Universitas Sebelas Maret Bidikmisi Applicant's Classification using C4.5 Algorithm

Muh. Safri Juliardi
Program Studi Informatika
Universitas Sebelas Maret
Jl. Ir. Sutami No. 36 A Surakarta
juliardi@student.uns.ac.id

Ristu Saptono
Program Studi Informatika
Universitas Sebelas Maret
Jl. Ir. Sutami No. 36 A Surakarta
ristu.saptono@staff.uns.ac.id

Denis Eka Cahyani
Program Studi Informatika
Universitas Sebelas Maret
Jl. Ir. Sutami No. 36 A Surakarta
denis.eka@staff.uns.ac.id

ABSTRACT

Bidikmisi scholarship is a scholarship for poor but outstanding students. Because of the amount applicants, there is a need to use an accurate method in the selection process of Bidikmisi scholarship, especially in Universitas Sebelas Maret's (UNS) environment. In this paper, C4.5 algorithm is proposed as a method to help on Bidikmisi recipients selection process. The dataset which is used is Bidikmisi applicants data from 2013 to 2015. The applicant's data from 2013 and 2014 is used as training data and the applicant's data from 2015 is used as testing data. Furthermore, oversampling and undersampling technique is used to address the class imbalance problem in training data. Finally the accuracy for each decision trees are compared to see which sampling method is better. The result of this study shows that the accuracy of the C4.5 algorithm decision tree with the applicant's data from 2015 as testing data is 79,80% and Area Under Curve (AUC) value 0.5539. Meanwhile, to compare the sampling method, the best decision tree based on testing result is chosen. Oversampling technique produce 82,69 % for precision, 91,22 % for recall, and 77,16 % for accuracy. While undersampling technique produce 82,78 % for precision, 91,22 % for recall, and 77,27 % for accuracy. Therefore it is concluded that undersampling technique gives a better accuracy than oversampling technique.

Keywords

Bidikmisi*, C4.5 algorithm, Decision Tree, Oversampling, Undersampling*

1. PENDAHULUAN

Beasiswa Bidikmisi adalah bantuan biaya pendidikan dari pemerintah untuk calon mahasiswa kurang mampu secara ekonomi dan memiliki potensi akademik baik untuk menempuh pendidikan di perguruan tinggi sampai lulus tepat waktu[1]. Universitas Sebelas Maret Surakarta (UNS) sebagai salah satu Perguruan Tinggi di Indonesia turut berpartisipasi dalam penyelenggaraan program Bidikmisi. Biro Administrasi Kemahasiswaan Pusat selanjutnya disebut sebagai Biro Mawa Pusat adalah biro yang ditunjuk untuk menyelenggarakan proses administrasi terkait beasiswa Bidikmisi dalam lingkungan UNS. Proses administrasi yang menjadi tanggung jawab Biro Mawa Pusat ini meliputi proses pendaftaran, verifikasi pendaftaran, penentuan penerima beasiswa, hingga monitoring penerima beasiswa Bidikmisi[2].

Metode yang digunakan dalam proses seleksi penerima beasiswa Bidikmisi oleh Biro Mawa Pusat saat ini adalah metode pembobotan kriteria. Bobot dari masing-masing kriteria ditentukan oleh Biro Mawa Pusat. Proses ini dilakukan dengan mengkonversi nilai atribut setiap pendaftar sesuai dengan kriteria yang telah ditentukan kemudian nilai hasil konversi digunakan untuk menentukan penerima beasiswa Bidikmisi. Pada metode pembobotan kriteria ini tidak ada pemanfaatan data dari tahun sebelumnya sebagai tolak ukur dalam proses seleksi penerima beasiswa Bidikmisi. Pada penelitian ini diusulkan penggunaan metode alternatif untuk proses seleksi penerima beasiswa Bidikmisi dengan memanfaatkan data seleksi dari tahun sebelumnya.

Salah satu metode yang dapat digunakan sebagai alternatif dalam proses seleksi penerima beasiswa Bidikmisi adalah metode klasifikasi. Klasifikasi adalah metode *data mining* yang dapat digunakan untuk mengelompokkan data ke dalam kelas-kelas yang telah ditentukan. Metode klasifikasi memiliki 2 tahapan, yaitu tahap pembelajaran untuk membangun model klasifikasi dan tahap klasifikasi yaitu ketika model klasifikasi digunakan untuk menentukan kelas suatu data[3]. Pada tahap pembelajaran dapat digunakan data

pendaftar dari tahun sebelumnya untuk membangun model klasifikasi. Sedangkan pada tahap klasifikasi digunakan data pendaftar tahun yang sedang berjalan untuk dikelompokkan menjadi diterima atau tidak diterima.

