JEPIN

(Jurnal Edukasi dan Penelitian Informatika)

ISSN(e): 2548-9364 / ISSN(p): 2460-0741

Vol. 4
No. 2
Desember
2018

Perbandingan Akurasi Regresi Logistik dengan Regresi Multinomial untuk Prediksi Kategori IP Mahasiswa Jalur Prestasi

R. Gunawan Santosa[#1], Antonius Rachmat Chrismanto[#2]

[#]*Fakultas Teknologi Informasi, Program Studi Informatika, Universitas Kristen Duta Wacana Yogyakarta*
Jl. Dr. Wahidin Sudirohusodo No. 5-25 Yogyakarta

[1]gunawan@staff.ukdw.ac.id
[2]anton@ti.ukdw.ac.id

***Abstrak*— Penelitian ini membandingkan akurasi prediksi antara model regresi logistik dengan model regresi multinomial terhadap kategori Indeks Prestasi semester pertama (IPS1) mahasiswa jalur prestasi di Fakultas Teknologi Informasi (FTI) UKDW untuk mengetahui akurasi yang terbaik. Pada kedua model tersebut yang digunakan sebagai variabel dependen adalah indeks prestasi semester pertama (IPS1) dan sebagai variabel independen yang digunakan adalah lokasi, kategori, dan status SMA serta level kemampuan bahasa Inggris. Dataset untuk membuat model regresi logistik maupun regresi multinomial digunakan data mahasiswa FTI angkatan 2008 sampai dengan angkatan 2015 secara kumulatif, sedangkan data ujinya menggunakan data IPS1 mahasiswa angkatan 2016. Akurasi kedua model diukur dengan menggunakan tabel Crosstab. Tabel *Crosstab* menunjukkan nilai akurasi antara kategori data hasil prediksi dengan kategori data IPS1 mahasiswa FTI angkatan 2016 yang sebenarnya. Rata-rata akurasi prediksi kategori IPS1 Mahasiswa Angkatan 2016 melalui Jalur Prestasi dengan dengan model regresi logistik adalah 67,80125%, sedangkan rata-rata akurasi hasil prediksi dengan model regresi Multinomial adalah 67,175 %. Selisih akurasi dari kedua metode tersebut ternyata tidak jauh berbeda.**

***Kata kunci*— Regresi logistik, Regresi multinomial, Prediksi, Tabel crosstab**

I. PENDAHULUAN

Indeks prestasi mahasiswa merupakan evaluasi keberhasilan belajar di perguruan tinggi. Ada beberapa mahasiswa yang mendapat Indeks Prestasi tinggi, namun ada pula mahasiswa-mahasiswa yang memperoleh indeks prestasi yang sangat rendah. Pada pengamatan yang terjadi di Universitas Kristen Duta Wacana (UKDW) terutama di Fakultas Teknologi Informasi (FTI), ternyata di tahun pertama ada beberapa mahasiswa yang pindah ke fakultas lain. Selain itu ada beberapa mahasiswa yang sudah tidak melakukan registrasi pada semester-semester berikutnya yaitu sudah tidak aktif dalam perkuliahan atau mengundurkan diri sebagai mahasiswa FTI UKDW.

Ada banyak faktor internal yang mempengaruhi nilai indeks prestasi mahasiswa, misalnya minat, bakat, kemampuan, kerajinan, motivasi diri dalam mahasiswa. Selain faktor internal, terdapat juga faktor eksternal yang akan menjadi perhatian: 1) lokasi sekolah menengah atas (SMA), yang dapat dikategorikan menjadi SMA di Jawa dan SMA di luar Jawa, 2) kategori sekolah menengah atas yaitu SMA dan SMK, 3) status SMA, yaitu SMA dengan Negeri atau Swasta, dan 4) level bahasa Inggris dari calon mahasiswa (level dasar, menengah, dan tinggi).

Pada penelitian [1] telah diselidiki tentang faktor-faktor yang mempengaruhi IPK mahasiswa secara signifikan, namun belum melihat hubungan yang jelas antara IPK dan faktor-faktor tersebut. Ternyata terdapat korelasi antara Indeks Prestasi Semester 1 (IPS1) dengan Indeks Prestasi kumulatif mahasiswa. Terbukti bahwa IPS1 paling berpengaruh pada Indeks Prestasi semester semester sesudahnya.

