REPOSITOR, Vol. 1, No. 2, Desember 2019, Pp. 131-140
ISSN : 2714-7975
E-ISSN : 2716-1382

Penerapan Algoritma C5.0 Pada Analisis Faktor-Faktor Pengaruh Kelulusan Tepat Waktu Mahasiswa Teknik Informatika Universitas Muhammadiyah Malang

Vinna Rahmayanti Setyaning Nastiti[*1], Yufis Azhar[2], Andriani Eka Pramudita[3]
[1,2,3]Teknik Informatika/Universitas Muhammadiyah Malang
vinastiti@umm.ac.id[*1], yufis@umm.ac.id[2], andrianiekaprmdt@gmail.com[3]

Abstrak

Kelulusan tepat waktu mahasiswa merupakan salah satu permasalahan yang sulit untuk diatasi oleh setiap pihak perguruan tinggi, begitu pula pada jurusan Teknik Informatika Universitas Muhammadiyah Malang. Permasalahan ini harus segera diatasi mengingat kualitas mahasiswa akan mempengaruhi sebuah akreditasi perguruan tinggi maupun jurusan. Oleh karena itu, perlu dilakukan analisis faktor-faktor pengaruh kelulusan tepat waktu mahasiswa Teknik Informatika UMM. Penelitian ini menggunakan algoritma C5.0 untuk melakukan seleksi fitur penting dan analisis regresi untuk melakukan estimasi peluang kelulusan tepat waktu mahasiswa. Variabel bebas yang digunakan adalah jenis kelamin, asal daerah, status masuk, SKS semester 4, SKS semester 6, IP semester 2, IP semester 4, IP semester 6, IPK semester 2, IPK semester 4, IPK semester 6, jenis SMA, status SMA, pendidikan orang tua, dan pekerjaan orang tua. Hasil implementasi algoritma C5.0 pada penelitian ini mampu melakukan seleksi fitur dengan menghasilkan 8 dari total keseluruhan 15 fitur dengan nilai akurasi yang lebih baik dibandingkan nilai akurasi yang menggunakan keseluruhan fitur. Serta, penelitian ini mampu memberikan model regresi dengan nilai akurasi sebesar 82%.

Kata Kunci: *Algoritma C5.0, Analisis Regresi, Kelulusan Tepat Waktu*

Abstract

Timely graduation of college students is one of the problems that is difficult to overcome by each college, as well as in the Department of Informatics, University of Muhammadiyah Malang. This problem must be resolved immediately, considering the quality of students will affect the accreditation of university and its majors. So, it is necessary to analyze the factors that influence the timely graduation of Informatics Engineering students in UMM. This study uses the C5.0 algorithm to do feature selection and regression analysis to estimate the opportunities of timely graduation. The independent variables used are gender, regional origin, entry status, academic credit system in 4th semester, academic credit system in 6th semester, grade point of 2nd semester, grade point of 4th semester, grade point of 6th semester, grade point average of 2nd semester, grade point average of 4th semester, grade point average of 6th semester, type of senior high school, status of senior high school, parent's education, and parent's job. The results of the implementation of the C5.0 algorithm in this study were able to do feature selection by producing 8 out of total 15 features with better accuracy than the value of accuracy using all features. And this study is able to provide a regression model with an accuracy value of 82%.

Keywords*: C5.0 Algorithm, Regression Analysis, Timely Graduation*

1. Pendahuluan

Seiring dengan meningkatnya jumlah peminat pendidikan di jenjang perguruan tinggi, maka setiap perguruan tinggi sudah seharusnya memiliki tindakan khusus dalam menyeimbangkan antara jumlah mahasiswa yang masuk dengan jumlah mahasiswa yang keluar. Salah satu cara dalam menyikapi hal tersebut adalah dengan mengontrol ketepatan waktu kelulusan mahasiswa.

Jurusan Teknik Informatika merupakan salah satu jurusan di Universitas Muhammadiyah Malang yang memiliki tingkat kelulusan tepat waktu rendah, yaitu sekitar kurang dari 15% yang lulus tepat waktu per angkatan. Rata-rata, mahasiswa jurusan teknik informatika lulus di semester 9, yang artinya mahasiswa tersebut menyelesaikan studi dengan waktu 4.5 tahun dan telah dikatakan lulus tidak tepat waktu. Hal ini telah menjadi perhatian lebih bagi pihak jurusan. Tetapi,

ternyata pihak jurusan salah menduga akar dari masalah kelulusan tidak tepat waktu ini, yaitu lamanya pengerjaan tugas akhir mahasiswa. Oleh karena itu, dilakukan penelitian untuk mengetahui faktor-faktor utama yang mempengaruhi kelulusan tepat waktu mahasiswa jurusan Teknik Informatika Universitas Muhammadiyah Malang.

