Komparasi Model Prediksi Daftar Ulang Calon Mahasiswa Baru Menggunakan Metode Decision Tree Dan Adaboost

Muhammad Naufal Rabbani *[1], Ahmad Yusuf [2], Dwi Rolliawati [3]
Departemen Sistem Informasi[1][2][3]
Fakultas Sains dan Teknologi
Universitas Islam Negeri Sunan Ampel, Surabaya, Indonesia
bosnaufalemail@gmail.com[1], ahmadyusuf@uinsby.ac.id[2], dwi_roll@uinsby.ac.id[3]

***Abstract*—Every year, all the colleges hold new student enrollment. It is needed to start a new school academic year. Unfortunately, the number of students who resigned is considerably high to reach 837 students and caused 324 empty seats. The college's stakeholders can minimize the resignation number if the selection phase of new students is done accurately. Making a machine learning-based model can be the answer. The model will help predict which candidates who potentially complete the enrollment process. By knowing it in the first place will help the management in the selection process. This prediction is based on historical data. Data is processed and used to train the model using the Adaboost algorithm. The performance comparison between Adaboost and Decision Tree model is performed to find the best model. To achieve the maximum performance of the model, feature selection is performed using chi-square calculation. The results of this research show that the performance of Decision Tree is lower than the performance of the Adaboost algorithm. The Adaboost model has f-measure score of 90.9%, precision 83.7%, and recall 99.5%. The process of analyzing the data distribution of prospective new students was also conducted. The results were obtained if prospective students who tended to finish the enrollment process had the following characteristics: graduated from an Islamic school, 19-21 years old, parents' income was IDR 1,000,000 to IDR. 5,000,000, and through the SBMPTN program.**

***Keywords*— Classification, Adaboost, Ensemble Learning, Decision Tree, Enrollment**

***Abstrak*—Pada tiap tahunnya, semua perguruan tinggi mengadakan penerimaan mahasiswa baru. Hal ini dilakukan untuk membuka tahun ajaran baru. Sayangnya jumlah mahasiswa yang mengundurkan diri cukup tinggi mencapai 837 mahasiswa dan menyebabkan adanya 324 sisa kursi kosong. Perguruan tinggi dapat meminimalisir pengunduran diri yang terjadi bila melakukan tahap seleksi calon mahasiswa baru secara tepat. Membuat model berbasis *machine learning* bisa menjadi salah satu jawabannya. Model dapat membantu memprediksi kandidat mana yang berpotensi melakukan daftar ulang pada proses penerimaan mahasiswa baru. Dengan mengetahui lebih awal, dapat membantu pihak manajemen dalam proses seleksi. Prediksi ini dibuat berdasarkan data yang telah ada sebelumnya. Data diolah dan digunakan untuk melatih model dengan menggunakan algoritma *Adaboost.* Selain itu juga dilakukan perbandingan performa dengan model algoritma *Decision Tree* guna mencari model terbaik. Demi mendapatkan peforma model yang maksimal, maka dilakukan proses *feature selection* menggunakan *chi square*. Hasil riset ini memperlihatkan bahwa peforma *Decision Tree* lebih rendah daripada algoritma *Adaboost.* Model *Adaboost* memiliki skor *f-measure* 90.9%, *precision* 83.7% dan *recall* 99.5%. Proses analisis sebaran data calon mahasiswa baru juga dilakukan. Hasilnya didapatkan jika calon mahasiswa yang cenderung melakukan daftar ulang memiliki ciri: lulusan pondok, usia 19-21 tahun, penghasilan Orang tua Rp. 1,000,000 hingga Rp. 5,000,000, dan melalui jalur SBMPTN.**

***Kata Kunci*—Classification, Adaboost, Ensemble Learning, Decision Tree, Enrollment**

I. PENDAHULUAN

Berdasarkan kacamata *value chain* perguruan tinggi, hal yang menjadi input logistik dalam kegiatan utama perguruan tinggi adalah proses penerimaan mahasiswa baru. Awal keberhasilan perguruan tinggi dimulai dari adanya manajemen yang baik dan efektif dalam proses penerimaan mahasiswa baru [1]. Namun tidak mudah untuk melakukan seleksi calon mahasiswa baru. Ini disebabkan karena masih banyak calon mahasiswa yang lolos seleksi tapi memilih untuk mengundurkan diri. Data dari proses penerimaan mahasiswa baru di Universitas Islam Negeri Sunan Ampel Surabaya Tahun 2019 pada Tabel I mendukung pernyataan bahwa banyak calon mahasiswa baru yang tidak melakukan heregistrasi hingga tahap akhir. Hal ini menyebabkan adanya jumlah kursi kosong yang mencapai 324 kursi. Apabila jumlah kursi kosong dapat diminimalisir, maka pemasukan dari uang kuliah tunggal (UKT) bisa dimaksimalkan.