Pada penelitian [2] telah didapatkan hasil prediksi kategori IPS1 dengan model regresi Logistik untuk mahasiswa angkatan 2015 yang melalui jalur prestasi dengan rata-rata hasil akurasinya adalah 67,857%.

Pada penelitian [3], Santosa & Chrismanto telah menyelidiki tentang prediksi kategori IP semester 1 dengan menggunakan model regresi logistik. Dan hasilnya ternyata rata-rata akurasi prediksi kategori IP semester 1 mahasiswa angkatan tahun 2016 jalur prestasi adalah 67,8013%. Hasil ini tidak terlalu berbeda dengan penelitian sebelumnya [2].

Penelitian ini merupakan kelanjutan dari penelitian sebelumnya yang pada hakikatnya mencoba membandingkan akurasi model regresi multinomial dengan model regresi logistik untuk memprediksi kategori IPS 1 mahasiswa FTI UKDW. Penelitian ini diharapkan dapat meningkatkan akurasi prediksi terhadap kategori IPS1 mahasiswa angkatan 2016.

Submitted 28-08-2018; Revised 20-12-2018; Accepted 28-12-2018

II. TINJAUAN PUSTAKA

A. Educational Data Mining

Penelitian ini pada hakikatnya termasuk dalam bidang data mining. Data mining mempunyai akar pada ilmu pembelajaran mesin (*machine learning*), kecerdasan buatan (a*rtificial intelligence*), ilmu komputer (c*omputer science*), dan statistik [4]. Berbagai pendekatan metode dalam data mining yang berbeda adalah pengelompokan, klasifikasi, dan pembangunan aturan asosiasi. Masing-masing pendekatan ini dapat digunakan untuk menganalisis secara kuantitatif sekumpulan data besar untuk mendapatkan makna (pola) tersembunyi pada suatu sistem. Data mining biasanya adalah proses eksplorasi, tetapi ia dapat juga digunakan untuk penyelidikan konfirmasi [5] [6] [7] [8].

Pada penelitian [9] data mining adalah suatu ilmu yang sedang berkembang pesat sebagai bagian dari ilmu komputer dan teknik Informatika. Data mining yang diterapkan dalam pendidikan atau pengajaran adalah suatu penelitian yang menyangkut berbagai bidang ilmu interdisipliner yang saat ini berkembang sangat pesat. Bagian data mining ini, juga dikenal dengan istilah Educational Data Mining (EDM). Hal ini berkaitan dengan pengembangan berbagai metode untuk mengeksploitasi jenis data unik yang berasal dari lingkungan pendidikan. Tujuannya adalah untuk memahami lebih baik bagaimana proses belajar dari siswa dan mengidentifikasi keadaan atau kondisi dimana mereka dapat meningkatkan hasil pendidikan serta untuk mendapatkan wawasan dan mendapatkan kejelasan pada fenomena pendidikan yang sedang terjadi.

Menurut penelitian [9], ada beberapa metode populer yang ada dalam EDM, yaitu: prediksi, *clustering* (pengelompokan), deteksi *outlier*, *relationship mining*, *Social Network Analysis* (analisis jaringan sosial), *process mining*, dan *text mining*. Pada penelitian [10] dilakukan penelitian yang berhubungan dengan data mining dalam konteks pendidikan dengan menawarkan model untuk sistem pendidikan tinggi di universitas. Dalam penelitian ini, metode klasifikasi digunakan untuk mengevaluasi kinerja siswa dan karena ada banyak pendekatan yang digunakan untuk klasifikasi data, metode pohon keputusan (*decision tree*) digunakan pada penelitian ini. Dengan menggunakan klasifikasi data maka dapat mengekstrak pengetahuan yang menggambarkan kinerja siswa dalam ujian akhir semester. Hal ini membantu dalam mengidentifikasi siswa putus sekolah dan siswa yang membutuhkan perhatian khusus serta memungkinkan pengajar untuk memberikan bimbingan atau konseling yang sesuai.

