3_Data Induk Mahasiswa sebagai Prediktor

by Lastri Suastri

Submission date: 11-Apr-2023 01:21AM (UTC+0700)
Submission ID: 2060768662
File name: 3_Data_Induk_Mahasiswa_sebagai_Prediktor.pdf (495.84K)
Word count: 5079
Character count: 28581

Jurnal SISFOKOM (Sistem Informasi dan Komputer), Volume 10, Nomor 01, PP 71 - 78

Data Induk Mahasiswa sebagai Prediktor Ketepatan Waktu Lulus MenggunakanAlgoritma CART Klasifikasi Data Mining

Arief Jananto[1]*, Sulastri[2], Eko Nur Wahyudi[3], Sunardi[4]
Program Studi Sistem Informasi[1],[2], Program Studi Manajemen Informatika[3],[4]
Fakultas Teknologi Informasi Universitas Stikubank, UNISBANK
Semarang, Indonesia
ajananto09@edu.unisbank.ac.id[1], sulastri@edu.unisbank.ac.id[2], ekowahyudi157@edu.unisbank.ac.id[3], sunardi@edu.unisbank.ac.id[4]

***Abstract*— Fakultas Teknologi Informasi Universitas Stikubank (UNISBANK) as one of the faculties in higher education in implementing learning activities has produced a lot of stored data and has graduated many students. The level of timeliness of graduation is important for study programs as an assessment of success. This research tries to dig up the pile of student parent data and graduation data in order to get the pass rate and graduation prediction of active students. By implementing the classification data mining technique and the CART algorithm, it is hoped that a decision tree can be used to predict the class timeliness of graduating from active students. By using the graduation data and student parent data totaling 1018 records, a decision tree model was obtained with an accuracy rate of 63% from the data testing test. Determination of split nodes using the Gini Index which breaks the dataset based on its impurity value. Tests conducted in this study show that the order of the variables in the decision tree is gender, origin school status, parental education, age at entry, city of birth, parent's occupation. The prediction with the resulting model is that 71% of active S1 Information Systems students can graduate on time and 51% for S1 Informatics Engineering students.**

Keywords*— *Klasifikasi, CART, Gini Index, Data Mining

***Abstrak*— Fakultas Teknologi Informasi Universitas Stikubank (UNISBANK) sebagai salah satu fakultas di perguruan tinggi dalam pelaksanaan kegiatan pembelajaran telah banyak memproduksi banyak data yang disimpan dan telah banyak meluluskan mahasiswa. Tingkat ketepatan waktu lulus menjadi penting bagi program studi sebagai penilaian kesuksesan. Penelitian ini mencoba menggali tumpukan data induk mahasiswa dan data wisuda guna mendapatkan tingkat kelulusan dan prediksi kelulusan dari mahasiswa aktif. Dengan mengimplementasikan teknik data mining klasifikasi dan algoritma CART yang menggunakan teknik partisi biner berulang diharapkan dapat diperoleh sebuah pohon keputusan yang mampu digunakan memprediksi kelas ketepatan waktu lulus dari mahasiswa aktif. Dengan menggunakan data wisuda dan data induk mahasiswa sejumlah 1018 record, diperoleh sebuah model pohon keputusan dengan tingkat akurasi 63% dari uji data testing. Penentuan node split menggunakan Gini Index yang memecah dataset berdasarkan nilai impuritasnya. Uji coba yang dilakukan dalam penelitian ini menunjukan bahwa urutan variabel dalam pohon keputusan adalah jenis kelamin, status sekolah asal, pendidikan orang tua, usia saat masuk, kota kelahiran, pekerjaan orang tua. Prediksi dengan model yang dihasilkan adalah 71% mahasiswa S1 Sistem Informasi aktif dapat lulus tepat waktu dan 51% untuk mahasiswa S1 Teknik Informatika.**

Kata Kunci*— *Klasifikasi, CART, Gini Index, Data Mining

I. PENDAHULUAN

Dengan berlandaskan kebudayaan bangsa Indonesia, perguruan tinggi adalah penyelenggara pendidikan tinggi yang melaksanakan jenjang pendidikan setelah pendidikan menengah dimana dapat berbentuk berbagai program mulai dari diploma hingga program spesialis [1].

Fakultas Teknologi Informasi Universitas Stikubank (UNISBANK) sebagai salah satu fakultas dalam perguruan tinggi dalam pelaksanaan kegiatannya dapat memproduksi data yang cukup besar dan selalu bertambah setiap saat. Mulai dari data mahasiswa regristrasi hingga data kelulusan, yang selama ini masih banyak yang hanya bersifat sebagai data histori atau hanya disimpan saja. Salah satunya yaitu angka kelulusan yang menjadi nilai ukur tingkat kesuksesan program studi dalam perguruan tinggi pada pelaksanaan aktivitas pembelajaran [2].

Setiap mahasiswa terutama program sarjana diharapkan dapat selesai tepat waktu dalam 8 semester atau 4 tahun sesuai dengan kurikulumnya. Sampai saat ini waktu lulus dari mahasiswa di Fakultas Teknologi Informasi masih terdapat mahasiswa yang lulus tidak tepat waktu, hal ini tentunya dapat menjadi masalah tersendiri dan perhatian yang lebih bagi fakultas maupun program studi. Permasalahan ini sebenarnya bisa diantisipasi dengan melakukan prediksi pada mahasiswa aktif dengan mengenali pola dari data kelulusan mahasiswa yang telah lulus. Namun dengan jumlah data yang cukup banyak dan jumlah atribut yang cukup bervariasi sebagai pertimbangan maka terkadang hal ini sering sulit dilakukan.