Berdasarkan penelitian-penelitian sebelumnya, kelulusan tepat waktu dapat dipengaruhi oleh berbagai macam faktor. Seperti pada penelitian Risnawati [1] yang melakukan analisis faktor-faktor kelulusan mahasiswa dengan menggunakan 4 atribut, yaitu IPK, prestasi, etika, dan SKS. Penelitian Suniantara dan Rusli [2] melakukan klasifikasi variabel-variabel yang mempengaruhi lama studi mahasiswa 7 atribut, yaitu status kelulusan mahasiswa, jenis kelamin, program studi, lama skripsi, IPK, IP semester 6, serta nilai ujian masuk. Penelitian Suniantara [3] melakukan klasifikasi faktor-faktor yang mempengaruhi lama studi mahasiswa dengan 5 atribut, yaitu asal daerah, jurusan, IPK, lama penyusunan skripsi, dan jenis kelamin. Sedangkan penelitian Rizki [4] melakukan analisis survival faktor-faktor yang mempengaruhi lama studi mahasiswa menggunakan 7 atribut, yaitu jenis kelamin, IPK, asal daerah, penghasilan orang tua, jalur masuk, pekerjaan orang tua, dan status sekolah menengah atas.

Berdasarkan berbagai macam faktor pengaruh kelulusan tepat waktu yang digunakan pada penelitian-penelitian sebelumnya, maka dalam penelitian ini menggunakan kombinasi beberapa faktor tersebut untuk dijadikan sebagai atribut penelitian ini dalam menyelesaikan masalah ketepatan waktu kelulusan mahasisiswa Teknik Informatika Universitas Muhammadiyah Malang dengan memanfaatkan Algoritma C5.0 dan Analisis Regresi. Algoritma C5.0 digunakan untuk menentukan faktor-faktor yang paling berpengaruh terhadap kelulusan tepat waktu mahasiswa dengan melakukan seleksi fitur dan Analisis Regresi digunakan untuk melakukan estimasi peluang kelulusan tepat waktu mahasiswa.

2. Metode Penelitian

2.1 Studi literatur

Pada tahap studi literatur dilakukan pemahaman konsep algoritma yang digunakan dalam penelitian ini. Literatur yang digunakan adalah berupa buku dan jurnal yang membahas tentang klasifikasi data mining, decision tree, algoritma C5.0, analisis regresi, dan faktor-faktor yang mempengaruhi lama studi mahasiswa.

2.2 Pengumpulan data

Pada tahap pengumpulan data dilakukan pengajuan permintaan data yang diperlukan untuk penelitian ini, yaitu data alumni mahasiswa jurusan Teknik Informatika Universitas Muhammadiyah Malang angkatan 2011 hingga 2014 kepada kantor jurusan. Data yang didapatkan berjumlah 558 data mahasiswa dengan rincian 65 mahasiswa lulus tepat waktu dan 493 mahasiswa lulus tidak tepat waktu. Gambaran data penelitian dapat dilihat pada Gambar 1 berikut.

Gambar 1. Gambaran Data Penelitian

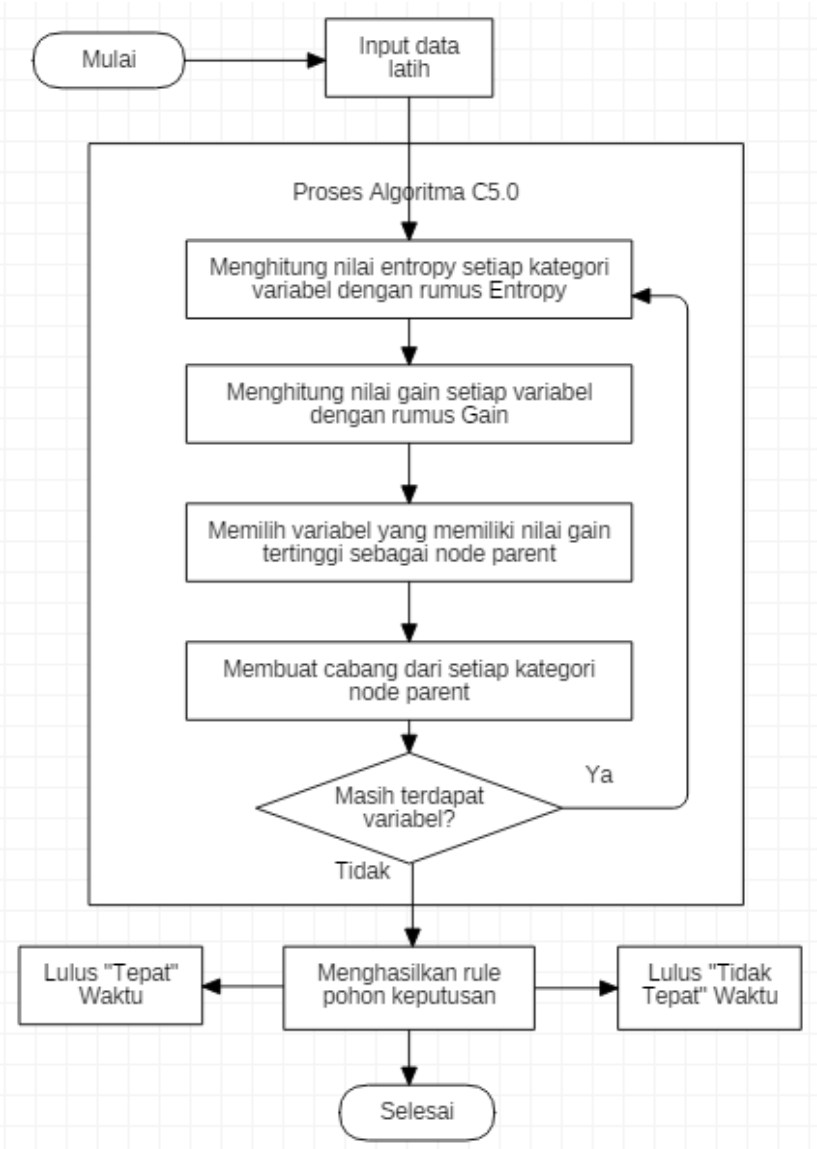
NIM	SEX	ID ASAL	STATUS MASUK	SKS_S4	SKS_S6	IP_S2	IP_S4	IP_S6	IPK_S2	IPK_S4	IPK_S6	JENIS SMA	STATUS SMA	PEND. ORTU	JOB ORTU	STATUS LULUS
201110370311027	1	16	1	86	127	4	4	3.79	3.82	3.87	3.87	2	1	6	1	TEPAT
201110370311028	1	24	1	82	124	3.18	3.45	3.48	3.04	3.13	3.23	1	2	6	4	TEPAT
201110370311048	1	24	2	82	124	2.92	3.09	3.19	2.91	2.93	3	2	1	7	1	TEPAT
201110370311050	1	16	2	86	127	3.95	3.92	3.55	3.63	3.77	3.72	2	2	2	4	TEPAT
201110370311004	2	20	1	82	124	3.47	3.41	3.76	3.37	3.35	3.52	2	1	4	4	TIDAK TEPAT
201110370311006	1	16	1	86	127	3.97	3.88	3.89	3.8	3.81	3.85	1	1	5	9	TIDAK TEPAT
201110370311014	1	16	1	82	126	3.39	3.77	3	3.24	3.39	3.28	1	1	2	4	TIDAK TEPAT
201110370311015	1	16	1	86	127	3.66	3.67	3.53	3.59	3.6	3.54	3	2	1	4	TIDAK TEPAT
201110370311019	2	16	2	82	124	3.18	3.41	3.71	3.08	3.21	3.36	1	1	6	5	TIDAK TEPAT