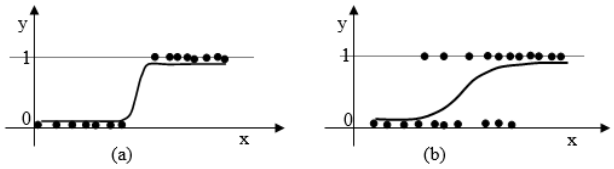
Raheela Asif, Agathe Merceron, Syed Abbas Ali, dan Najmi Ghani Haider [11] mengadakan penelitian menggunakan metode data mining untuk mempelajari dan mengamati kinerja mahasiswa. Ada dua aspek kinerja siswa telah diamati dan dipelajari. Pertama, adalah memprediksi pencapaian akademik siswa pada akhir empat tahun program studi. Kedua, mempelajari perkembangan khusus dan menggabungkannya dengan hasil prediksi tersebut. Dua kelompok siswa telah diidentifikasi yaitu: siswa berprestasi tinggi dan rendah. Hasilnya menunjukkan bahwa dengan berfokus pada sejumlah mata kuliah khusus yang merupakan indikator kinerja yang sangat baik atau buruk, maka adalah mungkin untuk menyediakan peringatan atau pesan yang tepat waktu bagi siswa. Dengan adanya pesan tersebut, maka dapat diberikan peringatan DO (*Drop Out*) dan dukungan untuk siswa berprestasi rendah sedangkan untuk siswa yang berprestasi tinggi dapat diberikan peluang dan nasehat lebih lanjut yang berguna untuk karir mereka.

Terakhir, institusi pendidikan mulai menggunakan teknik data mining untuk meningkatkan layanan yang mereka sediakan dan untuk meningkatkan nilai dan pengetahuan siswa, oleh karena itu S. Lakshmi Prabha dan A.R. Mohamed Shanavas [12] melakukan penelitian yang membahas tentang aplikasi *Educational Data Mining* (EDM) yang dapat membuat prediksi dan rekomendasi tindakan berdasarkan peningkatan pengetahuan ke dalam tindakan siswa. Pengajar mendapatkan pemahaman dan pandangan terhadap kinerja siswa yang akan membantu mereka dalam menyesuaikan pengajaran, tutorial dan tugas yang akan diberikan kepada siswa. Dengan menggunakan data yang ada, maka institusi dapat membuat kebijakan, melaksanakan program, dan menyesuaikan kebijakan dan program untuk kemajuan pengajaran dan pembelajaran.

B. Regresi Logistik dan Multinomial

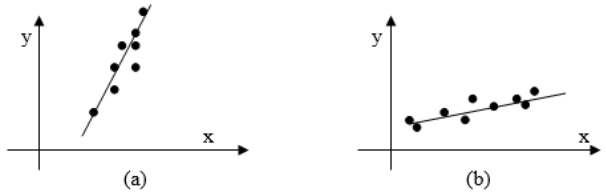
Regresi logistik mulai digunakan pada riset epidemiologi. Regresi ini membuat atau membentuk hubungan antara satu atau lebih variabel independen (*eksposure*) terhadap kondisi penyakit yang diderita. Dalam hal ini penyakit yang diderita merupakan variabel dependen yang bernilai dikotomi (biner) yaitu 0 mewakili "tidak sakit" dan 1 mewakili "sakit".

Bentuk regresi logistik sederhana dan regresi multinomial sederhana (regresi linier) dapat digambarkan secara mudah pada keterangan berikut ini. Untuk bentuk regresi logistik (2 variabel x dan y): $y = \frac{1}{1+e^{-(\alpha+\beta x)}}$ atau $y = \frac{e^{(\alpha+\beta x)}}{1+e^{(\alpha+\beta x)}}$. Data yang mempunyai titik *cut off* yang tajam antara nilai y=0 dan y=1 akan mempunyai nilai β yang cenderung besar, seperti pada Gambar 1 (a). Data yang mempunyai transisi yang panjang dari y=0 dan y=1 akan mempunyai nilai β yang cenderung kecil, seperti pada Gambar 1 (b) [13]. Untuk melakukan proses klasifikasi data, secara umum Gambar 1 (a) lebih baik karena lebih tegas.