Kegiatan menemukan pola atau informasi penting di suatu kumpulan data yang sudah ditentukan dengan memakai teknik atau cara spesifik disebut data mining. Adapun teknik, cara maupun prosedur algoritma yang dapat digunakan dalam

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v10i1.991, Copyright ©2020
Submitted : September 13, 2020; Revised : February 11, 2020; Accepted : February 16, 2020; Published : February 22, 2021

penambangan data sangat banyak ragamnya dengan pertimbangan pada tujuan dan kegiatan penambangan pengetahuan dalam basis data [3]. Teknologi data mining dapat digunakan sebagai suatu *tools* eksplorasi kumpulan data yang mampu menghasilkan suatu informasi yang mungkin selama ini tidak diduga.

Pada penelitian sebelumnya disampaikan bahwa untuk memprediksi masa studi dari mahasiswa di Universitas Sam Ratulangi dengan menggunakan algoritma *naïve bayes* dapat didasarkan pada pilihan program studi, angka indek prestasi semester dan jumlah kredit yang diambil tiap semeternya yang merupakan aspek akademik [4].

Dalam penelitian lainnya menyatakan bahwa jurusan dan IPK adalah vaiabel yang paling signifikan dalam membedaan waktu kelulusan mahasiswa antara tepat waktu dan tidak tepat waktu pada Fakultas MIPA Universitas Pattimura. Dengan jumlah sampel sebanyak 585 lulusan periode januari 2010 sampai dengan April 2014. Ketepatan dari pohon klasifikasi yang terbentuk adalah 85,5% [5].

Penggunaan data mining dengan algoritma C4.5 dapat diterapkan untuk memperkirakan 4 jenis kinerja studi mahasiswa berupa lulus dengan waktu cepat, tepat, terlambat maupun drop-out. Dimana faktor akademik paling berpengaruh adalah indek prestasi pada semester [6].

Karakteristik dari tiap atribut pada variable yang mempengaruhi tingkat kelulusan, dengan menggunakan teknik klasifikasi dengan menerapkan metode *Decision Tree* dan algoritma J48 sebagai alat bantu pada weka 3.6.8 bahwa variabel tempat lahir, pekerjaan orang tua, asal sekolah dan jenis kelamin adalah variabel yang menentukan tingkat kelulusan mahasiswa pada jurusan Sistem Informasi Universitas Binadarma Palembang [7].

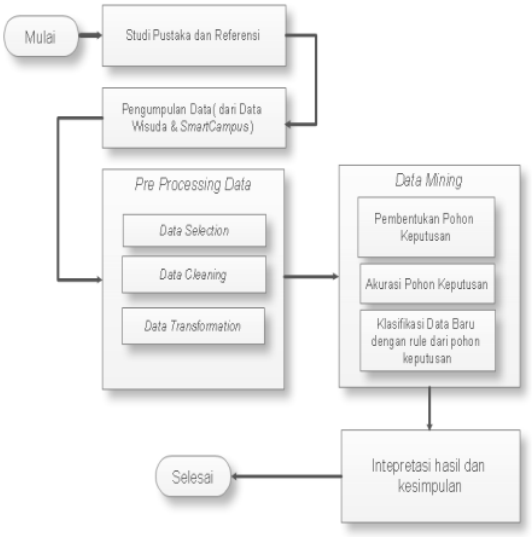
Pada penelitian ini berusaha mengimplementasikan teknik data mining klasifikasi untuk mengetahui urutan atribut dari variabel data induk mahasiswa terhadap ketepatan waktu lulus dengan algoritma CART *(Classifcation And Regression Trees)*. Kemudahan intepretasi, cepat dan akurat serta dapat diimplementasikan pada dataset berukuran besar, field berjumlah besar pada skala field campuran dengan tahapan pembelahan dua sisi (biner) menjadi kelebihan algoritma CART dalam kelompok algoritma klasifikasi [8]. Dengan kelebihan tersebut maka peneliti tertarik untuk mengaplikasikannya dalam penelitian ini. Selanjutnya pohon keputusan atau model yang dihasilkan dapat digunakan untuk mengklasifikasikan dan memprediksi jumlah lulusan yang tepat waktu mapun tidak tepat waktu. Ruang lingkup penelitian adalah dengan menggunakan data kelulusan selama 5 tahun di Fakultas Teknologi Informasi dan data induk mahasiswa bersangkutan serta alat bantu berupa Bahasa R dan *Microsoft Excel*.

Penelitian ini bertujuan untuk menghasilkan model pohon keputusan untuk menemukan urutan atribut klasifikasi dari variabel data induk mahasiswa dalam suatu data kelulusan menggunakan teknik klasifikasi data mining terhadap ketepatan waktu lulus.

Dengan menggunakan model pohon keputusan yang dihasilkan maka program studi dapat membuat suatu prediksi terhadap ketepatan waktu lulus pada data mahasiswa aktif. Diperolehnya informasi mengenai perkiraan ketepatan waktu lulus mahasiswa yang tepat waktu dan tidak tepat waktu, maka dengan demikian bisa segera diambil suatu perlakuan terhadap mahasiswa yang diperkirakan tidak tepat waktu untuk mengantisipasi ketidaktepatan waktu lulusnya. Dalam penelitian ini selain menggunakan data induk tentang profil mahasiswa dan data wisuda tapi juga profil orang tua mahasiswa sehingga diharapkan dengan lebih banyak variasi atribut yang digunakan maka akan diperoleh hasil model yang lebih rinci.

II. METODE PENELITIAN

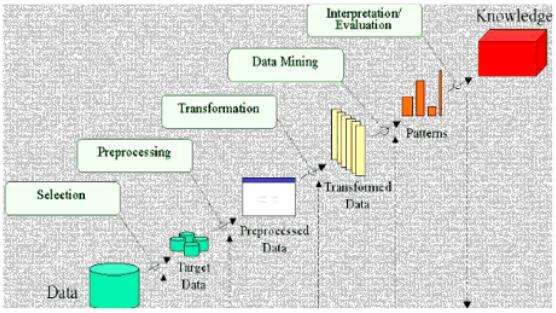
Tahapan yang dilakukan dalam penelitian ini dapat digambarkan seperti tampak pada Gambar 1.