2.3 Implementasi algoritma

Pada tahap implementasi algoritma dilakukan proses klasifikasi data dengan menggunakan algoritma C5.0 untuk menghasilkan fitur-fitur penting yang paling berpengaruh. Atribut yang digunakan adalah 15 atribut, yaitu jenis kelamin, asal daerah, jalur masuk, SKS semester 4, SKS semester 6, IP semester 2, IP semester 4, IP semester 6, IPK semester 2, IPK semester 4, IPK semester 6, jenis SMA, status SMA, pendidikan orang tua, dan pekerjaan orang tua. Atribut-atribut ini merupakan kombinasi dari berbagai macam atribut yang telah digunakan dalam penelitian sebelumnya, yaitu dalam penelitian Risnawati[1], Suniantara dan Rusli[2],

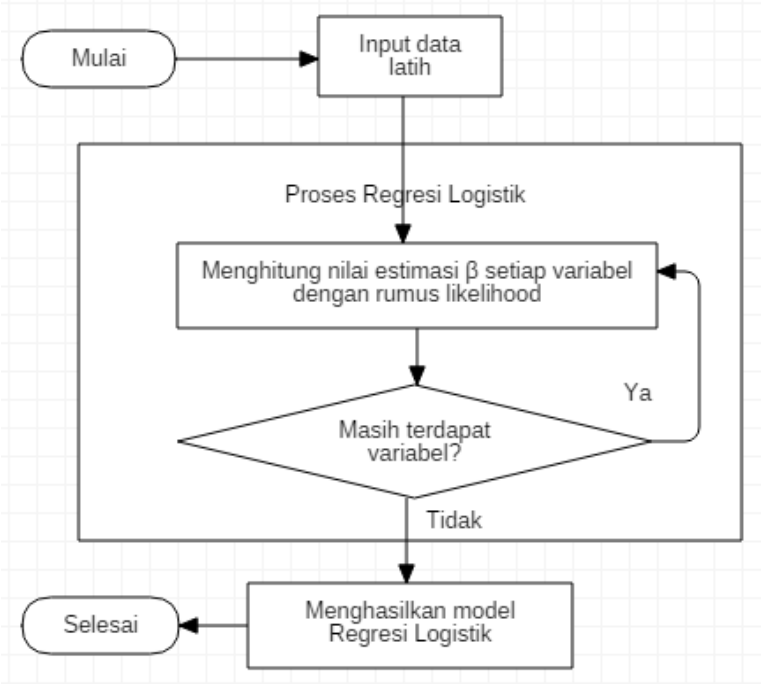
Suniantara[3], serta Rizki[4], meskipun ada beberapa atribut yang tidak digunakan dan terdapat atribut tambahan yang disesuaikan dengan ketersediaan data pada jurusan Teknik Informatika UMM.

Dalam melakukan implementasi algoritma C5.0 terdapat beberapa tahapan, yaitu mulai dari penginputan data, pemilihan node parent, hingga menghasilkan sebuah rule pohon keputusan dengan 2 kategori, yaitu “Tepat” dan “Tidak Tepat”. Rule pohon keputusan inilah yang merupakan hasil dari model klasifikasi C5.0. Rincian tahapan implementasi algoritma C5.0 dapat dilihat pada Gambar 2 berikut.