TABLE I. RANGKUMAN PAGU PENDAFTARAN TAHUN 2019

Fakultas	Total Pagu	Daftar Ulang	Tidak Daftar Ulang	Sisa pagu
Adab dan Humaniora	460	434	69	26
Dakwah dan Komunikasi	665	645	175	20
Ekonomi dan Bisnis Islam	630	591	116	39
Ilmu Sosial dan Ilmu Politik	333	310	31	23
Psikologi dan Kesehatan	140	134	11	6
Sains dan Teknologi	420	362	27	58

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v10i1.939, Copyright ©2020
Submitted : July 21, 2020; Revised : December 14, 2020; Accepted : January 8, 2021; Published : January 14, 2021

TABLE I. RANGKUMAN PAGU PENDAFTARAN TAHUN 2019

Fakultas	Total Pagu	Daftar Ulang	Tidak Daftar Ulang	Sisa pagu
Syariah dan Hukum	760	707	146	53
Tarbiyah dan Keguruan	770	729	131	41
Ushuluddin dan Filsafat	595	537	131	58
			837	324

Upaya pengurangan jumlah kursi kosong dapat dilakukan dengan pendekatan *data mining* [2]. Salah satu fungsi *data mining* adalah untuk melakukan prediksi melalui klasifikasi berdasarkan data yang telah ada sebelumnya. Sehingga model prediksi ini nantinya dapat membantu proses seleksi [3]. Model akan memberi tanda pada calon mahasiswa yang berpotensi untuk mengundurkan diri dan mahasiswa yang berpotensi tinggi untuk melanjutkan hingga tahap akhir.

Data mining pada bidang pendidikan (*Educational Data Mining*) merupakan hal perlu dirperhatikan agar bisa menggali informasi baru yang membantu pihak penyelenggara [23]. Baik untuk kepentingan *knowledge discovery* maupun klasifikasi. Data yang bisa dimanfaatkan untuk proses klasifikasi diambil dari data orang tua mahasiswa, data diri mahasiswa, data demografis, dan riwayat pendidikan [4]. Klasifikasi dapat dilakukan dengan banyak metode namun metode yang umum digunakan pada *Educational Data Mining* (EDM) adalah *Decision Tree* [19].

Pada penelitian sebelumnya untuk kasus EDM, metode *Decision Tree* mendapat skor peforma yang lebih baik daripada metode lainnya yaitu: *Logistic Regression* dan *Naïve bayes* [5][6]. Agar meningkatkan peforma *Decision* Tree lebih baik lagi, dapat dilakukan *Boosting* menggunakan metode *Adaboost*. Selain mampu menambah peforma, *Adaboost* juga mampu mengatasi masalah data yang tidak seimbang (*imbalance*) sehingga prediksi yang dihasilkan menjadi lebih akurat [7][8].

Oleh karena itu, penelitian ini melakukan pembuatan model prediksi (klasifikasi) menggunakan metode *Adaboost* guna meningkatkan efektifitas proses penerimaan mahasiswa baru. Proses *Boosting* menggunakan metode *Adaboost* pada model *Decision Tree* diharapkan mampu meningkatkan peforma dari model. Model yang dihasilkan akan digunakan untuk memprediksi kandidat yang berpotensi untuk melanjutkan proses daftar ulang hingga akhir . Sehingga dengan mengetahui lebih awal, dapat membantu pihak manajemen dalam proses seleksi dan merencanakan kebutuhan sumber daya [20].

II. METODOLOGI PENELITIAN

Metode yang digunakan berjenis metode kuantitatif yang disesuaikan. Adapun beberapa langkah penelitian yang dilakukan untuk mencapai tujuan akhir penelitian sebagai berikut:

A. Pemilihan Atribut

Langkah paling awal adalah dengan menentukan atribut atau *feature* dari data yang hendak digunakan untuk proses pengembangan model prediksi. Pemilihan atribut ini diambil berdasarkan irisan dari penelitian-penlitian sebelumnya yang dipresentasikan pada Tabel II.

TABLE II. ATRIBUT TERPILIH BERDASARKAN PENELITIAN TERDAHULU

Atribut	Penelitian Terdahulu					
	Wanjau & Muketha [9]	*Melati, dkk [10]*	*Yahya & Jananto [2]*	*Aradea, dkk. [11]*	*Rozi [12]*	*Ab Ghani, dkk. [5]*
Jenis Kelamin	✔		✔			
Usia	✔				✔	
Kota Asal			✔	✔		✔
Jenis Sekolah Asal		✔	✔			
Pekerjaan Orang Tua	✔	✔				
Penghasilan Orang Tua	✔				✔	
Fakultas			✔	✔		✔
Gelombang		✔				

B. Pengolahan Data

Data yang bersih dan *valid* tentu didapatkan melalui tahap-tahap pengolahan data, yaitu: *Data Validation*, *Data Integration and Transformation*, *Data Size Reduction and discretization* [13]. Berikut penjelasan tahapannya:

1) Data Validation

Ekstraksi *database* merupakan tahap pertama yang dilakukan. Data dari ekstraksi akan menjadi input utama. Namun tidak semua data dapat langsung digunakan. *Data Validation* dilakukan untuk melakukan seleksi pada data dengan menghilangkan *noise* dan *outliers*. Hasil akhirnya berupa data yang konsisten dan data lengkap atau tidak ada *missing value* [13].