Terdapat beberapa algoritma yang dapat digunakan untuk membangun model klasifikasi, diantaranya adalah C4.5, *Artifical Neural Network, Random Tree,* dan *Naive Bayes*. Pada penelitian ini algoritma C4.5 dipilih sebagai algoritma untuk membangun model klasifikasi. Algoritma C4.5 ini dipilih karena berdasarkan penelitian yang dilakukan Khairul Sani *et al.*[4] dan Özsoy *et al.*[5], algoritma C4.5 memiliki tingkat akurasi yang lebih baik dari beberapa algoritma lain yang diuji, meliputi *Naive Bayes, Artificial Neural Network* (ANN), *DecisonStump*, FT, dan *REPTree*.

Dalam himpunan data yang digunakan untuk klasifikasi, terdapat kondisi yang disebut *class imbalance problem*. *Class imbalance problem* adalah ketidakseimbangan pada dataset karena terdapat kelas yang memiliki data jauh lebih banyak daripada kelas lain[3]. *Class imbalance problem* dapat menyebabkan bias terhadap kelas mayoritas, menurunnya performa klasifikasi, dan meningkatnya jumlah *false negative*. Pada kasus data pendaftar Bidikmisi, jumlah pendaftar yang diterima jauh lebih besar daripada jumlah pendaftar yang tidak diterima. Kasus ini terjadi karena tingkat *confidence* pendaftar untuk diterima beasiswa Bidikmisi sangat tinggi sehingga jumlah pendaftar yang diterima dan pendaftar yang ditolak menjadi tidak imbang. Untuk mengatasi kasus *class imbalance problem* ini dapat digunakan teknik *sampling* yang sesuai sehingga model hasil *training* dapat mendeteksi kelas sebuah data dengan akurat[3]. Pada penelitian ini untuk mengatasi *class imbalance problem* digunakan teknik *sampling* untuk membangkitkan data *training*.

2. BEASISWA BIDIKMISI

Beasiswa adalah bantuan seluruh biaya pendidikan sejak awal mengikuti seleksi penerimaan mahasiswa baru (SPMB) UNS sampai dengan batas waktu tertentu kepada para calon/mahasiswa tidak mampu/miskin yang mempunyai prestasi akademik atau non akademik[6].

Pemerintah melalui Direktorat Jenderal Pembelajaran dan Kemahasiswaan, Kementerian Riset Teknologi dan Pendidikan Tinggi mulai tahun 2010 meluncurkan Program Bantuan Biaya Pendidikan Bidikmisi yaitu bantuan biaya pendidikan bagi calon mahasiswa tidak mampu secara ekonomi dan memiliki potensi akademik baik untuk menempuh pendidikan di perguruan tinggi pada program studi unggulan sampai lulus tepat waktu[1].

3. ALGORITMA C4.5

Berdasarkan penjelasan Quinlan dalam bukunya C4.5 : Programs For Machine Learning[7], algoritma C4.5 untuk membangun pohon keputusan jika diberikan sebuah *dataset* T dengan kelas {C1, C2,....,Ck} adalah sebagai berikut :

a. Jika dataset T memenuhi kriteria berhenti, maka pohon keputusan untuk T adalah sebuah *leaf node* dengan kelas paling umum pada T sebagai hasil klasifikasi. Salah satu kriteria berhenti adalah jika semua kasus pada T termasuk pada satu kelas Ck.
b. Jika dataset T tidak memenuhi kriteria berhenti, maka pilih salah satu atribut X pada T, kemudian untuk tiap-tiap nilai X1, X2, ..., Xn, pecah T menjadi *subset* T1, T2, ..., Tn dimana Ti adalah *subset* dari T dengan atribut X hanya bernilai Xi.

   Untuk menentukan atribut X, digunakan nilai *gain ratio* terbesar diantara semua atribut yang ada. Tapi sebelumnya perlu menghitung nilai *Info* atau *Entropy* dari *dataset* T terlebih dahulu menggunakan persamaan berikut :

$$Info(T) = -\sum_{j=1}^{k} \frac{\text{freq}(C_j, T)}{|T|} * \log_2 \left(\frac{\text{freq}(C_j, T)}{|T|} \right) \quad (1)$$

   dimana :

   T = himpunan kasus pada dataset
   freq(Cj , T) = jumlah kasus pada T yang diklasifikasi sebagai kelas Cj
   |T| = jumlah kasus pada dataset T

   Kemudian hitung nilai *Info* dari atribut X dengan persamaan berikut :

$$Info_x(T) = -\sum_{i=1}^{n} \frac{|T_i|}{|T|} * info(T_i) \quad (2)$$

   dimana :

   |Ti| = jumlah kasus pada *subset* Ti
   i = jenis nilai dari atribut X
   n = jumlah nilai dari atribut X

