulus tepat waktu untuk angkatan 2016-2018 adalah jurusan akuntansi, manajemen, ilmu komunikasi, Teknik Sipil, Ilmu Komunikasi dan Teknik informatika dan untuk yang lulus tidak tepat waktu ada jurusan Akuntansi yang berada di 3 pie chart Angkatan 2016-2018.

7. Mode variable per lama studi group by lama studi (semua data yang telah memiliki label lama studi)

Menunjukkan data yang paling sering muncul pada variabel dan label lama studi tepat tepat waktu maupun tidak tepat waktu. Ditemukan dari proses ini untuk Angkatan yang paling banyak memiliki mahasiswa lulus tepat waktu adalah 2016 sedangkan tidak tepat waktu adalah 2017, SKS Tempuh Semester 5 untuk mahasiswa yang lulus tepat waktu adalah 112 dan tidak tepat waktu adalah 79 serta mahasiswa yang tepat waktu didominasi oleh mahasiswa yang berasal dari program kelas Reguler 1.

8. Nilai Rata-Rata variable per jurusan dan lama studi group by lama studi(semua data yang telah memiliki label lama studi)

Menunjukkan nilai rata-rata pada variabel dan label lama studi tepat tepat waktu maupun tidak tepat waktu. Untuk nilai indeks prestasi mahasiswa yang lulus tepat waktu memiliki nilai >=3.5 sedangkan yang lulus tidak tepat waktu <3.5 dengan umur mahasiswa yang lulus tidak tepat waktu rata-rata 20 dan yang tepat waktu 19.

V. KESIMPULAN

Proses pengujian klasifikasi pola data kelulusan mahasiswa dengan hasil akurasi terbaik adalah menggunakan metode algoritma *Ensemble Learning Bagging* atau *Random Forest* dengan melakukan *cross validation* dan *hyperparameter tuning (Grid Search CV)* dengan akurasi 96.1%. Penggunaan *cross validation* dan *hyperparameter tuning* terbukti dapat mempengaruhi dan mengoptimalkan akurasi learning. Faktor yang paling mempengaruhi pola kelulusan mahasiswa adalah jumlah SKS semester 5, Angkatan, Indeks Prestasi, Program Kelas dan Total Cuti. Dan untuk data yang digunakan dalam penelitian ini menunjukan persentase kelulusan mahasiswa yang kurang baik karena persentase mahasiswa yang lulus tidak tepat waktu perbandingannya di atas 50% pada 3 data angkatan 2016-2018. Hasil ini diharapkan dapat menjadi masukan untuk dosen ataupun perguruan tinggi dalam mengambil keputusan atau memberlakukan pembelajaran pada mahasiswa agar dapat meningkatkan kembali kualitas persentase mahasiswa yang dapat lulus dengan tepat waktu. Serta saran untuk penelitian selanjutnya adalah menggunakan data yang lebih lebgkap lagi seperti ada kolom prestasi, pekerjaan, pekerjaan orang tua, pendapatan, jam belajar atau kegiatan sehari-hari dan sebagainya dengan jumlah data yang lebih banyak dan beragam.

DAFTAR PUSTAKA

[1] M. B. Musthafa, C. Rahmad, R. A. Asmara, F. Rahutomo, and P. N. Malang, "Pemanfaatan Data Pddikti Sebagai Pendukung Keputusan Utilization of Pddikti Data As a Higher Education Management Decission Support," *J. Teknol. Inf. dan Ilmu Komput.*, vol. x, no. 30, pp. 1–11, 2018, doi: 10.25126/jtiik.202072585.

[2] A. Azahari, Y. Yulindawati, D. Rosita, and S. Mallala, "Komparasi Data Mining Naive Bayes dan Neural Network memprediksi Masa Studi Mahasiswa S1," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 7, no. 3, p. 443, 2020, doi: 10.25126/jtiik.2020732093.

[3] D. Fitrianah, S. Dwiasnati, H. Hikmayanti, and K. A. Baihaqi, "Penerapan Metode Machine Learning untuk Prediksi Nasabah Potensial menggunakan Algoritma Klasifikasi Naïve Bayes," *Fakt. Exacta*, vol. 14, no. 2, pp. 1979–276, 2021.

[4] E. Etriyani, "Perbandingan Tingkat Akurasi Metode Knn Dan Decision Tree Dalam Memprediksi Lama Studi Mahasiswa," *J. Ilm. Bin. STMIK Bina Nusant. Jaya Lubuklinggau*, vol. 3, no. 1, pp. 6–14, 2021, doi: 10.52303/jb.v3i1.40.