Gambar 1. Tahapan penelitian

Tahapan penelitian dimulai setelah studi pustaka dan referensi yang didalamnya termasuk perumusan masalah hingga survei ke lokasi penelitian. Selanjutnya dilakukan pengumpulan data dari pihak terkait yang kemudian data memasuki tahapan *preprosesing* data. Tahapan berikutnya meliputi data *selection*, data *cleaning/preprosesing* dan data *transformation*. Data yang telah siap kemudian dilakukan proses *data mining* yaitu implementasi dari algoritma CART yang terdiri dari sejumlah langkah didalamnya hingga menghasilkan pohon keputusan. Dari pohon keputusan ini digunakan untuk melakukan prediksi terhadap label kelas dari data baru. Tahapan terakhir adalah memberikan intepretasi hasil yang diperoleh.Tahap-tahap proses aliran informasi dalam data mining dapat diilustrasikan pada Gambar 2.

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v10i1.991, Copyright ©2020
Submitted : September 13, 2020; Revised : February 11, 2020; Accepted : February 16, 2020; Published : February 22, 2021



Gambar 2. Aliran Informasi dalam data mining [9]

A. *Klasifikasi (Classification)*

Klasifikasi adalah bagaimana melaksanakan proses penggerombolan dengan mengacu pada karakeristik tertentu [10]. Proses klasifikasi dalam penelitian ini menggunakan pohon keputusan (*Decision Tree)*.

Menurut Han & Kamber dalam [11], bahwa pohon keputusan adalah diagram alir yang menyerupai struktur pohon, dengan tiap *internal node* sebagai simbol atribut yang dites, memiliki cabang yang melambangkan hasil dari atribut yang diuji dan *leaf node* melambangkan kelas-kelas khusus atau distribusi dari kelas.

B. *Algoritma CART*

Pohon (Tree) yang dibangun menggunakan data latih *(training set)* dapat dimungkinkan untuk memposisikan suatu nilai kelas berdasarkan nilai variable lain maupun independen pada variable target dari record baru. Pembangunan pohon binary *(binary tree)* berdasarkan fungsi variabel input tunggal dengan memecah record di setiap node dikenal sebagai CART [12].

Adapun tahapan algoritma CART [13] terdiri dari :

1. Tahap I, membuat daftar calon cabang semua variabel prediktor secara keseluruhan sehingga menghasilkan daftar calon cabang mutakhir.
2. Tahap II, menilai kinerja keseluruhan calon cabang yang ada pada daftar calon cabang tahap I dengan jalan menghitung nilai besaran kesesuaian *(goodness)*.
3. Tahap III, memilih calon cabang yang akan dijadikan cabang dengan memilih calon cabang yang memiliki nilai kesesuaian *(goodness)* terbesar.

Kemudian buat percabangan, jika tidak ada lagi noktah keputusan, proses algoritma CART selesai namun jika tidak maka proses algoritma dilanjutkan dengan kembali ke tahap II dengan tidak memperhitungkan lagi calon cabang yang telah dijadikan cabang sebelumnya sehingga mendapatkan daftar calon cabang yang baru.

Pada fase pembangunan pohon keputusan sebagai model CART pemilihan *split* atribut memakai nilai *impurity*. Nilai *impurity* pada CART dapat diformulasikan berupa nilai *Gini Index* [14].

C. *Gini Index*

Untuk dataset yang mempunyai *m* kelas, maka G*ini Index* untuk atribut *D* dapat diformulasikan [15], sebagai berikut :

$$Gini(D) = 1 - \sum_{i=1}^{m} p_i^2 \qquad (1)$$

Dimana *m* adalah banyaknya atribut kelas pada dataset, p_i adalah probabilitas suatu record pada dataset memiliki atribut kelas *Ci* dan dihitung dengan membagi banyaknya atribut kelas *Ci* pada dataset terhadap jumlah record dataset.

Pembelahan dataset menjadi 2 bagian dengan *Gini Index* terendah. Jika data dipilah menjadi 2 subset yaitu *D1* dan *D2*, *Gini Index* dapat dirumuskan [15] sebagai berikut :

$$Gini_A = \frac{|D_1|}{|D|} Gini(D_1) + \frac{|D_2|}{|D|} Gini(D_2) \qquad (2)$$

III. HASIL DAN PEMBAHASAN

A. *Persiapan Data*

Data induk mahasiswa pada tahap persiapan diambil dari bagian Biro Administrasi Akademik Fakultas yang telah lulus pada data wisuda berupa nim, nama mahasiswa, jenis kelamin, kota lahir, tanggal lahir, nama sekolah asal selama 5 tahun yang terdistribusi dalam setiap semester. Dari proses ini diperoleh data sejumlah 1151 record seperti tampak pada tabel I.

TABEL I. JUMLAH DATA WISUDAWAN

No	Tahun	Periode Wisdua	Jumlah Wisudawan FTI	Jumlah Wisudawan SI & TI
1	I-Ganjil	67	112	102
2	I-Genap	68	168	150
3	II-Ganjil	69	121	105
4	II-Genap	70	178	172
5	III-Ganjil	71	89	88
6	III-Genap	72	149	143
7	IV-Ganjil	73	92	82
8	IV-Genap	74	138	120
9	V-Ganjil	75	78	76
10	V-Genap	76	135	112
		Total	1260	1151

Data tersebut di atas digabungkan menjadi satu tabel dengan mengambil hanya pada beberapa *field* yang digunakan yaitu : NIM, Nama Mahasiswa, Jenis Kelamin, Tempat Lahir, Tanggal Lahir, Nama Sekolah Asal, Tanggal Lulus, Periode Wisuda.