2) Data Integration and Transformation

Langkah kedua adalah melakukan transformasi data sehingga nilai dari masing-masing atribut sesuai dengan ekspektasi. Misalnya pada atribut usia, di database tidak ada kolom usia melainkan hanya tanggal lahir dan periode masuk. Jadi, pada tahap ini akan dihitung usianya berdasarkan selisih tahun masuk dengan tanggal lahir.

3) Data Size Reduction and Discretization

Tahap akhir pengolahan data adalah melakukan optimiasi pada ukuran data. Jika proses sebelumnya memerlukan kolom periode masuk untuk menghitung usia, maka pada tahap ini kolom periode masuk dihapus karena tidak termasuk atribut terpilih. Hal ini bertujuan agar mempercepat eksekusi pelatihan model yang dilakukan.

C. Analisis Data

Analisis dilakukan dengan menggunakan metode statistik *chi square* yaitu perhitungan statistik untuk mengetahui ada atau tidaknya hubungan antara dua variabel. Proses perhitungannya menggunakan *library python pingouin* [21]. Pemilihan metode *chi square* disebabkan karena jenis variabel yang ada pada *dataset* berjenis kategorial yang setara. Setelah mengetahui pengaruhnya, maka dapat dijadikan rujukan untuk proses pemilihan fitur (*feature selection*) pada proses pelatihan model. Rumus perhitungan *chi square* tertulis pada persamaan (1) [14].

$$X^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \quad (1)$$

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v10i1.939, Copyright ©2020
Submitted : July 21, 2020; Revised : December 14, 2020; Accepted : January 8, 2021; Published : January 14, 2021

Keterangan:

E_i = nilai ekspektasi ke-i

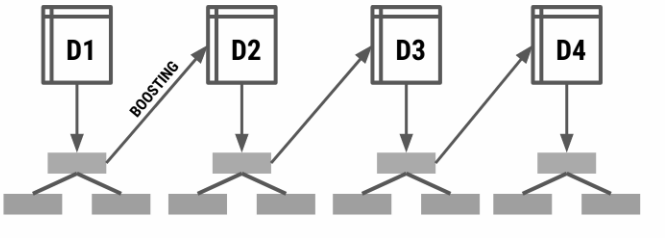
O_i = nilai observasi ke-i

Namun ada beberapa syarat wajib sebelum menghitung nilai *chi square* dari data [14]:

- Tidak boleh ada data dengan frekuensi aktual (F_0) yang bernilai 0 (nol).
- Jika tabel kontingensi data memiliki bentuk 2x2, maka harus tidak ada data yang memiliki nilai frekuensi harapan (F_h) kurang dari 5.
- Jika bentuk tabel kontingensi lebih dari 2x2 (contoh: 2x3), maka proporsi jumlah sel dengan nilai frekuensi harapan (F_h) kurang dari 5 tidak lebih dari 20%.

D. Pelatihan Model

Pelatihan model yang dilakukan menggunakan bahasa *python dan library scikit-learn* [22]. Data hasil *preprocessing* digunakan untuk melatih model dengan metode *Adaboost* yaitu teknik *ensemble learning* yang memiliki *weak learner* berupa model *decision tree*. *weak learner (Decision Tree)* yang digunakan memiliki 1 tingkat percabangan atau biasa disebut *decision stump* dan dilatih secara bergantian dengan data *training* yang sama. Tiap proses pelatihan akan memperbaiki distribusi data pada proses pelatihan selanjutnya. Sehingga meningkatkan peforma *weak learner* berikutnya. Iterasi akan berhenti apabila data *training* tidak bisa diperbaiki lagi atau apabila tercapai jumlah model maksimal [7]. Contoh alur pelatihan *Adaboost* dapat dilihat pada Gambar 1.



Gambar 1. Contoh alur pelatihan model *Adaboost*

Terdapat beberapa skenario pelatihan yang dilakukan sesuai dengan Tabel III. Penggunaan skenario seperti pada Tabel III bertujuan untuk mengetahui apakah jumlah data training mempengaruhi peforma model secara signifikan atau tidak. Skenario terbaik akan dipilih berdasarkan skor *Confusion Matrix: Precision, Recall,* dan *F Measure* tertinggi [15].

TABLE III. SKENARIO PELATIHAN MODEL

Skenario	Data Training	Data Testing
1	70%	30%
2	50%	50%
3	60%	40%

Confusion Matrix dapat dihitung apabila sudah diketahui hasil prediksinya. Ada beberapa kategori hasil prediksi yaitu: *True Positive (TP), True Negative (TN), False Positive (FP),* dan *False Negative (FN)* seperti yang dijelaskan pada Tabel IV.

TABLE IV. HASIL PREDIKSI

Confusion Matrix		**Kelas hasil prediksi**	
		Positif	***Negatif***
Kelas Sebenarnya	***Positif***	TP	FP
	Negatif	FN	TN

Masing-masing skor memiliki rumus yang berbeda. Berikut adalah rumus perhitungan dari *precision*, *recall*, dan *F*-measure secara berturut-turut [16]:

$$Precision = \frac{TN}{TN + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

$$F = 2\,\frac{Precision\ .Recall}{Precision + Recall} \quad (4)$$

F-measure bukanlah hasil perhitungan *mean* biasa, melainkan perhitungan *harmonic mean* dari *recall* dan *precision*. Penggunaan *harmonic mean* lebih kredibel dan proporsional daripada menghitung *mean* biasa dalam evaluasi model [16].