Gambar 1. Grafik Regresi Logistik

Untuk bentuk regresi multinomial (2 variabel x dan y): $y = \alpha + \beta x$. Data yang mempunyai gradien atau kemiringan yang besar (mendekati tegak) akan mempunyai nilai β yang cenderung besar, seperti pada Gambar 2 (a). Data yang mempunyai gradien atau kemiringan yang kecil (mendekati datar) akan mempunyai nilai β yang cenderung kecil, seperti pada Gambar 2 (b). Untuk melakukan proses klasifikasi data, secara umum Gambar 2 (a) lebih baik karena jarak antara titik tertinggi dan terendah lebih terdeteksi.

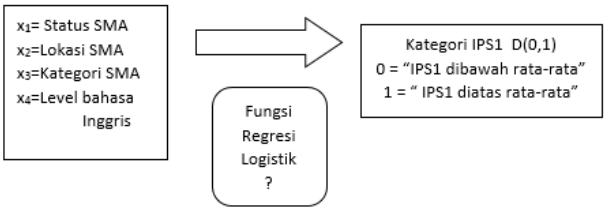


Gambar 2. Grafik Regresi Multiinomial

Penelitian yang dilakukan pada tulisan ini difokuskan pada analisis mengenai prediksi data. Prediksi indeks prestasi (IP) mahasiswa semester satu di suatu perguruan tinggi perlu dilakukan secara dini untuk menanggulangi masalah-masalah yang mungkin akan ditimbulkan oleh mahasiswa di kemudian hari. Penelitian yang berhubungan dengan prediksi IP semester satu mahasiswa telah dilakukan, diantaranya adalah:

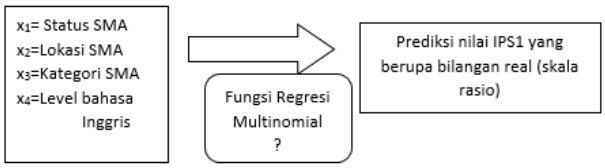
1. Telah dilakukan penelitian perbandingan akurasi algoritma C4.5 dan CART dalam memprediksi kategori IP mahasiswa, hasilnya Algoritma C4.5 dan CART memiliki akurasi yang sama untuk memprediksi kategori IP mahasiswa baru angkatan 2016 pada jalur prestasi (data non-numerik), yaitu sebesar 86,86%. Untuk memprediksi kategori IP mahasiswa baru angkatan 2016 pada jalur non-prestasi (data numerik), algoritma CART memberikan akurasi lebih baik daripada C4.5, yaitu 63,16% berbanding 61,54% [14].
2. Telah dilakukan penelitian tentang perbandingan akurasi model regresi logistik untuk prediksi kategori IP mahasiswa jalur prestasi dengan non jalur prestasi, hasilnya hasil rata-rata akurasi model prediksi kategori IPS1 mahasiswa angkatan 2016 jalur prestasi setelah uji signifikansi adalah 70,94% sedangkan hasil akurasi rata-rata model prediksi kategori IPS1 mahasiswa angkatan 2016 non jalur prestasi setelah uji signifikansi adalah 65,8%. Jadi rata-rata persentase akurasi model regresi logistik dalam memprediksi kategori IPS1 mahasiswa baru angkatan 2016 melalui jalur prestasi lebih tinggi daripada persentase akurasi model regresi Logistik dalam memprediksi kategori IPS1 mahasiswa baru melalui non jalur prestasi [15].

Pada penelitian ini dibuat model regresi logistik dan model regresi multinomial untuk memprediksi kategori IPS1 mahasiswa FTI UKDW. Pada model regresi logistik variabel independennya adalah status SMA (x_1), lokasi SMA (x_2), kategori SMA (x_3), level bahasa Inggris (x_4). Sedangkan variabel dependennya adalah kategori IPS1 yaitu variabel dikotomi yang berupa status subjek yang diklasifikasikan sebagai 0 ("IPS1 dibawah rata-rata") dan 1 ("IPS1 diatas rata-rata"). Gambar 3 menggambarkan skema model regresi logistik pada penelitian ini.



Gambar 3. Skema model regresi logistik

Gambar 4 menggambarkan skema model regresi multinomial pada penelitian ini.