B. *Pembersihan data (Cleanning)*

Permasalahan data yang ada adalah tidak diisi atau diisi dengan nilai 0, usia saat masuk diatas 80 tahun dan dibawah 16 tahun serta masa studi dibawah 3,4 tahun.

Proses pembersihan data yang bermasalah pada penelitian ini dilakukan dengan menghilangkan data dari tabel data utama.

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v10i1.991, Copyright ©2020
Submitted : September 13, 2020; Revised : February 11, 2020; Accepted : February 16, 2020; Published : February 22, 2021

C. Transformasi Data

Proses transformasi berupa diskritisasi menjadi data kategorikal digunakan untuk lebih memudahkan proses mining selanjutnya.

1. Kota kelahiran ditransformasi ke dalam 3 kategori, yaitu untuk kota kelahiran (KK) Semarang dan Kab Semarang menjadi Dalam Kota (DK), kota kelahiran diluar Semarang namun masih dalam pulau Jawa maka dikonversi menjadi Luar Kota (LK) sedangan yang diluar Semarang dan diluar pulau jawa sebagai Luar Pulau (LP).
2. Jenis kelamin (JK) dari data set tidak dilakukan konversi. P untuk jenis kelamin putri dan L untuk jenis kelamin laki-laki.
3. Kategori sekolah asal (STA) dikelompokan menjadi dua nilai yaitu N untuk sekolah berstatus negeri dan S untuk sekolah berstatus swasta.
4. Usia Saat Masuk (USM) Kuliah adalah usia mahasiswa saat masuk atau terdaftar sebagai mahasiswa di Fakultas Teknologi Informasi. Pada data usia saat masuk kuliah ini hanya dibagi dalam 2 kategori yaitu kurang dari sama dengan 19 Tahun (<=19) dan lebih besar dari 19 Tahun (>19).
5. Pendidikan Orang Tua (PDO) adalah pendidikan orang tua mahasiswa saat siswa tersebut terdaftar sebagai mahasiswa. Untuk data pendidikan orang tua dikelompokan menjadi 4 kategori yaitu S0 untuk pendidikan orang tua yang Tidak Tamat SD, Tamat SD, Tamat SMP, Tamat SMP, Tamat SMA. Kategori S1 untuk pendidikan orang tua D3 atau S1. Kategori S2 untuk pendidikan orang tua S2 dan S3 untuk penddidikan S3.
6. Pekerjaan orang tua (PKO) adalah pekerjaan orang tua siswa saat siswa tersebut terdaftar sebagai mahasiswa. Untuk data pendidikan orang tua dikelompokan menjadi 2 kategori yaitu Masih Bekerja (MB) untuk (PNS, ABRI, Pegawai Swasta, Usaha Sendiri) dan Tidak Bekerja (TB) untuk (Tidak Bekerja, Pensiunan, Lain-lain).
7. Sedangkan untuk masa studi (SL) yang merupakan variabel yang akan digunakan sebagai variabel terikat menjadi nilai kelas dari setiap record data tersebut. Data masa studi dikonversikan menjadi dua kelas yaitu Tepat Waktu (Tepat) untuk masa studi dibawah 5 Tahun dan sisanya diatas 5 Tahun dianggap Tidak Tepat Waktu (Tidak).

Pada tabel II ditampilkan contoh dari sebagian perolehan dataset sejumlah 1018 record yang siap untuk proses datamining selanjutnya.

TABEL II. CONTOH DATA SIAP PROSES DATAMINING

Kota Kelahiran	Jenik Kelamin	Status Sekolah Asal	Usia Saat Masuk	Pendidikan Orang Tua	Pekerjaan Orang Tua	Status Masa Studi
(KK)	(JK)	(STA)	(USM)	(PDO)	(PKO)	(SL)
LK	P	S	>19	S0	MB	Tidak
DK	P	S	<=19	S1	MB	Tidak
LK	P	S	>19	S0	TB	Tidak
DK	L	N	<=19	S0	TB	Tidak
LK	L	N	<=19	S0	MB	Tidak
LK	L	S	<=19	S0	MB	Tidak
LK	L	N	>19	S0	TB	Tidak
DK	L	S	>19	S0	TB	Tidak
DK	L	S	<=19	S1	MB	Tidak
LK	L	S	<=19	S0	MB	Tidak

D. Pembentukan Pohon Keputusan

Data latih *(training dataset)* dipakai sebagai *set* data untuk membangun pohon keputusan. Pemeriksaan model dilakukan dengan menguji model memakai data tes *(test dataset)* [16].

Test dataset digunakan untuk mengukur tingkat ketepatan dari model *(classifier)*, dengan menghitung presentase klasifikasi benarnya. Jika tingkat akurasi model dapat diterima maka model dapat digunakan pada klasifikasi data baru.

Variabel dalam dataset terdiri atas variabel bebas yang meliputi 6 variabel yaitu kota kelahiran, jenis kelamin, status sekolah asal, usia saat masuk kuliah, pendidikan orang tua dan pekerjaan orang tua. Sedangkan variabel terikat dalam penelitian ini adalah ketepatan waktu lulus yang terdiri dari dua kategori yaitu tepat waktu (tepat) dan tidak tepat waktu (tidak).

Jumlah data awal sebelum *preprosesing* adalah sebanyak 1151 record yang kemudian dilakukan *preprosesing* menjadi 1018 record.