   Selanjutnya dari persamaan (1) dan (2) diatas dapat dihitung nilai *gain* dari atribut X menggunakan persamaan berikut :

$$gain(X) = info(T) - info_x(T_i) \quad (3)$$

Sebelum menghitung nilai *gain ratio,* perlu dihitung terlebih dahulu nilai *splitinfo* menggunakan persamaan berikut :

$$splitinfo(X) = -\sum_{i=1}^{n} \frac{|T_i|}{|T|} * log_2\left(\frac{|T_i|}{|T|}\right) \quad (4)$$

Kemudian dari persamaan (3) dan (4) dapat dihitung nilai *gain ratio* dari atribut X menggunakan persamaan berikut :

$$gainratio(X) = gain(X)/splitinfo(X) \quad (5)$$

c. Untuk tiap Ti, ulangi langkah pertama sampai tidak ada atribut lagi untuk dipilih atau mencapai kriteria berhenti.

4. *DATA PREPROCESSING*

Data Preprocessing adalah proses untuk mengubah data agar dapat diolah sesuai dengan metode yang ditentukan. Menurut Han *et al*.[3] terdapat 4 metode *data preprocessing* yaitu :

a. *Data Cleaning*

Pada penelitian ini, *data cleaning* digunakan untuk membersihkan data dari *missing values,* yaitu kondisi dimana suatu atribut memiliki nilai yang tidak valid. Terdapat beberapa teknik untuk menangani *missing values*. Diantaranya adalah dengan menghapus data yang memiliki banyak atribut kosong atau dengan mengisi data yang kosong dengan suatu konstanta[3].

b. *Data Integration*

Data integration adalah proses penggabungan data dari beberapa sumber data. Proses integrasi harus dilakukan secara hati-hati untuk menghindari redundansi dan inkonsistensi pada dataset yang dihasilkan.

c. *Data Reduction*

Dataset yang dihasilkan dari proses *data cleaning* dan *data integration* terkadang memiliki jumlah yang besar. Proses analisa dan *mining* pada *dataset* yang besar dapat memakan waktu yang lama. Sehingga perlu diterapkan proses *data reduction* untuk mengurangi jumlah *dataset* yang digunakan namun dengan menjaga integritas dari data asli. Proses mining pada dataset hasil *data reduction* seharusnya lebih efisien namun dengan hasil analisa yang sama.

d. *Data Transformation*

Data transformation adalah proses mengubah data menjadi bentuk yang sesuai untuk diproses menggunakan algoritma *data mining*. Terdapat beberapa strategi pada *data transformation* diantaranya adalah *normalization*, *discretization*, dan *concept hierarchy* (hirarki konsep). *Normalization* atau normalisasi berlaku pada data numerik dimana data angka diubah menjadi suatu skala tertentu seperti antara -1 sampai 1 atau antara 0 sampai 1. *Discretization* berlaku untuk data numerik, dimana data numerik diubah menjadi data label interval atau label konseptual. Contoh untuk data umur bisa diubah menjadi label interval "20-30 tahun", atau label konseptual seperti "muda", "tua", dan sebagainya. Label yang dihasilkan kemudian dapat dikonversi lagi menjadi konsep label yang lebih tinggi untuk lebih menyederhanakan data. Proses konversi label menjadi konsep yang lebih tinggi ini disebut *concept hierarchy*[3].

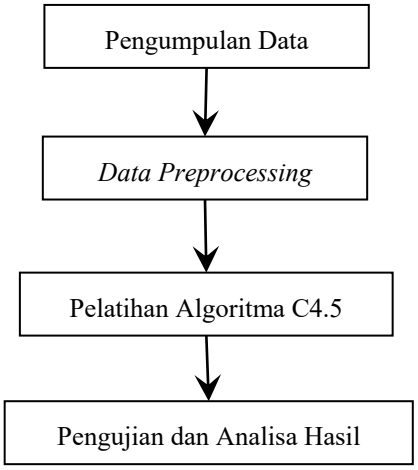
5. *OVERSAMPLING* DAN *UNDERSAMPLING*

Terdapat beberapa pendekatan yang dapat digunakan untuk meningkatkan akurasi algoritma klasifikasi pada kasus *class imbalance problem*, diantaranya adalah *oversampling* dan *undersampling*. Kedua pendekatan tersebut mengubah distribusi data *training* sehingga kelas yang lebih kecil dapat direpresentasikan dengan baik. Teknik *oversampling* dilakukan dengan menggandakan data pada kelas yang lebih kecil sehingga *data training* memiliki kelas yang seimbang. Sedangkan teknik *undersampling* dilakukan dengan mengurangi jumlah data pada kelas yang lebih besar sehingga *data training* memiliki kelas yang seimbang [3]. Contoh penggunaan teknik *oversampling* dan *undersampling* dapat dilihat pada contoh kasus berikut :