Gambar 2. Skenario Implementasi Algoritma C5.0

Setelah mendapatkan hasil dari implementasi algoritma C5.0, lalu dilakukan proses implementasi analisis regresi untuk melakukan estimasi peluang kelulusan tepat waktu mahasiswa dengan mengacu pada hasil selesi fitur algoritma C5.0. Dalam melakukan implementasi analisis regresi dengan menggunakan metode regresi logistik biner, terdapat beberapa tahapan, yaitu mengubah nilai kategorik dari masing-masing variabel menjadi nilai dummy (bernilai 0 atau 1), lalu dilanjutkan penginputan data, perhitungan estimasi β setiap variabel, hingga menghasilkan sebuah model regresi $\pi(\boldsymbol{\chi})$. Rincian tahapan implementasi algoritma C5.0 dapat dilihat pada Gambar 3 berikut.



Gambar 3. Skenario Implementasi Analisis Regresi

2.4 Pengujian

Pengujian dibagi menjadi 2, yaitu pengujian model klasifikasi dan pengujian model regresi. Pada tahap pengujian model klasifikasi dilakukan dengan mengimplementasikan algoritma klasifikasi yang lain, yaitu algoritma naïve bayes, dengan menggunakan seluruh atribut (fitur) yang ada dan dengan menggunakan hasil seleksi atribut dari algoritma C5.0. Hasil keduanya akan dibandingkan dan dihitung perbedaan nilai akurasinya dengan menggunakan confusion matrix. Tabel confusion matrix dapat dilihat pada Tabel 1 berikut [5].

Tabel 1. Tabel Confusion Matrix

		Actual	
		Yes	No
Prediksi	Yes	TP	FP
	No	FN	TN

Dimana:

a. TP (True Positive) = Jumlah sampel bernilai true yang diprediksi benar
b. TN (True Negative) = Jumlah sampel bernilai false yang diprediksi secara benar
c. FP (False Positive) = Jumlah sampel bernilai false yang salah diprediksi sebagai sampel bernilai true
d. FN (False Negative) = Jumlah sampel bernilai true yang salah diprediksi sebagai sampel bernilai true

Berikut Persamaan 1, Persamaan 2, dan Persamaan 3 untuk menghitung nilai akurasi, precision, dan recall berdasarkan tabel Confusion Matrix yang telah dibuat sebelumnya.

$$Akurasi = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

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

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

Sedangkan untuk pengujian model regresi dilakukan dengan uji serentak chi square, uji parsial, uji prediksi model, dan perhitungan root mean square error (RMSE). Berikut Persamaan 4 perhitungan RMSE.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (X_i - F_i)^2}{n}} \quad (4)$$

Dimana:
Xi = nilai prediksi
Fi = nilai sebenarnya
n = jumlah data yang ada

2.5 Analisa hasil dan penarikan kesimpulan

Pada tahap analisa hasil dilakukan analisa terhadap faktor-faktor yang telah ditemukan dan terhadap hasil prediksi lama masa studi mahasiswa untuk ditarik sebuah kesimpulan dari penelitian yang dilakukan.

3. Hasil Penelitian dan Pembahasan

Penelitian ini dibuat untuk menganalisis faktor-faktor pengaruh kelulusan tepat waktu mahasiswa Teknik Informatika UMM dengan menerapkan algoritma C5.0 dan analisis regresi. Hasil akhir dalam penelitian ini adalah berupa model klasifikasi yang menerangkan faktor-faktor utama yang mempengaruhi dan model regresi yang menerangkan estimasi peluang kelulusan tepat waktu mahasiswa.

3.1 Preprocessing Data

Dalam penelitian ini hanya dilakukan 2 teknik preprocessing data, yaitu data integration dan data cleaning. Data integration dilakukan penyatuan data yang berasal dari database yang berbeda-beda menjadi satu database yang baru untuk mempermudah pengelolaan data. Data cleaning dilakukan untuk menghilangkan dan memperbaiki data-data yang mengandung noise atau kosong. Jumlah keseluruhan data sebelum dilakukan data cleaning adalah 558 rule. Dan setelah melalui tahapan ini berkurang menjadi 557 rule.

3.2 Pembagian Data

Dalam penelitian ini menggunakan presentase pembagian data sebesar 75% untuk data latih dan 25% untuk data uji. Sehingga di dapatkan 418 data yang akan digunakan sebagai data latih dalam pembangunan model dan 139 data yang akan digunakan sebagai data uji dalam evaluasi model.

Dalam membagi data, penelitian ini menggunakan metode partisi data K-fold cross validation dengan K=4. Sehingga kesuluruhan jumlah data akan dibagi menjadi 4 partisi untuk mendapatkan hasil model terbaik. Pembagian partisi data dalam penelitian ini dapat dilihat pada Tabel 2 berikut.