Pembentukan pohon keputusan dengan menggunakan algoritma CART pada dasarnya terdiri dari 3 tahap. Berikut langkah-langkah algoritma CART dalam pembentukan pohon keputusan dimana digunakan percobaan pertama dengan proporsi 70% data *training* dan 30% data *testing*.

1) Menyusun Calon Cabang (Candidat Split)

Penyusunan calon cabang dilakukan terhadap seluruh variabel prediktor. Daftar calon cabang mutakhir dapat dilihat pada Tabel III.

TABEL III. DAFTAR CALON CABANG

Calon Cabang		
Nomor	**Kiri**	**Kanan**
1	KK=DK	KK=LK,LP
2	KK=LK	KK=DK,LP
3	KK=LP	KK=DK,LK
4	JK=P	JK=L
5	STA=N	STA=S
6	USM<=19	USM>19
7	PDO=S0	PDO=S1,S2,S3
8	PDO=S1	PDO=S0,S2,S3
9	PDO=S2	PDO=S0,S1,S3
10	PDO=S3	PDO=S0,S1,S2
11	PKO=MB	PKO=TB

2) Menilai Kinerja Keseluruhan Calon Cabang

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v10i1.991, Copyright ©2020
Submitted : September 13, 2020; Revised : February 11, 2020; Accepted : February 16, 2020; Published : February 22, 2021

Perhitungan kinerja setiap calon cabang (Tabel III) harus dilakukan terlebih dahulu sebagai langkah kedua. Hasil tahap ini akan digunakan untuk penentuan atribut pembelahan dataset dengan menggunakan homogenitas yang tinggi.

a) Menghitung Gini Index Dataset

Sebelum digunakan membangun pohon keputusan maka tingkat homogenitas kelas atribut harus diperhitungkan terlebih dahulu menggunakan *Gini Index*. Jika diperoleh nilai 0 hal ini berarti dataset sudah sampai pada terminal node yang artinya dataset tidak perlu dibelah lagi.

Pada dataset yang digunakan dalam penelitian ini terdiri dari 711 record, dengan dua nilai atribut kelas Tepat dan Tidak. Jumlah record atribut kelas Tepat adalah sebanyak 400 record dan 311 record untuk kelas Tidak. *Gini Index* dataset dapat diperoleh sebagai berikut :

Gini(D) = 1 – (400/711)^2 – (311/711)^2 = 0.492166

b) Menghitung Gini Index untuk masing-masing atribut.

Untuk atribut Kota Kelahiran (KK) mempunyai 3 nilai yaitu Dalam Kota (DK), Luar Kota (LK) dan Luar Pulau(LP). Dengan demikian maka atribut KK kemungkinan mempunyai subset sebanyak (2^3-2)/2=(8-2)/2=3, dimana subset yang mungkin adalah {(DK),(LK,LP)},{(LK),(DK,LP)} dan {(LP),(DK,LK)}. Untuk subset {(DK),(LK,LP)} dibagi menjadi partisi *D1* untuk subset DK dan partisi *D2* untuk subset (LK,LP). Ada 316 record memenuhi kondisi *D1* dan *D2* sebanyak 395 record yang memenuhi. Dimana nilai *Gini Index* dapat disajikan perhitungannya sebagai berikut :

Ginikk(DK)(D) = (316/711)*Gini(D1)* + (395/711)**Gini(D2)*

Ginikk(DK) (D) = (316/711)*(1-(177/316)^2-(139/316)^2) + (395/711)*(1-(223/395)^2-(172/395)^2)

Ginikk(DK) (D) = 0.492156

Tabel IV menunjukan daftar nilai *Gini Index* untuk masing-masing atribut dengan setiap subset kemungkinannya.

TABEL IV. NILAI *GINI INDEX* ATRIBUT PREDIKTOR

Atribut	D1	D2	*Gini Index*
KK	KK=DK	KK=LK,LP	0.492155815
	KK=LK	KK=DK,LP	0.491916182
	KK=LP	KK=DK,LK	0.490310986
JK	***JK=P***	***JK=L***	***0.447179670***
STA	STA=N	STA=S	0.483679762
USM	USM<=19	USM>19	0.491252023
PDO	PDO=S0	PDO=S1,S2,S3	0.491542709
	PDO=S1	PDO=S0,S2,S3	0.491469827
	PDO=S2	PDO=S0,S1,S3	0.491793581
	PDO=S3	PDO=S0,S1,S2	0.491059865
PKO	PKO=MB	PKO=TB	0.492129697

c) Menentukan Calon Cabang yang Menjadi Cabang

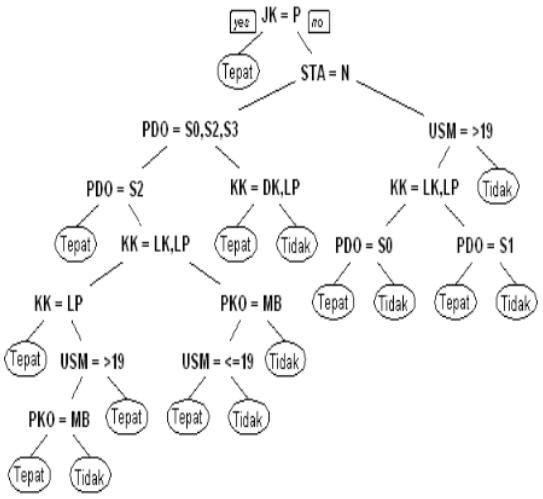
Dengan melihat perhitungan nilai Gini index dari semua atribut dengan semua kemungkinan subset yang terbentuk, untuk iterasi-1 adalah atribut Jenis Kelamin (JK) dengan Gini Index terendah yaitu 0.447179670. Dengan demikian untuk iterasi-1 sebagai atribut pembelah adalah Jenis Kelamin (JK) dengan subset JK=P dan JK=L.