E. Evaluasi

Evaluasi hasil menjadi tahap paling akhir. Pada tahap ini diambil kesimpulan berdasarkan hasil analisis data dan pelatihan model yang dilakukan. Evaluasi pada bagian analisis data menjelaskan atribut apa saja yang memiliki pengaruh terhadap status daftar ulang mahasiswa serta memaparkan ciri mahasiswa yang berpontensi daftar ulang. Sedangkan pada pelatihan model, akan dievaluasi juga pengaruh *feature selection* dan mencari tahu perbandingan peforma *Adaboost* dan *Decision Tree*. Tidak lupa menjelaskan skenario terbaik, dan model terbaik yang dihasilkan.

III. HASIL DAN PEMBAHASAN

A. Pengolahan Data

Total jumlah keseluruhan data awal yang didapat adalah **11735** data. Kemudian berkurang karena proses validasi menjadi **7216** data. Komposisi data dipresentasikan pada Tabel V, baik sebelum dan sesudah data diolah. Tabel V menunjukkan perbandingan jumlah data yang timpang atau tidak seimbang (*imbalance*) antara yang tidak melanjutkan daftar ulang dan yang melanjutkan daftar ulang setelah diolah.

TABLE V. KOMPOSISI SEBARAN DATA

Status	**Sebelum**			**Sesudah**		
	2018	***2019***	***Total***	***2018***	***2019***	***Total***
Daftar Ulang	3921	4444	8365	2739	3296	6035
Tidak Daftar Ulang	2102	1266	3370	615	566	1181
Total Data			11735			7216

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v10i1.939, Copyright ©2020
Submitted : July 21, 2020; Revised : December 14, 2020; Accepted : January 8, 2021; Published : January 14, 2021

B. Analisis Data

Tabel VI merupakan hasil perhitungan *chi square* pada masing-masing atribut. Atribut dikatakan memiliki hubungan apabila nilai *p* kurang dari *a* (alpha = 0.05). Berdasarkan hasil yang dipresentasikan, atribut *jenis_kelamin* dan *pekerjaan_orang_tua* tidak memiliki hubungan dengan status daftar ulang calon mahasiswa baru.

TABLE VI. HASIL PERHITUNGAN *CHI SQUARE* TIAP ATRIBUT

Atribut	chi2	dof	p	memiliki hubungan
usia	12.92	4	0.012	Ya
jenis_kelamin	3.638	1	0.056	Tidak
kabupaten_asal	33.659	20	0.029	Ya
jenis_asal_sekolah	70.622	10	0	Ya
pekerjaan_orang_tua	18.492	14	0.185	Tidak
penghasilan_orang_tua	35.875	11	0	Ya
fakultas _terpilih	37.895	12	0	Ya
jalur_masuk	158.72	4	0	Ya

Meskipun ada sedikitnya 3 referensi penelitian terdahulu [2][9][10] yang menggunakan jenis kelamin dan pekerjaan orang tua sebagai atribut terpilih, penelitian ini menyatakan jika variabel *jenis_kelamin* dan *pekerjaan_orang_tua* tidak ada hubungannya dengan penerimaan mahasiswa baru karena nilai *p* lebih besar dari nilai *a*. Hasil ini berbeda dengan penelitian [2] yang menjelaskan jika terdapat hubungan antara variabel pekerjaan orang tua dengan penerimaan mahasiswa baru. Untuk memastikan signifikansi hubungan dari kedua atribut tersebut, maka ditambahkan beberapa skenario pelatihan dengan dan tanpa atribut *jenis_kelamin* maupun *pekerjaan_orang_tua*. Model terbaik dari salah satu skenario pelatihan akan menunjukkan apakah kedua atribut tersebut lebih baik digunakan atau tidak dalam proses pelatihan model.

C. Training Model

Analisis data menjelaskan bahwa status daftar ulang calon mahasiswa baru tidak memiliki hubungan dengan atribut *jenis_kelamin* dan *pekerjaan_orang_tua*. Oleh sebab itu, diadakan skenario *training model* tambahan tanpa atribut *pekerjaan_orang_tua, jenis_kelamin,* maupun keduanya. Hal ini biasa disebut *feature selection* (FS). Dari proses *training* didapatkan hasil skor model *Adaboost* dan model *Decision Tree* seperti yang ditunjukkan pada Tabel VII dan Tabel VIII.