Tabel 2. Pembagian Partisi Data

Data	K1	K2	K3	K4
Data Test	Data 1-139	Data 14-278	Data 279-417	Data 418-556
Data Train	Data 140-557	Data 1-139 Data 279-557	Data 1-278 Data 418-557	Data 1-417 Data 557

3.3 Implementasi Algoritma C5.0

Implementasi algoritma C5.0 dilakukan dengan bantuan perangkat lunak RStudio dan menggunakan bahasa pemograman R. Pembuatan model C5.0 dapat dilihat pada Gambar 3 dengan rincian langkah-langkah pembuatan model sebagai berikut:

1. Menghubungkan database ke dalam R
2. Membaca dataset pada database yang telah terhubung
3. Merubah bentuk dataset menjadi Data Frame yang merupakan kerangka data berisi variabel yang memiliki karakteristik
4. Merubah tipe data Status Lulus (sebagai variabel 'Y') menjadi label faktor untuk menginisialisasi level, yaitu "Tepat" dan "Tidak Tepat"
5. Inisialisasi variabel X dan Y
6. Inisialisasi data train dan data test
7. Pembuatan model C5.0

```
dblatih = dbConnect(MySQL(),
                    user = "root",
                    password = "",
                    host = "127.0.0.1",
                    dbname = "skripsiku")
dt = dbSendQuery(dblatih, "select * from  trainc5_k1")
data = fetch(dt, n=-1)
df = data.frame(data)
data$Status_Lulus <- as.factor(data$Status_Lulus)

X <- data[,2:16]
Y <- data[,17]
trainX <- X[140:557,]
trainY <- Y[140:557]
testX <- X[1:139,]
testY <- Y[1:139]

model_C5 <- C50::C5.0( trainX, trainY )
summary(model_C5)
plot(model_C5)
```

Gambar 4. Source Code Pembuatan Model C5.0

Source code pada Gambar 4 diatas diulang sebanyak 4 kali dengan menggunakan data pada masing-masing partisi yang telah dijelaskan dalam pembagian data sebelumnya. Hasilnya, data pada partisi K2 memiliki nilai akurasi terbaik, yaitu 92.8% dengan menghasilkan 8 fitur dari keseluruhan 15 fitur. Hasil model keputusan if-then pada partisi K2 dapat dilihat pada Gambar 5 berikut.

```
Console   Terminal
C:/xampp/htdocs/skripsweet/Skripsi/

Decision tree:

IPK_S6 <= 3.68: TIDAK TEPAT (343/18)
IPK_S6 > 3.68:
:...SKS_S6 <= 124: TIDAK TEPAT (25/3)
    SKS_S6 > 124:
    :...IP_S4 <= 3.6: TEPAT (4)
        IP_S4 > 3.6:
        :...ID_Sex > 1:
            :...ID_JenisSMA <= 1: TEPAT (10/4)
            :   ID_JenisSMA > 1: TIDAK TEPAT (7/1)
            ID_Sex <= 1:
            :...IP_S2 <= 3.84: TIDAK TEPAT (15/1)
                IP_S2 > 3.84:
                :...ID_Asal > 17: TEPAT (2)
                    ID_Asal <= 17:
                    :...IPK_S4 <= 3.78: TEPAT (3)
                        IPK_S4 > 3.78: TIDAK TEPAT (9/3)
```

Gambar 5. Hasil Model C5.0 Partisi K2

Model keputusan pada Gambar 5 menghasilkan 8 fitur dari total keseluruhan 15 fitur, yaitu IPK semester 6, SKS semester 6, IP semester 4, jenis kelamin, IP semester 2, jenis SMA, asal daerah, dan IPK semester 4. Ini artinya, faktor-faktor tersebut merupakan faktor utama yang memiliki pengaruh signifikan terhadap kelulusan tepat waktu mahasiswa Teknik Informatika UMM.

3.4 Pengujian Model Klasifikasi

Evaluasi model klasifikasi dalam penelitian ini adalah dengan mengimplementasikan kembali menggunakan algoritma klasifikasi yang lain, yaitu algoritma Naïve Bayes. Pengujiannya dibagi menjadi 2 jenis, yaitu dengan data yang menggunakan seluruh fitur dan dengan data yang menggunakan hasil seleksi fitur algoritma C5.0. Kemudian kedua hasil tersebut dilakukan perbandingan nilai akurasinya dengan menggunakan Confusion Matrix.

Berdasarkan implementasi algoritma klasifikasi Naïve Bayes dengan menggunakan seluruh fitur dan hanya menggunakan fitur hasil seleksi algoritma C5.0, maka didapatkan perbandingan nilai akurasi model sebagai berikut.

Tabel 3. Perbandingan Nilai Akurasi Pengujian Model Klasifikasi

	Semua Fitur	Hasil Seleksi Fitur
Akurasi	67.6%	69.8%

Dari perbandingan nilai akurasi pada Tabel 3, dapat diketahui bahwa model dengan menggunakan fitur hasil seleksi algoritma C5.0 memiliki nilai akurasi lebih tinggi jika dibandingkan dengan model yang menggunakan keseluruhan fitur yang ada. Maka, hasil pemangkasan fitur dengan algoritma C5.0 dianggap berhasil untuk menghasilkan model klasifikasi yang baik.

3.5 Implementasi Analisis Regresi

Implementasi analisis regresi dengan menggunakan metode regresi logistik biner dalam penelitian ini digunakan untuk melakukan estimasi peluang kelulusan tepat waktu mahasiswa Teknik Informatika UMM dengan menggunakan variabel bebas yang telah melalui tahap seleksi fitur pada algoritma C5.0. Variabel bebas tersebut adalah IPK semester 6, SKS semester 6, IP semester 4, jenis kelamin, IP semester 2, jenis SMA, asal daerah, dan IPK semester 4.