Selanjutnya untuk iterasi-2 dan seterusnya dilakukan hal yang sama tetapi dengan membuang atribut (tidak mengikutkan kembali dalam perhitungan Gini Index) untuk atribut yang telah menjadi node split atau pembelah dataset.

Pembentukan pohon keputusan dibantu dengan menggunakan perangkat lunak Bahasa R dengan memanfaatan fungsi package rpart.

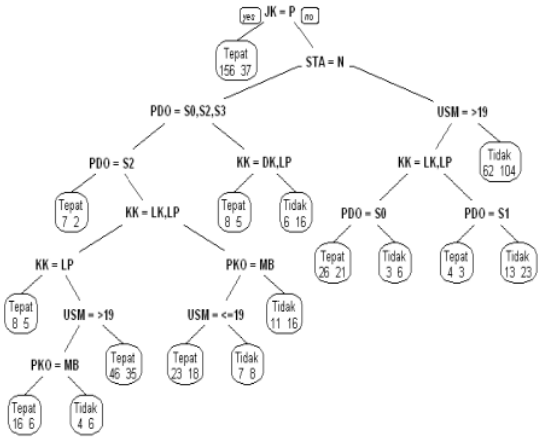
Adapun tahapan klasifikasi pembentuk pohon keputusan dengan menggunakan fungsi *rpart* adalah:

- Memuat(load) data sampel
  Data1 <- read.csv("~/cobar7/Data1.csv")
- Membagi data sampel menjadi 2 subset
  ind<- sample(2,nrow(Data1),replace=TRUE, prob=c(0.70,0.30))
  train.data <- Data1[ind == 1,]
  test.data <- Data1[ind == 2,]
- Memanggil/mengaktifkan library/package rpart
  library(rpart)
- Pembentukan pohon keputusan dengan mode spliting gini index
  pohontrain<- rpart(SL~KK+JK+STA+USM+PDO+PKO,train.data, parms=list(split="gini"),cp=0.0003)
- Menampilkan summary
  summary(pohontrain)
- Menampilkan pohon keputusan dalam bentuk split node
  print(pohontrain)
- Menampilkan pohon keputusan yang terbentuk dengan menggunakan fungsi **plot grafik dari rpart.**
  library(rpart.plot)
  prp(pohontrain,faclen=9,cex=0.8,extra=0)



Gambar 3. Pohon keputusan dengan ujung berupa nama kelas

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v10i1.991, Copyright ©2020
Submitted : September 13, 2020; Revised : February 11, 2020; Accepted : February 16, 2020; Published : February 22, 2021



Gambar 4. Pohon keputusan dengan Jumlah data

- Melakukan dan menampilkan uji akurasi model pohon keputusan *pohontest* <-
  predict(pohontrain,test.data,type="class")
  library(caret)
  confusionMatrix(pohontest,test.data$SL)

  Hasil eksekusi :

```
        Confusion Matrix and Statistics
            Reference
Prediction Tepat Tidak
     Tepat  121    61
     Tidak   52    73
Accuracy : 0.6319
                95% CI : (0.5753, 0.686)
Sensitivity : 0.6994
Specificity : 0.5448
```

E. Uji Akurasi Pohon Keputusan

Pada penelitian ini dilakukan 5 kali percobaan dengan mengubah proporsi jumlah data training dan data testing dari dataset yang ada. Dari hasil uji dengan mengubah proporsi data training dan testing maka diperoleh tingkat akurasi model pohon keputusan seperti tampak pada table V.

TABEL V. NILAI AKURASI MODEL BERDASAR DATA *TESTING*

Percobaan Ke	Jumlah Data			*acurracy*
	Probabilitas	*Training*	*Testing*	
I	70 : 30	711	307	0.6319
II	75 : 25	755	263	0.6236
III	80 : 20	806	212	0.6274
IV	85 : 15	856	162	0.6235
V	90 : 10	896	122	0.6230

Dengan melihat tingkat akurasi model melalui *confusion matrix* pada tabel V, maka pada proporsi data *training* dan *testing* 70 : 30 adalah yang tertinggi yaitu sebesar 63%.

Berdasarkan pohon keputusan yang terbentuk terlihat bahwa atribut jenis kelamin (JK) masih menduduki peringkat pertama dalam pemecahan(*split*) data berkaitan dengan ketepatan waktu lulus. Atribut JK ="P" menunjukkan kelulusan tepat waktu dengan tingkat akurasi record data sebanyak 156 Tepat (81%) dan 37 Tidak (19%). Dari proses pemecahan data (*split*) pada pohon keputusan disimpulkan menjadi daun berlabel Tepat. Sedangkan untuk JK="L" atau untuk mahasiswa berjenis kelamin Laki-Laki, ketepatan waktu lulus masih dipengaruhi oleh atribut prediktor lainnya.

Urutan *node* selanjutnya adalah Status sekolah Asal (STA) yaitu apakah status sekolah asal adalah Negeri (N) atau Swasta (S). Atribut ini mempengaruhi tingkat ketepatan waktu lulus pada kelompok jenis kelamin "L", dimana untuk sekolah asal bersatus "N" akan dipengaruhi lagi oleh faktor pendidikan orang tua (PDO), Kota Kelahiran (KK), Pekerjaan Orang Tua (PKO) dan usia saat masuk (USM). Sedangkan untuk sekolah yang berstatus "S" maka tingkat ketepatan waktu lulus dipengaruhi oleh usia saat masuk (USM), dimana untuk yang berusia saat masuk > 19 Tahun lebih cenderung banyak tidak tepat waktu kelulusannya (63%). Untuk status sekolah asa "S" dan usia saat masuk <=19 selanjutnya dipengaruhi oleh kota kelahiran (KK) yang selanjutnya untuk kota kelahiran (KK) dalam kota (DK) dan pendidikan orang tuanya (PDO) adalah "S0" atau non perguruan tinggi mempunyai tingkat ketepatan waktu lulus mencapai 55%. Sedangkan yang kota lahir dalam kota dan pendidikan orang tua (PDO) nya selain "S0" mempunyai tingkat ketepatan waktu lulus yang rendah dan lebih banyak yang tidak tepat waktu (67%) . Untuk yang kota kelahirannya (KK) luar kota maupun luar pulau pada pendidikan orang tua yang "S1" mempunyai jumlah ketepatan waktu lulus mencapai 57% sedangkan yang berpendidikan selain "S1" mempunyai tingkat ketepatan waktu lulus yang lebih rendah yaitu hanya 36%.