Diberikan *training set* yang mengandung 100 data positif dan 1000 data negatif. Pada teknik *oversampling*, maka dilakukan replikasi data pada kelas yang lebih kecil sehingga membentuk *training set* baru yang mengandung 1000 data positif dan 1000 data negatif. Jika menggunakan teknik *undersampling*, maka dilakukan eliminasi data secara acak pada kelas yang lebih besar sehingga membentuk *training set* baru yang mengandung 100 data positif dan 100 data negatif[3].

6. METODOLOGI PENELITIAN

Alur penelitian klasifikasi pendaftar beasiswa Bidikmisi digambarkan pada gambar 1.



Gambar 1. Diagram Alir Metodologi Penelitian

a. Mengajukan permintaan data pendaftar Bidikmisi kepada Biro Mawa Pusat UNS.
b. Melakukan *preprocessing* terhadap data pendaftar Bidikmisi.
c. Menggunakan algoritma C4.5 untuk membuat pohon keputusan.
d. Melakukan pengujian dan analisa hasil terhadap pohon keputusan yang dibuat.

7. HASIL DAN PEMBAHASAN

7.1 Deskripsi Data

Data yang digunakan pada penelitian ini adalah data pendaftar Bidikmisi tahun 2013, 2014, dan 2015 yang didapatkan dari Biro Mawa Pusat Universitas Sebelas Maret (UNS). Terdapat 4034 data pendaftar yang didapatkan dengan rincian 1336 data pendaftar tahun 2013, 1752 data pendaftar tahun 2014, dan 946 data pendaftar tahun 2015.

Atribut yang ada pada data yang didapatkan meliputi nama pendaftar, status ayah, status ibu, pekerjaan ayah, pekerjaan ibu, penghasilan ayah, penghasilan ibu, jumlah tanggungan, status rumah, dan status diterima. Khusus untuk data tahun 2015 tidak memiliki atribut jumlah tanggungan. Sehingga atribut jumlah tanggungan pada data pendaftar tahun 2014 tidak digunakan sebagai data *training*.

Tabel 2. Jumlah Data Pendaftar Bidikmisi Tahun 2013, 2014, dan 2015

Data Tahun	Jumlah Diterima	Jumlah Ditolak	Total
2013	1320	16	1336
2014	1346	406	1752
2015	775	171	946
Total Data			4034

7.2 *Data Preprocessing*

7.2.1 Data Cleaning

Pada tahap *data cleaning* dilakukan pembersihan data dari *missing values* dan duplikasi. *Missing values* adalah kondisi dimana suatu atribut memiliki nilai yang tidak valid. Pada penelitian ini, data dengan kasus *missing values* tidak digunakan sebagai data training. Hasil pengecekan *missing values* menemukan 6 kasus *missing values* pada keseluruhan *dataset* dengan rincian 3 kasus pada data 2013 dan 3 kasus pada data 2014.

Kasus duplikasi data adalah kondisi dimana semua atribut suatu data termasuk nama pendaftar memiliki kesamaan dengan semua atribut suatu data yang lain. Pengecekan duplikasi data dilakukan pada data antar tahun. Hasil pengecekan duplikasi menemukan bahwa data pendaftar tahun 2014 sebanyak 1336 data merupakan duplikat dari seluruh data pendaftar tahun 2013. Karena semua data pendaftar tahun 2013 memiliki duplikat pada data tahun 2014 maka data pendaftar tahun 2013 tidak digunakan pada penelitian ini.

7.2.2 Data Transformation

Pada tahap ini digunakan strategi *concept hierarchy* untuk menyederhanakan data atribut penghasilan ayah dan penghasilan ibu. Atribut penghasilan ayah dan penghasilan ibu memiliki nilai berbentuk *range* penghasilan. Terdapat 18 nilai unik pada atribut tersebut. Nilai unik yang banyak akan menghasilkan cabang pohon keputusan yang banyak dan menambah waktu *training*. Sehingga perlu dilakukan penyederhanaan nilai untuk mengurangi ukuran pohon keputusan yang dihasilkan.

Proses penyederhanaan data menggunakan *concept hierarchy* dilakukan dengan pertama mengurutkan data penghasilan dari *range* terkecil sampai *range* terbesar.Kemudian untuk tiap *range* diberikan label yang sesuai. Pemberian label didasarkan pada hasil Survey Sosial Ekonomi Nasional 2015 oleh Badan Pusat Statistik[8]. Tabel 2 menunjukkan konversi nilai atribut penghasilan.