Pembuatan model regresi dapat dilihat pada Gambar 5 dengan rincian langkah-langkah pembuatan model sebagai berikut:

1. Menghubungkan database ke dalam R
2. Membaca dataset pada database yang telah terhubung
3. Merubah bentuk dataset menjadi Data Frame yang merupakan kerangka data berisi variabel yang memiliki karakteristik
4. Inisialisasi data train dan data test
5. Pembuatan model regresi

```
5  dblatih = dbConnect(MySQL(),
6                      user = "root",
7                      password = "",
8                      host = "127.0.0.1",
9                      dbname = "skripsiku")
10
11 dt5 = dbSendQuery(dblatih, "select * from dummy_traink2")
12 data5 = fetch(dt5, n=-1)
13 df5 = data.frame(data5)
14
15 trainRL <- data5[,2:34]
16 trainRL = trainRL[1:418, ]
17
18 RLmodel <- glm(Status_Lulus ~., data = trainRL, family = binomial)
19 RLmodel
20 summary(RLmodel)
```

Gambar 6. Source Code Pembuatan Model Regresi

Berdasarkan pembuatan model regresi dengan menggunakan bantuan perangkat lunak RStudio dan menggunakan bahasa pemograman R yang dapat dilihat pada Gambar 6, maka menghasilkan model regresi logistik seperti pada Gambar 7.

$$\pi(\boldsymbol{\chi}) = \frac{exp\left(\begin{array}{c} -5.4352 + 0.1527(\mathrm{ID_{Sex1}}) - 35.5508(\mathrm{ID_{Asal2}}) - 34.5271(\mathrm{ID_{Asal4}}) \\ - 32.2371(\mathrm{ID_{Asal5}}) - 19.2115(\mathrm{ID_{Asal8}}) - 0.9431(\mathrm{ID_{Asal9}}) \\ - 32.6092(\mathrm{ID_{Asal10}}) - 36.6972(\mathrm{ID_{Asal11}}) - 19.9449(\mathrm{ID_{Asal12}}) \\ - 36.3320(\mathrm{ID_{Asal13}}) - 36.1103(\mathrm{ID_{Asal14}}) - 20.6223(\mathrm{ID_{Asal16}}) \\ - 36.8284(\mathrm{ID_{Asal17}}) - 20.7173(\mathrm{ID_{Asal18}}) - 36.5505(\mathrm{ID_{Asal19}}) \\ - 36.4328(\mathrm{ID_{Asal20}}) - 20.3539(\mathrm{ID_{Asal21}}) - 21.1292(\mathrm{ID_{Asal22}}) \\ - 20.7992(\mathrm{ID_{Asal23}}) - 19.9880(\mathrm{ID_{Asal24}}) - 36.4695(\mathrm{ID_{Asal28}}) \\ - 37.3511(\mathrm{ID_{Asal29}}) - 35.4187(\mathrm{ID_{Asal32}}) + 0.1170(\mathrm{SKS_{S6}}) \\ - 0.4480(\mathrm{IP_{S2}}) + 0.3004(\mathrm{IP_{S4}}) - 5.7232(\mathrm{IPK_{S4}}) + 9.7921(\mathrm{IPK_{S6}}) \\ - 5.0212(\mathrm{ID_{JenisSMA1}}) - 4.6212(\mathrm{ID_{JenisSMA2}}) - 5.2652(\mathrm{ID_{JenisSMA3}}) \\ - 21.0154(\mathrm{ID\_JenisSMA4}) \end{array}\right)}{1+exp\left(\begin{array}{c} -5.4352 + 0.1527(\mathrm{ID_{Sex1}}) - 35.5508(\mathrm{ID_{Asal2}}) - 34.5271(\mathrm{ID_{Asal4}}) \\ - 32.2371(\mathrm{ID_{Asal5}}) - 19.2115(\mathrm{ID_{Asal8}}) - 0.9431(\mathrm{ID_{Asal9}}) \\ - 32.6092(\mathrm{ID_{Asal10}}) - 36.6972(\mathrm{ID_{Asal11}}) - 19.9449(\mathrm{ID_{Asal12}}) \\ - 36.3320(\mathrm{ID_{Asal13}}) - 36.1103(\mathrm{ID_{Asal14}}) - 20.6223(\mathrm{ID_{Asal16}}) \\ - 36.8284(\mathrm{ID_{Asal17}}) - 20.7173(\mathrm{ID_{Asal18}}) - 36.5505(\mathrm{ID_{Asal19}}) \\ - 36.4328(\mathrm{ID_{Asal20}}) - 20.3539(\mathrm{ID_{Asal21}}) - 21.1292(\mathrm{ID_{Asal22}}) \\ - 20.7992(\mathrm{ID_{Asal23}}) - 19.9880(\mathrm{ID_{Asal24}}) - 36.4695(\mathrm{ID_{Asal28}}) \\ - 37.3511(\mathrm{ID_{Asal29}}) - 35.4187(\mathrm{ID_{Asal32}}) + 0.1170(\mathrm{SKS_{S6}}) \\ - 0.4480(\mathrm{IP_{S2}}) + 0.3004(\mathrm{IP_{S4}}) - 5.7232(\mathrm{IPK_{S4}}) + 9.7921(\mathrm{IPK_{S6}}) \\ - 5.0212(\mathrm{ID_{JenisSMA1}}) - 4.6212(\mathrm{ID_{JenisSMA2}}) - 5.2652(\mathrm{ID_{JenisSMA3}}) \\ - 21.0154(\mathrm{ID\_JenisSMA4}) \end{array}\right)}$$

Gambar 7. Model Regresi

3.6 Pengujian Analisis Regresi

Evaluasi model regresi dalam penelitian ini adalah melakukan uji serentak chi square, uji parsial, uji prediksi model, dan perhitungan root mean square error (RMSE).