[5] P. Meilina, "Penerapan Data Mining dengan Metode Klasifikasi Menggunakan Decision Tree dan Regresi," *J. Teknol. Univ. Muhammadiyah Jakarta*, vol. 7, no. 1, pp. 11–20, 2015, [Online]. Available: jurnal.ftumj.ac.id/index.php/jurtek.

[6] R. Sudiyarno, A. Setyanto, and E. T. Luthfi, "Peningkatan Performa Pendeteksian Anomali Menggunakan Ensemble Learning dan Feature Selection," *Creat. Inf. Technol. J.*, vol. 7, no. 1, p. 1, 2021, doi: 10.24076/citec.2020v7i1.238.

[7] A. K. Hamoud, A. S. Hashim, and W. A. Awadh, "Predicting Student Performance in Higher Education Institutions Using Decision Tree Analysis," *Int. J. Interact. Multimed. Artif. Intell.*, vol. 5, no. 2, p. 26, 2018, doi: 10.9781/ijimai.2018.02.004.

[8] Saruni Dwiasnati and Yudo Devianto, "Classification of forest fire areas using machine learning algorithm," *World J. Adv. Eng. Technol. Sci.*, vol. 3, no. 1, pp. 008–015, 2021, doi: 10.30574/wjaets.2021.3.1.0048.

[9] L. Wen and M. Hughes, "Coastal wetland mapping using ensemble learning algorithms: A comparative study of bagging, boosting and stacking techniques," *Remote Sens.*, vol. 12, no. 10, 2020, doi: 10.3390/rs12101683.

[10] R. Puspita and A. Widodo, "Perbandingan Metode KNN, Decision Tree, dan Naïve Bayes Terhadap Analisis Sentimen Pengguna Layanan BPJS," *J. Inform. Univ. Pamulang*, vol. 5, no. 4, p. 646, 2021, doi: 10.32493/informatika.v5i4.7622.

[11] A. Ikhlas, A. Abdullah, and D. Y. Prasetyo, "Mesin Pembelajaran Ensemble Untuk Identifikasi Varietas Padi," *Inform. Pertan.*, vol. 29, no. 2, p. 123, 2020, doi: 10.21082/ip.v29n2.2020.p123-130.

[12] I. Ahmad, M. Basheri, M. J. Iqbal, and A. Rahim, "Performance Comparison of Support Vector Machine, Random Forest, and Extreme Learning Machine for Intrusion Detection," *IEEE Access*, vol. 6, pp. 33789–33795, 2018, doi: 10.1109/ACCESS.2018.2841987.

[13] H. Ding, G. Li, X. Dong, and Y. Lin, "Prediction of Pillar Stability for Underground Mines Using the Stochastic Gradient Boosting Technique," *IEEE Access*, vol. 6, pp. 69253–69264, 2018, doi: 10.1109/ACCESS.2018.2880466.

[14] B. H. Shekar and G. Dagnew, "Grid search-based hyperparameter tuning and classification of microarray cancer data," *2019 2nd Int. Conf. Adv. Comput. Commun. Paradig. ICACCP 2019*, pp. 1–8,

2019, doi: 10.1109/ICACCP.2019.8882943.

[15] Puneet and A. Chauhan, "Detection of Lung Cancer using Machine Learning Techniques Based on Routine Blood Indices," *2020 IEEE Int. Conf. Innov. Technol. INOCON 2020*, pp. 1–6, 2020, doi: 10.1109/INOCON50539.2020.9298407.

[16] V. Cox, "Translating Statistics to Make Decisions," *Transl. Stat. to Make Decis.*, 2017, doi: 10.1007/978-1-4842-2256-0.

[17] P. R. Anisha, C. Kishor Kumar Reddy, K. Apoorva, and C. Meghana Mangipudi, "Early Diagnosis of Breast Cancer Prediction using Random Forest Classifier," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 1116, no. 1, p. 012187, 2021, doi: 10.1088/1757-899x/1116/1/012187.

[18] C. Reimann, P. Filzmoser, K. Hron, P. Kynčlová, and R. G. Garrett, "A new method for correlation analysis of compositional (environmental) data – a worked example," *Sci. Total Environ.*, vol. 607–608, pp. 965–971, 2017, doi: 10.1016/j.scitotenv.2017.06.063.