Secara umum bila dilihat dari pohon keputusan yang terbentuk, peringkat dari masing-masing variabel data induk mahasiswa (atribut prediktor) terhadap ketepatan waktu lulus adalah jenis kelamin (JK) – (peringkat I), status sekolah asal (STA) – (peringkat II), pendidikan orang tua (PDO) – (peringkat III.1), usia saat masuk (USM) – (peringkat III.2), kota kelahiran (KK) – (peringkat IV), pekerjaan orang tua (PKO) – (peringkat V).

Dari model pohon keputusan yang dihasilkan berdasarkan percobaan yang ke I yaitu pada proporsi 70% data *training* dan 30% data *testing*, diperoleh aturan klasifikasi sebagai berikut:

1. IF JK=P THEN SL = Tepat
2. IF JK=L AND STA=N AND PDO=S2 THEN SL = Tepat
3. IF JK=L AND STA=N AND (PDO=S0 OR S3) AND KK=LP THEN SL = Tepat
4. IF JK=L AND STA=N AND (PDO=S0 OR S3) AND KK=LK AND USM>19 AND PKO=MB THEN SL = Tepat
5. IF JK=L AND STA=N AND (PDO=S0 OR S3) AND KK=LK AND USM>19 AND PKO=TB THEN SL = Tidak

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v10i1.991, Copyright ©2020
Submitted : September 13, 2020; Revised : February 11, 2020; Accepted : February 16, 2020; Published : February 22, 2021

6. IF JK=L AND STA=N AND (PDO=S0 OR S3) AND KK=LK AND USM<=19 THEN SL = Tepat
7. IF JK=L AND STA=N AND (PDO=S0 OR S3) AND KK=DK AND PKO=MB AND USM<=19 THEN SL = Tepat
8. IF JK=L AND STA=N AND (PDO=S0 OR S3) AND KK=DK AND PKO=MB AND USM>19 THEN SL = Tidak
9. IF JK=L AND STA=N AND (PDO=S0 OR S3) AND KK=DK AND PKO=TB THEN SL = Tidak
10. IF JK=L AND STA=N AND PDO=S1 AND (KK=DK OR KK=LP) THEN SL = Tepat
11. IF JK=L AND STA=N AND PDO=S1 AND KK=LK THEN SL = Tidak
12. IF JK=L AND STA=S AND USM>19 AND (KK=LK OR LP) AND PDO=S0 THEN SL = Tepat
13. IF JK=L AND STA=S AND USM>19 AND (KK=LK OR LP) AND PDO=S1 THEN SL = Tidak
14. IF JK=L AND STA=S AND USM>19 AND (KK=DK OR LP) AND PDO=S1 THEN SL = Tepat
15. IF JK=L AND STA=S AND USM>19 AND (KK=DK OR LP) AND (PDO=S0 OR S2) THEN SL = Tidak

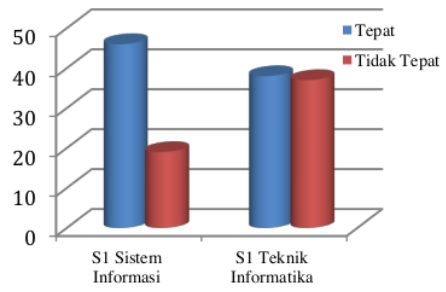
IF JK=L AND STA=S AND USM<=19 THEN SL = Tidak

F. Klasifikasi Data Baru

Aturan klasifikasi yang terbentuk kemudian digunakan untuk memprediksi atau memberi label kelas pada data mahasiswa aktif. Berdasarkan data dari SmartCampus untuk mahasiswa S1 Sistem Informasi yang aktif adalah sebanyak 65 Mahasiswa dan untuk S1 Teknik Informatika sebanyak 75 Mahasiswa sehingga total 140 Mahasiswa. Hasil prediksi kelas dari data uji coba didapatkan hasil prediksi seperti tampak pada tabel VI.

TABEL VI. HASIL PREDIKSI MAHASISWA AKTIF

Progdi	Jumlah	Tepat	Tidak Tepat
S1 Sistem Informasi	65	46	19
S1 Teknik Informatika	75	38	37
	140		



Gambar 5. Grafik Prediksi Ketepatan Waktu Lulus Mahasiswa Aktif

Dari hasil prediksi terlihat untuk program studi S1 Teknik Informatika memiliki jumlah mahasiswa yang tidak tepat waktu cukup besar dibandingkan dengan program studi S1 Sistem Informasi.

Dengan mengetahui prediksi ini diharapkan program studi dapat melakukan suatu tindakan prefentif khususnya terhadap mahasiswa yang diprediksi tidak tepat waktu. Sehingga dapat mengurangi tingkat keterlambatan dari kelulusan mahasiswa bersangkutan.