3.6.1 Uji Serentak Chi Square

Berikut hipotesis uji serentak chi square [6]:
H0 : β1 = β2 = … = βi = 0
H1 : paling sedikit terdapat satu βi ≠ 0

```
Console   Terminal
C:/xampp/htdocs/skripsweet/Skripsi/
> lrtest(RLmodel)
Likelihood ratio test

Model 1: Status_Lulus ~ ID_Sex1 + ID_Asal2 + ID_Asal4 + ID_Asal5 + ID_Asal8 +
    ID_Asal9 + ID_Asal10 + ID_Asal11 + ID_Asal12 + ID_Asal13 +
    ID_Asal14 + ID_Asal16 + ID_Asal17 + ID_Asal18 + ID_Asal19 +
    ID_Asal20 + ID_Asal21 + ID_Asal22 + ID_Asal23 + ID_Asal24 +
    ID_Asal28 + ID_Asal29 + ID_Asal32 + SKS_S6 + IP_S2 + IP_S4 +
    IPK_S4 + IPK_S6 + ID_JenisSMA1 + ID_JenisSMA2 + ID_JenisSMA3 +
    ID_JenisSMA4
Model 2: Status_Lulus ~ 1
  #Df   LogLik  Df  Chisq Pr(>Chisq)
1  33  -99.972
2   1 -134.118 -32 68.293  0.0001951 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Gambar 8. Hasil Uji Serentak Chi Square

Hasil uji serentak chi square pada Gambar 8, terlihat nilai chi square (Chisq) adalah sebesar 68.293 dengan p-value chi square (Pr(>Chisq)) adalah sebesar 0.0001951. Dengan menggunakan ∝ = 0.05 (5%), maka perbandingan nilai p-value chi square dengan ∝ adalah p-value < ∝, yaitu 0.0001951 < 0.05. Sehingga pengujian ini tolak H0, dimana paling sedikit terdapat satu variabel independen yang memiliki pengaruh signifikan terhadap kelulusan tepat waktu mahasiswa Teknik Informatika UMM.

3.6.2 Uji Parsial Wald

Berikut hipotesis untuk uji parsial setiap variabel [6]:

H0 : βi = 0

H1 : βi ≠ 0

Berdasarkan uji parsial wald yang didapatkan bersamaan dengan hasil model regresi logistik pada pembuatan model regresi yang dapat dilihat pada Gambar 5, maka didapatkan satu variabel yang memiliki nilai p-value dibawah nilai ∝ = 0.05 (5%), yaitu variabel IPK_S6 dengan nilai p-value wald sebesar 0.00303.

Maka, pengambilan keputusan uji parsial dalam regresi logistik ini adalah tolak H0, dimana variabel IPK_S6 secara parsial memiliki pengaruh signifikan terhadap kelulusan tepat waktu mahasiswa Teknik Informatika UMM.

3.6.3 Prediksi Model Regresi

Prediksi digunakan untuk mengevaluasi model regresi logistik yang telah didapatkan dengan menerapkan data testing ke dalam model tersebut. Pembuatan prediksi model regresi dapat dilihat pada Gambar 9 berikut.