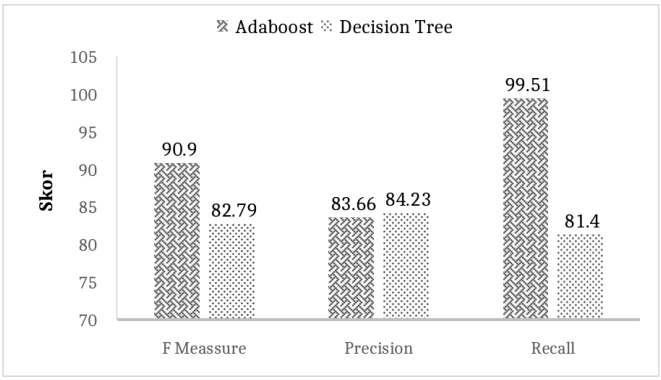
TABLE VII. SKOR TRAINING MODEL *ADABOOST*

Scenario	Train size	Test size	F Measure	Precision	Recall
Menggunakan semua atribut					
1	70%	30%	90.571792	83.505155	98.945031
2	60%	40%	90.573224	83.321654	99.207343
3	50%	50%	90.366059	83.593086	98.333333
Tanpa atribut jenis_kelamin dan pekerjaan_orang_tua					
1	70%	30%	90.756729	83.6219	99.222654
2	60%	40%	90.63214	83.362522	99.29078
3	50%	50%	90.23569	83.418223	98.266667
Tanpa atribut pekerjaan_orang_tua					
1	70%	30%	90.895257	83.660131	99.500278
2	60%	40%	90.621426	83.432783	99.165624
3	50%	50%	90.180483	83.323912	98.266667
Tanpa atribut jenis_kelamin					
1	70%	30%	90.651127	83.364399	99.333703
2	60%	40%	90.412529	83.374428	98.748436
3	50%	50%	90.398412	83.384962	98.7

TABLE VIII. SKOR TRAINING MODEL *DECISION TREE*

Scenario	Train size	Test size	F Measure	Precision	Recall
Menggunakan semua atribut					
1	70%	30%	81.756757	82.924043	80.621877
2	60%	40%	82.325383	84.080035	80.64247
3	50%	50%	82.356937	83.939079	80.833333
Tanpa atribut jenis_kelamin dan pekerjaan_orang_tua					
1	70%	30%	81.895093	83.209169	80.621877
2	60%	40%	81.717687	83.31166	80.183563
3	50%	50%	82.944162	84.226804	81.7
Tanpa atribut pekerjaan_orang_tua					
1	70%	30%	82.51551	83.839542	81.232649
2	60%	40%	82.283047	83.721934	80.892783
3	50%	50%	82.793694	84.235943	81.4
Tanpa atribut jenis_kelamin					
1	70%	30%	81.598199	82.715345	80.510827
2	60%	40%	82.312925	83.918509	80.767626
3	50%	50%	82.105263	83.66782	80.6

Tiap model memiliki skenario terbaiknya. Pada model *Adaboost,* skenario terbaiknya adalah skenario 1 tanpa menggunakan *pekerjaan_orang_tua.* Sedangkan pada model *Decision Tree,* skenario terbaiknya adalah skenario 3 tanpa menggunakan *pekerjaan_orang_tua*. Jika membandingkan skenario terbaik pada tiap model, maka didapatkan garfik seperti pada Gambar 2 yang menunjukkan bahwa peforma *Adaboost* lebih baik daripada *Decision* Tree.



Gambar 2. Perbandingan skenario terbaik dari *Adaboost* dan *Decision Tree*

Hasil *training model* membuktikan bahwa proses *feature selection* dengan *chi-square* dapat meningkatkan skor metriks

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v10i1.939, Copyright ©2020
Submitted : July 21, 2020; Revised : December 14, 2020; Accepted : January 8, 2021; Published : January 14, 2021

dari model sebesar 0.32% berdasarkan skor *F-measure* tertinggi dari setiap kelompok skenario. Selisihnya tidak terlalu besar atau signifikan, namun proses *feature selection* mengurangi jumlah atribut yang harus diproses. Hasil ini selaras dengan penelitian [13] dan berbeda dengan hasil penelitian [14] yang menjelaskan tidak adanya pengaruh proses *feature selection* menggunakan *chi square* terhadap peforma model.

Selaras dengan komparasi beberapa algoritma *machine learning* yang dilakukan oleh Srivastava dan Saiprasath G. dkk. menunjukkan bahwa algoritma *Adaboost* mengalahkan algoritma *decision tree* [17][18]. Hasil riset pada penelitian ini menjadi bukti tambahan bahwa *Adaboost* memliki peforma yang lebih baik dibandingkan dengan *Decision Tree*, terutama dalam kasus komposisi data yang tidak seimbang [8].

Untuk menambah pemahaman tentang model *Adaboost*, dilampirkan Gambar 3 yang merupakan contoh salah satu *decision stump* yang dihasilkan oleh model *Adaboost*.

```
Decision Stump:
|--- penghasilan_orang_tua <= 11.50
|   |--- jalur_masuk <= 3.50
|   |   |--- class: 1
|   |--- jalur_masuk >  3.50
|   |   |--- class: 0
|--- penghasilan_orang_tua >  11.50
|   |--- usia <= 19.50
|   |   |--- class: 1
|   |--- usia >  19.50
|   |   |--- class: 1

penghasilan_orang_tua categories:
11 | "10.000.001 - 15.000.00"
12 | "15.000.000 lebih"

jalur_masuk categories:
1 | "SNMPTN"
2 | "SBMPTN"
3 | "SPANPTKIN"
```

Gambar 3. Contoh *decision stump* dari model *Adaboost*

Potongan data yang ditampilkan pada Gambar 3 merupakan data yang sudah diberi label, sehingga perlu diterjemahkan agar lebih mudah untuk dipahami. Berdasarkan data yang ada, tidak ada kategori penghasilan orang tua dengan kode 11.50 yang ada hanyalah kategori 11 dan 12. maka maksud dari *penghasilan_orang_tua <= 11.50* adalah kategori 12 yaitu 15.000.000 lebih.