IV. PENUTUP

Adapun dari serangkaian uji coba yang telah dilakukan maka dapat disimpulkan beberapa hal yakni dalam data kelulusan/wisuda, terdapat informasi bahwa variabel data induk mahasiswa dapat digunakan untuk mengklasifikasikan terhadap pencapaian ketepatan waktu lulus. Jenis Kelamin menempati peringkat pertama yang kemudian diikuti berturut-turut status sekolah asal, pendidikan orang tua, usia saat masuk, kota kelahiran, pekerjaan orang tua. Dengan jumlah data 711 untuk *training set* dan 307 *test set*, dapat dibangun pohon keputusan klasifikasi menggunakan algoritma CART dan *Gini Index* untuk *node spliting* dengan 15 *node* serta mempunyai tingkat akurasi tertinggi 63% pada proporsi data *training* 70% dan data *testing* 30%. Model pohon keputusan yang sudah diubah ke dalam bentuk *rule* dapat digunakan untuk memprediksi kelas dari data baru mahasiswa aktif S1-SI dan S1-TI dengan hasil prediksi 71% mahasiswa S1 Sistem Informasi dapat lulus tepat waktu dan 51% untuk mahasiswa S1 Teknik Informatika.

REFERENSI

[1] Pemerintah-Indonesia, "Undang-Undang RI No 12 Tahun 2012 Tentang Pendidikan Tinggi." 2012.

[2] I. Farida and S. W. H. L. Hendric, "Prediksi Pola Kelulusan Mahasiswa Menggunakan Teknik Data Mining Classification Emerging Pattern," Petir, vol. 12, no. 1, pp. 1–17, 2019, doi: 10.33322/petir.v12i1.414.

[3] E. T. Kursini, Luthfi, Algoritma Data Mining. Yogyakarta: Andi Offset, 2009.

[4] M. winny Amelia, A. S. . Lumenta, and A. Jacobus, "Prediksi Masa Studi Mahasiswa dengan Menggunakan Algoritma Naïve Bayes," J. Tek. Inform., vol. 11, no. 1, 2017, doi: 10.35793/jti.11.1.2017.17652.

[5] M. Fendjalang, "Klasifikasi Variabel Penentu Kelulusan Mahasiswa FMIPA Unpatti Menggunakan Metode CHAID," Statistika, vol. 15, no. 1, pp. 1–6, 2015.

[6] D. H. Kamagi and S. Hansun, "Implementasi Data Mining dengan Algoritma C4.5 untuk Memprediksi Tingkat Kelulusan Mahasiswa," J. Ultim., vol. 6, no. 1, pp. 15–20, 2014, doi: 10.31937/ti.v6i1.327.

[7] S. M. Andri, Yesi Novaria Kunang, "Implementasi Teknik Data Mining Untuk Memprediksi Tingkat Kelulusan," vol. 2013, no. June 2016, pp. 56–63, 2013.

[8] E. T. Novalyn, G. Ginting, and H. K. Siburian, "Omset Pakaian Pria Remaja (Studi Kasus : Pt . Matahari Department Store Thamrin Plaza Medan)," vol. 17, pp. 436–443, 2018.

[9] F. Nasari and S. Darma, "PENERAPAN K-MEANS CLUSTERING PADA DATA PENERIMAAN MAHASISWA BARU (STUDI KASUS: UNIVERSITAS POTENSI UTAMA)," in Seminar Nasional Teknologi Informasi dan Multimedia, 2015, pp. 6–8.

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v10i1.991, Copyright ©2020
Submitted : September 13, 2020; Revised : February 11, 2020; Accepted : February 16, 2020; Published : February 22, 2021

[10] N. Wijaya and A. Ridwan, "Klasifikasi Jenis Buah Apel Dengan," Sisfokom, vol. 08, no. 1, pp. 74–78, 2019.

[11] A. Shiddiq, R. K. Niswatin, and I. N. Farida, "Analisa Kepuasan Konsumen Menggunakan Klasifikasi Decision Tree Di Restoran Dapur Solo (Cabang Kediri)," Gener. J., vol. 2, no. 1, p. 9, 2018, doi: 10.29407/gj.v2i1.12051.

[12] Y. Y. W, "Perbandingan Performansi Algoritma Decision Tree C5 . 0 , Cart ," Seminar, vol. 2007, no. Snati, pp. 0–3, 2007.

[13] N. Nafi'iyah, "Algoritma Cart Dalam Penentuan Pohon Keputusan," J. SPIRIT, vol. 7, no. 2, 2015.

[14] R. W. Ningrat and B. Santosa, "Pemilihan Diet Nutrien bagi Penderita Hipertensi Menggunakan Metode Klasifikasi Decision Tree (Studi Kasus: RSUD Syarifah Ambami Rato Ebu Bangkalan)," J. Tek. Its, vol. VOL.1 , no. 1, pp. 536–539, 2012.

[15] M. Yusa, E. Utami, and E. Luthfi. Taufiq, "Evaluasi Performa Algoritma Klasifikasi Decision Tree ID3, C4.5, dan CART Pada Dataset Readmisi Pasien Diabetes," InfoSys J., vol. 4, no. 1, pp. 23–34, 2016.

[16] S. Hendrian, "Algoritma Klasifikasi Data Mining Untuk Memprediksi Siswa Dalam Memperoleh Bantuan Dana Pendidikan," Fakt. Exacta, vol. 11, no. 3, pp. 266–274, 2018, doi: 10.30998/faktorexacta.v11i3.2777.

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v10i1.991, Copyright ©2020
Submitted : September 13, 2020; Revised : February 11, 2020; Accepted : February 16, 2020; Published : February 22, 2021

3_Data Induk Mahasiswa sebagai Prediktor

ORIGINALITY REPORT

9%	10%	5%	4%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

MATCH ALL SOURCES (ONLY SELECTED SOURCE PRINTED)

14%

★ www.researchgate.net

Internet Source

Exclude quotes	On	Exclude matches	< 2%
Exclude bibliography	On		