Tabel 2. Tabel Konversi Nilai Atribut Penghasilan

***Range* Penghasilan**	**Nilai Konversi**
< Rp. 1.000.001	rendah
Rp. 1.000.001 - Rp. 2.000.000	menengah
> Rp. 2.000.001	tinggi

Tabel 3 berikut menunjukkan jenis-jenis nilai tiap atribut setelah melewati tahap *data transformation.* Kolom kode diberikan untuk memudahkan pembacaan pada tabel 4.

Tabel 3. Jenis Nilai Tiap Atribut

Atribut	**Nilai**	**Kode**
Status Ayah & Status Ibu	hidup	H
	bercerai	B
	wafat	W
Pekerjaan Ayah & Pekerjaan Ibu	Tidak bekerja	TB
	Lainnya	L
	Petani	P
	Nelayan	N
	Wirausaha	W
	Pegawai Swasta	PS
	PNS	PNS
	TNI / POLRI	TNI
Penghasilan Ayah & Penghasilan Ibu	rendah	R
	menengah	M
	tinggi	T
Status Rumah	Sendiri	Sn
	Menumpang	Mn
	Menumpang	MTI
	Sewa Tahunan	ST
	Sewa Bulanan	SB
	Tidak Memiliki	TM
Status Diterima	Diterima	DTR
	Ditolak	DTK

Kemudian contoh data pendaftar Bidikmisi tahun 2014 yang telah melewati tahap *data cleaning* dan *data transformation* dapat dilihat pada tabel 4. Kode yang digunakan pada tabel tersebut disesuaikan dengan tabel 2 diatas.

Tabel 4. Contoh data yang telah melewati tahap *data cleaning* dan *data transformation*

A1	**A2**	**A3**	**A4**	**A5**	**A6**	**A7**	**A8**
H	W	W	TB	R	R	Mn	DTR
H	H	W	TB	M	R	Sn	DTR
H	H	P	PNS	R	M	Mn	DTR
H	H	PS	L	R	R	Sn	DTK
H	H	N	W	R	R	Sn	DTK
H	H	L	TB	M	R	Sn	DTK

Keterangan :

A1 = Atribut Status Ayah

A2 = Atribut Status Ibu

A3 = Atribut Pekerjaan Ayah

A4 = Atribut Pekerjaan Ibu

A5 = Atribut Penghasilan Ayah

A6 = Atribut Penghasilan Ibu

A7 = Atribut Status Rumah

A8 = Atribut Status Diterima

7.2.3 Sampling

Data yang telah melalui proses *data cleaning* dan *data transformation* telah siap digunakan pada proses *training* untuk membuat pohon keputusan menggunakan algoritma C4.5. Namun, karena jumlah data pendaftar yang diterima dan ditolak tidak seimbang, dilakukan proses *sampling* terlebih dahulu untuk membangkitkan data *training*. Teknik *sampling* yang digunakan pada penelitian adalah teknik *oversampling* dan *undersampling*. Pembangkitan data *training* dilakukan secara *random* dimana pada metode *oversampling* dilakukan duplikasi data kelas ditolak sehingga didapatkan rasio kelas yang diinginkan. Sedangkan pada metode *undersampling* dilakukan eliminasi data kelas diterima secara *random* sehingga didapatkan rasio kelas yang diinginkan.

Pada penelitian ini digunakan beberapa rasio kelas untuk melihat kinerja masing-masing teknik *sampling*. Tabel 5 menunjukkan rasio kelas yang digunakan untuk proses pelatihan pada penelitian ini.

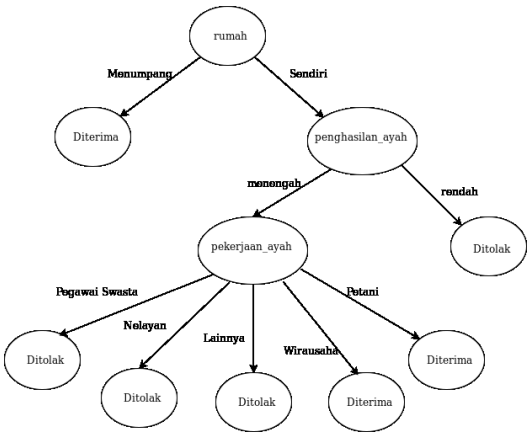
Tabel 5. Tabel Rasio Kelas

Nama	**Kelas**		**Rasio**	**Keterangan**
	Diterima	**Ditolak**		
Dataset 1	1346	406	3 : 1	Rasio Data Asli