```
24  dt6 = dbSendQuery(dblatih, "select * from dummy_testk2")
25  data6 = fetch(dt6, n=-1)
26  df6 = data.frame(data6)
27
28  testRL <- data6[,2:34]
29  testRL = testRL[1:137, ]
30
31  RLpred <- predict(RLmodel, newdata = testRL, type="response")
32  RLpred <- ifelse(RLpred>0.5, 1, 0)
33  tabpred <- table(Predicted=RLpred, Actual=testRL$Status_Lulus)
34  tabpred
35
```

Gambar 9. Pembuatan Prediksi Model Regresi

Pembuatan prediksi model regresi di atas mengatur tipe parameter sebagai 'respon', sehingga akan menampilkan nilai probabilitas dalam bentuk P (y = 1 | X). Batas keputusan yang diatur adalah 0.5. Jika P (y = 1 | X) > 0.5 maka akan menghasilkan y = 1 (lulus tepat waktu) dan sebaliknya apabila P (y = 1 | X) < 0.5 maka akan menghasilkan y = 0 (lulus tidak tepat waktu). Berikut hasil prediksi model regresi yang dihasilkan dalam bentuk confusion matrix.

Tabel 4. Confusion Matrix Hasil Prediksi Model Regresi

Predicted	Actual	
	0	1
0	112	24
1	1	0

Hasil confusion matrix pada Tabel 4 menunjukkan bahwa terdapat 122 kasus kelulusan tidak tepat waktu yang diprediksi secara benar dan 1 kasus kelulusan tidak tepat waktu yang diprediksi secara salah. Selain itu juga menunjukkan bahwa terdapat 24 kasus kelulusan tepat waktu yang dipredisi secara salah dan tidak ada kasus kelulusan tepat waktu yang diprediksi secara benar.

Tidak adanya kasus kelulusan tepat waktu yang diprediksi secara benar kemungkinan dapat dipengaruhi oleh sedikitnya kasus kelulusan tepat waktu yang ada pada data yang digunakan, yaitu dengan total hanya 65 kasus. Sedangkan kasus kelulusan tidak tepat waktu pada data yang digunakan sangatlah banyak, yaitu 492 kasus. Sehingga model yang dibentuk dapat terpengaruh oleh banyaknya kasus kelulusan tidak tepat waktu yang sangat banyak dan tidak seimbang dengan jumlah kasus kelulusan tepat waktu yang ada.

Tetapi, meskipun tidak adanya kasus kelulusan tepat waktu yang diprediksi secara benar, model regresi logistik yang dihasilkan memiliki nilai akurasi sebesar 82% dengan perhitungan Persamaan 1.

$$Akurasi = \frac{TP + TN}{TP + TN + FP + FN} = \frac{112 + 0}{112 + 24 + 1 + 0} = 0.82 = 82\% \quad (1)$$

4. Kesimpulan

Berdasarkan implementasi dan pengujian algoritma yang telah dilakukan dan dijelaskan pada bab hasil dan pembahasan, maka didapatkan kesimpulan sebagai berikut:

1. Algoritma klasifikasi C5.0 mampu melakukan seleksi fitur untuk menentukan pengaruh kelulusan tepat waktu mahasiswa Teknik Informatika UMM dengan menghasilkan 8 fitur dari total keseluruhan 15 fitur dengan nilai akurasi sebesar 91.9%.
2. Faktor-faktor utama yang dianggap mempengaruhi kelulusan tepat waktu mahasiswa Teknik Informatika UMM adalah IPK semester 6, SKS semester 6, IP semester 4, jenis kelamin, IP semester 2, jenis SMA, asal daerah, dan IPK semester 4.
3. Pengujian klasifikasi pada penelitian ini menghasilkan nilai akurasi sebesar 67.6% dengan menggunakan keseluruhan fitur dan 69.8% dengan menggunakan hasil seleksi fitur algoritma C5.0.
4. Analisis regresi logistik biner dalam penelitian ini mampu menghasilkan nilai akurasi sebesar 82% dengan model sebagai berikut:

Referensi

[1] Risnawati, “Analisis Kelulusan Mahasiswa Menggunakan Algoritma C.45,” *J. Mantik Penusa*, vol. 2, no. 1, pp. 71–76, 2018.
[2] I. K. P. Suniantara and M. Rusli, “Klasifikasi Waktu Kelulusan Mahasiswa Stikom Bali Menggunakan Chaid Regression – Trees Dan Regresi Logistik Biner,” vol. 5, no. 1, 2017.
[3] I. K. P. Suniantara, “Analisis Clasification and Regression Trees (CART) pada Lama Studi Mahasiswa STIKOM BALI,” in *SENAPATI 2016*, 2016, pp. 30–34.
[4] R. Fitriana, “Analisis Survival Faktor-Faktor Yang Mempengaruhi Lama Studi Mahasiswa Pendidikan Matematika Angkatan 2010 Dengan Metode Regresi Cox Proportional Hazard,” Universitas Negeri Semarang, 2016.
[5] Betrisandi, “Klasifikasi Nasabah Asuransi Jiwa Menggunakan Algoritma Naive Bayes Berbasis Backward Elimination,” *Ilk. J. Ilm.*, vol. 9, no. April, pp. 96–101, 2017.
[6] Y. A. Tampil, H. Komalig, and Y. Langi, “Analisis Regresi Logistik Untuk Menentukan Faktor-Faktor Yang Mempengaruhi Indeks Prestasi Kumulatif (IPK) Mahasiswa FMIPA Universitas Sam Ratulangi Manado,” *JdC*, vol. 6, no. 2, pp. 56–62, 2017.
[7] F. A. Hermawati, *Data Mining*. Yogyakarta: Penerbit ANDI Yogyakarta, 2013.
[8] E. Prasetyo, *Data Mining - Konsep dan Aplikasi Menggunakan MATLAB*. Yogyakarta: Penerbit ANDI Yogyakarta, 2012.
[9] H. Munawaroh, B. Khusnul, and Y. Kustiyahningsih, “Perbandingan Algoritma ID3 dan C5.0 dalam Indentifikasi Penjurusan Siswa SMA,” *J. Sarj. Tek. Inform.*, vol. 1, no. 1, pp. 1–12, 2013.
[10] C. Hutabarat, “Penerapan Data Mining Untuk Memprediksi Permintaan Produk Kartu Perdana Internet Menggunakan Algoritma C5.0 (Studi Kasus: Vidha Ponsel),” *J. Pelita Inform.*, vol. 17, no. 2, pp. 168–173, 2018.
[11] Sugiarto, *Tahap Awal + Aplikasi Analisis Regresi*. Yogyakarta: Andi Offset, 1992.
[12] Zakariyah and I. Zain, “Analisis Regresi Logistik Ordinal pada Prestasi Belajar Lulusan Mahasiswa di ITS Berbasis SKEM,” *J. Sains dan Seni ITS*, vol. 4, no. 1, pp. 121–126, 2015.
[13] H. Yuliansyah, “Perancangan Replikasi Basis Data Mysql Dengan Mekanisme Pengamanan Menggunakan Ssl Encryption,” *J. Inform.*, vol. 8, no. 1, pp. 826–836, 2014