Pada cabang keputusan selanjutnya, tidak ada kategori jalur masuk dengan kode 3.50 namun terdapat kategori jalur masuk 1, 2 dan 3. Maka maksud dari *jalur_masuk <= 3.50* adalah jalur masuk yang bukan dari SNMPTN, SBMPTN, dan SPANPTKIN. Sehingga apabila keseluruhan *decision stump* diterjemahkan akan menjadi seperti pada Gambar 4.

```
|--- penghasilan_orang_tua lebih dari 15 juta
|   |--- jalur_masuk SELAIN dari SNMPTN, SBMPTN, dan SPANPTKIN
|   |   |--- class: daftar ulang
|   |--- jalur_masuk dari SNMPTN, SBMPTN, dan SPANPTKIN
|   |   |--- class: tidak daftar ulang
|--- penghasilan_orang_tua kurang dari 15 juta
|   |--- class: daftar ulang
```

Gambar 4. Hasil terjemah *decision stump Adaboost*

Cara membacanya alur keputusan dari decision stump pada Gambar 4 adalah apabila calon mahasiswa memiliki data *penghasilan_orang_tua* lebih dari 15 juta dan masuk melalui jalur selain SNMPTN, SBMPTN, dan SPANPTKIN maka diprediksi akan melanjutkan daftar ulang hingga tahap akhir. Namun apabila penghasilan orang tua lebih dari 15 juta tapi melalui jalur SNMPTN, SBMPTN, dan SPANPTKIN maka diprediksi akan mengundurkan diri. Sedangkan mahasiswa dengan data *penghasilan_orang_tua* kurang dari 15 juta diprediksi daftar ulang.

D. Observasi Sebaran Data

Selain menghitung *chi square*, juga dilakukan observasi sebaran data untuk mengetahui probabilitas pada beberapa atribut berdasarkan frekuensi kemunculan dari tiap nilai (*value*) yang ada pada atribut. Contoh pada Tabel IX tentang atribut *penghasilan_orang_tua* yang memiliki 12 nilai unik berupa kategori penghasilan orang tua. Pada tiap kategori penghasilan orang tua dihitung frekuensi kemunculannya pada *class* daftar ulang maupun tidak daftar ulang. Lalu dihitung probabilitasnya.

TABLE IX. OBSERVASI PADA ATRIBUT PENGHASILAN ORANG TUA

penghasilan_orang_tua	daftar ulang	tidak daftar ulang	probabilitas daftar ulang
3.000.001 - 4.000.000	909	144	86.32
4.000.001 - 5.000.000	605	102	85.57
2.500.001 - 3.000.000	557	100	84.78
1.500.001 - 2.000.000	582	106	84.59
2.000.001 - 2.500.000	508	94	84.39
500.001 - 1.000.000	630	121	83.89
1.000.001 - 1.500.000	766	148	83.81
5.000.001 - 7.500.000	582	125	82.32
7.500.001 - 10.000.000	299	65	82.14
10.000.001 - 15.000.000	112	26	81.16
500.000 atau kurang	367	110	76.94
15.000.000 lebih	118	40	74.68

Tabel IX memperlihatkan bahwa calon mahasiswa yang memiliki probabilitas di atas rata-rata (82, 54%) untuk daftar ulang adalah yang orang tuanya berpenghasilan rentang Rp. 1,000,000 - Rp. 5,000,000. Sedangkan yang penghasilannya 1,000,000 ke bawah dan 15,000,000 keatas memiliki porbabilitas di bawah rata-rata untuk melanjutkan daftar ulang.

Berdasarkan rentang usia mahasiswa baru S1 yaitu usia diantara 19 dan 21 tahun, semakin tua usia calon, maka semakin tinggi kemungkinan untuk daftar ulang. Hal ini dibuktikan pada Tabel X.

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v10i1.939, Copyright ©2020
Submitted : July 21, 2020; Revised : December 14, 2020; Accepted : January 8, 2021; Published : January 14, 2021

TABLE X. OBSERVASI PADA ATRIBUT USIA

usia	daftar ulang	tidak daftar ulang	probabilitas daftar ulang
17	205	62	76.78
18	3336	638	83.95
19	2023	406	83.29
20	375	63	85.62
21	69	8	89.61
22	11	1	91.67
23	4	2	66.67
24	1	0	100
25	1	0	100
26	1	0	100
39	9	1	90

Tabel berikutnya adalah Tabel XI yang membuktikan jika calon mahasiswa dari lulusan pondok berpotensi paling tinggi untuk daftar ulang, berbeda dengan calon mahasiswa dari alumni SMAN (sekolah Menengah Atas Negeri) yang memiliki kemungkinan terkecil untuk daftar ulang.