Dataset 2	1346	1346	1 : 1	*Oversampling*
Dataset 3	1346	1009	4 : 3	
Dataset 4	1346	897	3 : 2	
Dataset 5	1346	673	2 : 1	
Dataset 6	406	406	1 : 1	*Undersampling*
Dataset 7	542	406	4 : 3	
Dataset 8	609	406	3 : 2	
Dataset 9	812	406	2 : 1	

7.3 Pelatihan Algoritma C4.5

Proses pelatihan dilakukan untuk membangkitkan pohon keputusan. Pelatihan dilakukan menggunakan algoritma C4.5 yang diimplementasikan dalam bahasa pemrograman PHP. Data yang digunakan pada proses pelatihan ini adalah data pendaftar tahun 2014 yang telah melalui tahap *preprocessing* serta data pendaftar setelah melalui tahap *sampling.* Gambar 2 menunjukkan visualisasi pohon keputusan hasil pelatihan menggunakan sampel data.



Gambar 2. Visualisasi Pohon Keputusan C4.5

Pohon keputusan yang dihasilkan tersebut kemudian dapat digunakan untuk melakukan klasifikasi pendaftar Bidikmisi kedalam kelas diterima atau ditolak.

7.4 Pengujian dan Analisa Hasil

Pengujian pohon keputusan dilakukan menggunakan data pendaftar Bidikmisi tahun 2015 dengan 946 data. Karena proses pelatihan dilakukan menggunakan 9 jenis *dataset* yang berbeda, yaitu 1 *dataset* dengan distribusi kelas asli dan 8 *dataset* hasil teknik *sampling,* maka akan dihasilkan 9 pohon keputusan yang berbeda. Masing-masing pohon keputusan tersebut melalui tahap pengujian untuk menguji akurasi tiap pohon keputusan dalam mengklasifikasi data pendaftar Bidikmisi 2015.

Pengujian akurasi pohon keputusan dilakukan menggunakan metode *confusion matrix*. Hasil pengujian akurasi didapatkan dengan cara membandingkan kelas hasil klasifikasi pohon keputusan dengan nilai atribut status penerimaan bidikmisi pada masing-masing data. Tabel 6 menunjukkan *confusion matrix* untuk tiap pohon keputusan.

Tabel 6. *Confusion Matrix* untuk tiap Pohon Keputusan

Actual Class	Diterima(DTR)		Ditolak(DTK)		Total Data
Predicted Class	DTR (TP)	DTK (FN)	DTR (FP)	DTK (TN)	
PKA	737	38	153	18	946
PKO1	532	243	104	67	946
PKO2	668	107	138	33	946
PKO3	674	101	140	31	946
PKO4	699	76	149	22	946
PKU1	513	262	109	62	946
PKU2	630	145	128	43	946
PKU3	640	135	123	48	946
PKU4	707	68	147	24	946

Keterangan :

PKA = Pohon Keputusan dari *dataset* asli

PKO = Pohon Keputusan dari *dataset* teknik *oversampling*

PKU = Pohon Keputusan dari *dataset* teknik *undersampling*

DTR = Diterima

DTK = Ditolak

Dari tabel 6 dapat dihitung nilai *precision*, *recall*, dan *accuracy* dari masing-masing pohon keputusan. Ketiga nilai tersebut dihitung untuk mengukur tingkat akurasi masing-masing pohon keputusan untuk mengklasifikasi data pendaftar Bidikmisi. Nilai *precision, recall,* dan *accuracy* dari masing-masing pohon keputusan dapat dilihat pada tabel 7, 8, dan 9 berikut.

Tabel 7. Nilai *Precision, Recall,* dan *Accuracy* pohon keputusan PKA

PK	Rasio	*Precision*	*Recall*	*Accuracy*
PKA	1346 : 406	82,80%	95,90%	79,80%

Tabel 8. Nilai *Precision, Recall,* dan *Accuracy* pohon keputusan PKO

PK	Rasio	*Precision*	*Recall*	*Accuracy*

PKO1	1346 : 1346	83,64 %	68,64 %	63,31 %
PKO2	1346 : 1009	82,87 %	86,19 %	74,10 %
PKO3	1346 : 897	82,80 %	86,96 %	74,52 %
PKO4	1346 : 673	82,43 %	90,19 %	76,21 %

Tabel 9. Nilai *Precision, Recall,* dan *Accuracy* pohon keputusan PKU

PK	Rasio	*Precision*	*Recall*	*Accuracy*
PKU1	406 : 406	82,47 %	66,19 %	60,78 %
PKU2	542 : 406	83,11 %	81,29 %	71,14 %
PKU3	609 : 406	83,37 %	82,58 %	72,72 %
PKU4	812 : 406	82,78 %	91,22 %	77,27 %

Dari tabel 7, 8, 9 diatas dapat dilihat perbandingan nilai *precision, recall,* dan *accuracy* masing-masing pohon keputusan. Pohon keputusan yang menghasilkan nilai *precision* terbaik adalah PKO1 dengan nilai *precision* 83,77 % sedangkan nilai *precision* paling rendah dihasilkan oleh pohon keputusan PKO4 yaitu 82,69 %.