TABLE XI. OBSERVASI PADA ATRIBUT JENIS ASAL SEKOLAH

jenis_asal_sekolah	daftar ulang	tidak daftar ulang	probabilitas daftar ulang
pondok	52	3	94.55
smks	174	19	90.16
pkbm	15	2	88.24
smk	117	18	86.67
smta	13	2	86.67
ma	216	35	86.06
smas	1455	241	85.79
smkn	165	29	85.05
man	1297	231	84.88
madrasah	77	14	84.62
mas	939	172	84.52
sman	1357	393	77.54

Selanjutnya Tabel XII menginformasikan jika jalur SPANPTKIN menghadirkan calon mahasiswa baru dengan probabilitas daftar ulang terendah. Berbeda dengan jalur SBMPTN yang lebih banyak membawa mahasiswa yang berpotensi daftar ulang.

TABLE XII. OBSERVASI PADA ATRIBUT JENIS ASAL SEKOLAH

jalur_masuk	daftar ulang	tidak daftar ulang	probabilitas daftar ulang
SBMPTN	1186	149	88.84
SNMPTN	455	70	86.67
UMPTKIN	2195	386	85.04

TABLE XII. OBSERVASI PADA ATRIBUT JENIS ASAL SEKOLAH

jalur_masuk	daftar ulang	tidak daftar ulang	probabilitas daftar ulang
MANDIRI	1460	272	84.3
SPANPTKIN	739	304	70.85

Tabel terakhir adalah Tabel XIII yang merupakan hasil observasi pada atribut *fakultas_terpilih.* Hasilnya, probabilitas paling rendah berasal dari fakultas Filsafat dan Ushuluddin, sedangkan probabilitas paling tinggi melanjutkan daftar ulang adalah fakultas Psikologi dan Kesehatan.

TABLE XIII. OBSERVASI PADA ATRIBUT JENIS ASAL SEKOLAH

fakultas _terpilih	daftar ulang	tidak daftar ulang	probabilitas daftar ulang
Psikologi dan Kesehatan	173	22	88.72
Adab dan Humaniora	515	81	86.41
Ilmu Sosial dan Politik	402	67	85.71
Ekonomi dan Bisnis Islam	960	171	84.88
Tarbiyah dan Keguruan	904	161	84.88
Dakwah dan Komunikasi	922	199	82.25
Syari'ah dan Hukum	603	136	81.6
Sains dan Teknologi	470	111	80.9
program_magister	182	46	79.82
Ushuluddin dan Filsafat	537	140	79.32

IV. KESIMPULAN

Berdasarkan hasil pengolahan data didapatkan sebaran data yang tidak seimbang *(imbalance)*. Namun sebaran data yang tak seimbang tersebut dapat ditangani lebih baik oleh model *Adaboost* daripada model *Decision Tree*. Hasil pelatihan model menunjukkan bahwa *Adaboost* memiliki peforma yang lebih baik daripada model *Decision Tree* dengan skor *precision* 83.66%, *recall* 99.50%, dan *f-measure* 90.89%. Hasil riset ini menjadi bukti tambahan bahwa proses *boosting* dapat menambah peforma model.

Implementasikan model dapat meminimalisir jumlah kursi kosong dengan memberi tanda pada calon mahasiswa yang berpotensi untuk datar ulang maupun tidak agar dapat dipertimbangkan ulang oleh *stakeholder*. Analisis sebaran data yang dilakukan telah berhasil mengetahui ciri calon mahasiswa baru yang berpotensi untuk melanjutkan proses pendaftaran hingga akhir yaitu: lulusan pondok, usia 19-21 tahun, penghasilan Orang tua Rp. 1,000,000 hingga Rp. 5,000,000, dan melalui jalur SBMPTN.

REFERENCES

[1] R. K. Niswatin, "Sistem Seleksi Penerimaan Mahasiswa Baru Menggunakan Metode Weighted Product (WP)," *Seminar Nasional Teknologi Informasi dan Multimedia*, 2016.

[2] I. S. Melati, L. Linawati, and I. A. D. Giriantari, "Knowledge Discovery Data Akademik Untuk Prediksi Pengunduran Diri Calon Mahasiswa," *JTE*, vol. 17, no. 3, p. 325, Dec. 2018, doi: 10.24843/MITE.2018.v17i03.P04.

[3] P. Saini and A. Kumar Jain, "Prediction using Classification Technique for the Students' Enrollment Process in Higher Educational Institutions," *IJCA*, vol. 84, no. 14, pp. 37–41, Dec. 2013, doi: 10.5120/14646-2966.

[4] S. K. Wanjau and G. M. Muketha, "Improving Student Enrollment Prediction Using Ensemble Classifiers," *IJCATR*, vol. 07, no. 03, pp. 122–128, Mar. 2018, doi: 10.7753/IJCATR0703.1003.

[5] N. L. Ab Ghani, Z. Che Cob, S. Mohd Drus, and H. Sulaiman, "Student Enrolment Prediction Model in Higher Education Institution: A Data Mining Approach," in *Proceedings of the 3rd International Symposium of Information and Internet Technology (SYMINTECH 2018)*, vol. 565, M. A. Othman, M. Z. A. Abd Aziz, M. S. Md Saat, and M. H. Misran, Eds. Cham: Springer International Publishing, 2019, pp. 43–52.

[6] N. Yahya and A. Jananto, "Komparasi Kinerja Algoritma C 45 Dan Naive Bayes Untuk Prediksi Kegiatan Penerimaanmahasiswa Baru (Studi Kasus Universitas Stikubank Semarang)," p. 8, 2019..

[7] F. Khan, J. Ahamed, S. Kadry, and L. K. Ramasamy, "Detecting malicious URLs using binary classification through adaboost algorithm," *IJECE*, vol. 10, no. 1, p. 997, Feb. 2020, doi: 10.11591/ijece.v10i1.pp997-1005.

[8] A. N. Rais and A. Subekti, "Integrasi SMOTE Dan Ensemble AdaBoost

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v10i1.939, Copyright ©2020
Submitted : July 21, 2020; Revised : December 14, 2020; Accepted : January 8, 2021; Published : January 14, 2021

Untuk Mengatasi Imbalance Class Pada Data Bank Direct Marketing," *JI. Jurnal. Informatika*, vol. 6, no. 2, pp. 278–285, Sep. 2019, doi: 10.31311/ji.v6i2.6186.

[9] S. K. Wanjau and G. M. Muketha, "Improving Student Enrollment Prediction Using Ensemble Classifiers," *IJCATR*, vol. 07, no. 03, pp. 122–128, Mar. 2018, doi: 10.7753/IJCATR0703.1003.

[10] I. S. Melati, L. Linawati, and I. A. D. Giriantari, "Knowledge Discovery Data Akademik Untuk Prediksi Pengunduran Diri Calon Mahasiswa," *JTE*, vol. 17, no. 3, p. 325, Dec. 2018, doi: 10.24843/MITE.2018.v17i03.P04.

[11] Aradea, Satriyo A., Ariyan Z., and Yuliana A., "Penerapan Decision Tree Untuk Penentuan Pola Data Penerimaan Mahasiswa Baru," *Jurnal Penelitian Sitrotika*, 2011.

[12] A. F. Rozi, "Sistem Pendukung Keputusan Seleksi Penerimaan Calon Siswa/i Baru Menggunakan Algoritma C4.5 (Studi Kasus: SDIT An-Najah Jatinom Klaten)," *teknoin*, vol. 21, no. 1, Feb. 2015, doi: 10.20885/teknoin.vol21.iss1.art2.

[13] Hoiriyah, "Algoritma C4.5 Berbasis Seleksi Atribut Menentukan Kemungkinan Pengunduran Diri Mahasiswa," vol. 9, p. 9, 2018.

[14] A. Nisa and E. Darwiyanto, "Analisis Sentimen Menggunakan Naive Bayes Classifier dengan Chi-Square Feature Selection Terhadap Penyedia Layanan Telekomunikasi," p. 10, 2019.

[15] S. Amornsamankul, B. Pimpunchat, W. Triampo, J. Charoenpong, and N. Nuttavut, "A Comparison of Machine Learning Algorithms and Their Applications," *International journal of simulation: systems, science & technology*, Aug. 2019, doi: 10.5013/IJSSST.a.20.04.08.

[16] Y. Sasaki, "The truth of the F-measure," p. 5, Oct. 2007.

[17] A. K. Srivastava, "Comparison Analysis of Machine Learning algorithms for Steel Plate Fault Detection," vol. 06, no. 05, p. 4, 2019.

[18] Saiprasath G., Naren Babu R., ArunPriyan J., Vinayakumar R., Sowmya V., and Soman K. P., "Performance Comparison Of Machine Learning Algorithms For Malaria Detection Using Microscopic Images" *IJRAR*, 2019.

[19] A. A. Saa, M. Al-Emran, and K. Shaalan, "Mining Student Information System Records to Predict Students' Academic Performance," in *The International Conference on Advanced Machine Learning Technologies and Applications (AMLTA2019)*, vol. 921, A. E. Hassanien, A. T. Azar, T. Gaber, R. Bhatnagar, and M. F. Tolba, Eds. Cham: Springer International Publishing, 2018, pp. 229–239.

[20] W. Handoko, "Prediksi Jumlah Penerimaan Mahasiswa Baru Dengan Metode Single Exponential Smoothing (Studi Kasus Amik Royal Kisaran)," *Jurnal Teknologi dan Sistem Informasi*, no. 2, p. 8, 2019.

[21] R. Vallat, "Pingouin: statistics in Python". *Journal of Open Source Software*, vol. 3, 2018, doi: https://doi.org/10.21105/joss.01026

[22] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, 2011, pp . 2825–2830.

[23] A. Yusuf and J. A. R. Hakim, "Pengembangan Perangkat Lunak Prediktor Nilai Mahasiswa Menggunakan Metode Spectral Clustering dan Bagging Regresi Linier," vol. 1, p. 5, 2012.

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v10i1.939, Copyright ©2020
Submitted : July 21, 2020; Revised : December 14, 2020; Accepted : January 8, 2021; Published : January 14, 2021