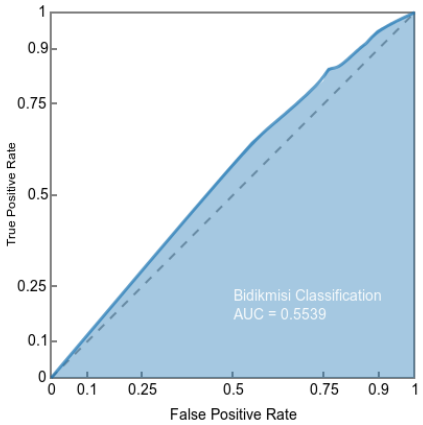
Sementara untuk nilai *recall* ditemukan hasil yang cukup menarik dimana terdapat beberapa variasi nilai yang dihasilkan. Dari variasi nilai *recall* yang dihasilkan, pohon keputusan dapat dikelompokkan menjadi 3 kelompok yaitu kelompok PKA, PKO4, dan PKU4 yang memiliki nilai diatas 90 %, kelompok kedua adalah PKO2, PKO3, PKU2, dan PKU3 dengan nilai *recall* antara 80 sampai 90 %, serta kelompok ketiga yaitu PKO1 dan PKU1 dengan nilai *recall* dibawah 70 %. Perbedaan ketiga kelompok tersebut adalah pada rasio kelas, dimana kelompok pertama dan kedua memiliki jumlah pendaftar yang ditolak lebih kecil daripada jumlah pendaftar yang diterima, sedangkan pada kelompok ketiga memiliki rasio yang seimbang antara pendaftar diterima dan ditolak. Rasio yang seimbang antara pendaftar diterima dan ditolak ini menyebabkan jumlah kasus *False Negative* menjadi besar karena memang data awal pendaftar Bidikmisi tahun 2014 memiliki jumlah pendaftar ditolak lebih kecil daripada pendaftar diterima.

Untuk nilai *accuracy* didapatkan nilai terbesar adalah 79,80% yang dihasilkan oleh PKA. Sedangkan untuk nilai *accuracy* terkecil adalah 61,09% yang dihasilkan oleh PKU1. Pada nilai *accuracy* ini juga ditemukan pola yang sama dengan nilai *recall* yaitu terdapat pola dimana nilai *accuracy* tiap pohon keputusan berbanding lurus dengan rasio kelas *dataset* yang artinya semakin mendekati rasio kelas asli nilai, *accuracy* pohon keputusan semakin tinggi.

Dari hasil diatas juga dapat dilihat perbandingan antara teknik *oversampling* dan *undersampling.* Perbandingan metode ini dilakukan dengan memilih pohon keputusan dengan komposisi nilai *precision, recall,* dan *accuracy* terbaik dari tiap kelompok pohon keputusan hasil *sampling.* Untuk metode *oversampling* dipilih pohon keputusan PKO4 sedangkan untuk metode *undersampling* dipilih pohon keputusan PKU4. Perbandingan hasil pengujian kedua pohon keputusan tersebut menunjukkan bahwa PKU4 memiliki nilai *precision, recall,* dan *accuracy* lebih baik daripada PKO4.

Untuk perbandingan keseluruhan antara pohon keputusan dari data asli dengan pohon keputusan dari *dataset* hasil *sampling* menunjukkan bahwa teknik *sampling* dapat meningkatkan *precision* dari pohon keputusan. Nilai *precision* terbesar yang dihasilkan oleh masing-masing kelompok pohon keputusan hasil *sampling* lebih besar daripada nilai *precision* PKA. Namun nilai *recall* dan *accuracy* PKA adalah yang terbesar diantara semua pohon keputusan. Sehingga untuk akurasi secara keseluruhan dapat disimpulkan bahwa pohon keputusan dari data asli memiliki akurasi yang lebih baik daripada pohon keputusan dari hasil *sampling.* Hal ini terjadi karena ada kemungkinan bahwa rasio pada *dataset* asli merupakan rasio yang paling optimal untuk membangkitkan pohon keputusan. Hasil ini didukung oleh penelitian sebelumnya[9] yang menunjukkan bahwa bisa jadi rasio asli dari suatu *dataset* merupakan rasio yang menghasilkan tingkat *error* terkecil. Meskipun demikian, hasil penelitian ini menunjukkan bahwa teknik *sampling* dapat meningkatkan tingkat deteksi pada data pendaftar ditolak.

Pengujian selanjutnya dilakukan menggunakan metode *Receiver Operating Characteristic* (ROC) *Curve*. Gambar 3 menunjukkan kurva ROC hasil *plotting.*



Gambar 2. Kurva ROC

Selanjutnya dihitung luas *Area Under Curve (AUC)* menggunakan *library RocChart*[10] dan didapatkan nilai 0,5539. Nilai AUC dari kurva ROC hasil pengujian pohon keputusan C4.5 hanya sedikit lebih baik daripada nilai AUC garis diagonal yaitu 0,5. Dari hasil tersebut maka dapat disimpulkan bahwa algoritma C4.5 dapat digunakan untuk mengklasifikasi pendaftar beasiswa Bidikmisi dengan catatan hasil klasifikasi yang dapat dipercaya adalah sebesar 79,80 % atau terdapat kemungkinan *error* pada hasil klasifikasi sebesar 20,20 %.

8. KESIMPULAN

Berdasarkan penelitian yang dilakukan, didapatkan hasil klasifikasi pendaftar Bidikmisi tahun 2015 menggunakan algoritma C4.5 serta data pendaftar Bidikmisi tahun 2014 sebagai data latih menghasilkan tingkat akurasi 79,80 %. Sedangkan hasil analisa menggunakan Kurva ROC didapatkan nilai AUC 0.5539 yang hanya sedikit lebih baik dari nilai AUC garis diagonal. Dengan demikian dapat diambil kesimpulan bahwa algoritma C4.5 dapat digunakan sebagai metode alternatif untuk membantu proses seleksi penerima beasiswa Bidikmisi.

Perbandingan nilai *precision, recall,* dan *accuracy* dari pohon keputusan dengan komposisi hasil pengujian terbaik dari masing-masing teknik *sampling* menunjukkan bahwa teknik *undersampling* dapat menghasilkan pohon keputusan dengan tingkat akurasi yang lebih baik daripada teknik *oversampling.* Meskipun secara akurasi pohon keputusan hasil teknik *sampling* tidak sebaik pohon keputusan dari rasio kelas asli, namun pohon keputusan hasil teknik *sampling* dapat mengklasifikasi data kelas ditolak dengan lebih baik.

Untuk penelitian selanjutnya perlu dilakukan penambahan data *training* menggunakan data hasil *survey* ke rumah pendaftar Bidikmisi serta data pendaftar dari beberapa tahun yang berbeda. Dari segi *sampling* dapat digunakan teknik *sampling* yang lain seperti teknik SMOTE (*Synthetic Minority Oversampling Technique*) untuk membangkitkan data *sampling.*

9. REFERENSI

[1] Republik Indonesia. (2015). Pedoman Penyelenggaraan Bantuan Bidikmisi Tahun 2015. Direktorat Jenderal Pembelajaran dan Kemahasiswaan Kementerian Riset Teknologi dan Pendidikan Tinggi

[2] Universitas Sebelas Maret. (2012). Prosedur Mutu Nomor UN27.14.2.PM03 tentang Beasiswa Khusus (Bidikmisi). Universitas Sebelas Maret.

[3] Han, J., Kamber, M., & Pei, J. (2012). Data Mining Concepts And Techniques 3rd Edition (3rd ed.). Morgan Kaufmann.

[4] Sani, K., Winarno, W. W., & Fauziati, S. (2016). Analisis Perbandingan Algoritma Classification Untuk Authentication Uang Kertas (Studi Kasus: Banknote Authentication). Jurnal Informatika, 10(1), 1130–1139.

[5] Özsoy, S., Gümüş, G., & KHALILOV, S. (2015). C4.5 Versus Other Decision Trees: A Review. Computer Engineering and Applications Journal, 4(3), 173–181

[6] Universitas Sebelas Maret. (2009). Peraturan Rektor Universitas Sebelas Maret Nomor 149A / H27/KM/2009 tentang Beasiswa Universitas Sebelas Maret. Universitas Sebelas Maret.

[7] Quinlan, J. R. (1993). C4.5 : Programs For Machine Learning. San Mateo, California: Morgan Kaufmann.

[8] Badan Pusat Statistik. (2016). Indikator Kesejahteraan Rakyat 2016. Badan Pusat Statistik.

[9] Weiss, G. M., & Provost, F. (2003). Learning When Training Data are Costly: The Effect of Class Distribution on Tree Induction. *Journal of Artificial Intelligence Research*, 19, 315–354.

[10] Stubbs, M. (2016). Static ROC Chart. Diambil dari https://bl.ocks.org/micahstubbs/5c3b87ce4bc247340186