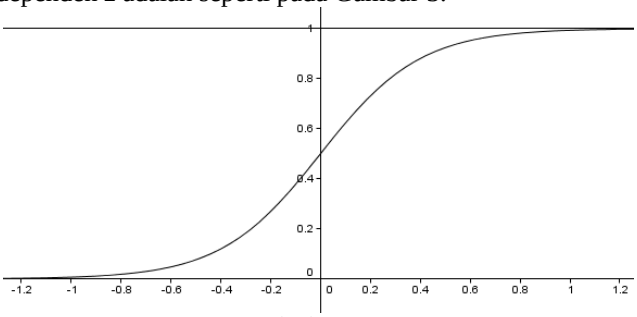


Gambar 4. Skema model regresi multinomial

Regresi logistik adalah pendekatan model yang dapat digunakan untuk menggambarkan relasi beberapa variabel independen X terhadap variabel dependen D yang bernilai dikotomi. Regresi logistik didasarkan pada fungsi logistik yang dapat dilihat pada persamaan (1).

$$f(z) = \frac{1}{1+e^{z}} \text{ dengan } -\infty < z < \infty \quad \ldots(1)$$

Bentuk grafik fungsi logistik dengan variabel independen z adalah seperti pada Gambar 3.



Gambar 5. Grafik fungsi logistik

Dari Gambar 5 terlihat sifat-sifat fungsi logistik sebagai berikut:

1. Range nilai f(z) terletak di antara 0 sampai dengan 1 tanpa memandang nilai z.
2. Untuk $z \rightarrow -\infty$ maka $f(-\infty) = 0$ dan untuk $z \rightarrow \infty$ maka $f(\infty) = 1$.
3. Alasan lain mengapa model logistik menjadi populer adalah karena bentuk dari fungsi logistik. Bentuk fungsi ini menggambarkan bentuk kurva S, sehingga sering disebut termasuk fungsi *sigmoid*. Seperti dilihat pada Gambar 3, jika

dimulai dari z = -∞ bergerak ke kanan maka nilai f(z) naik melengkung tajam mendekati z=0 untuk sementara, kemudian menaik secara dramatis mendekati 1 dan akhirnya akan mendekati di sekitar 1.

Karena nilai f(z) terletak di antara 0 sampai dengan 1 inilah maka alasan regresi logistik adalah populer. Model ini dirancang untuk menggambarkan probabilitas yang selalu terletak antara 0 sampai dengan 1 [16]. Dalam istilah epidemiologi probabilitas seperti ini memberikan nilai resiko dari seorang individu untuk mendapatkan penyakit tertentu.

Sedangkan dalam penelitian ini model regresi logistik adalah f dan sebagai fokus utama adalah z yang dapat dituliskan sebagai kombinasi linier dari x_1, x_2, x_3, …, x_k, yang dapat dilihat pada persamaan (2) dan (3).

$$z = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k \quad \ldots(2)$$

sehingga $\hat{f}(z) = \frac{1}{1+e^{-(\alpha+\sum \beta_i x_i)}}$ …(3)

Nilai α dan βi merupakan parameter yang tidak diketahui dan akan diestimasi dari dataset yang berupa sampel dengan menggunakan *Maksimum Likelihood Estimator* (MLE). Sehingga parameter α dan βi akan diestimasi oleh $\hat{\alpha}$ dan $\widehat{\beta_i}$.

Sedangkan, regresi multinomial merupakan pengembangan dari regresi linier sederhana. Regresi multinomial mempunyai bentuk model pada formula (4):

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k \quad \ldots(4)$$

dengan y adalah variabel dependen dan x_1, x_2, x_3, x_4, … x_k adalah variabel independen. Dengan menggunakan estimasi LSE (*Least Square Estimation*) maka akan didapat nilai $\hat{\beta}$ dalam bentuk persamaan normal pada persamaan 5:

$$\hat{\beta} = (\bar{x}'\bar{x})^{-1}(\bar{x}'\bar{y}) \quad \ldots(5)$$

Nilai y pada formula (4) akan berbentuk skala rasio sehingga perlu ditransformasikan menjadi bentuk nominal atau ordinal supaya dapat dibandingkan dengan regresi logistik yang outputnya berbentuk variabel kategori (nominal atau ordinal).

Transformasi untuk mengubah skala rasio menjadi skala nominal atau ordinal menggunakan pengkodean ulang, yaitu bila $\hat{y}$ < rata-rata IPS1 maka dikodekan dengan "0" sedangkan bila $\hat{y}$ ≥ rata-rata IPS1 akan dikodekan dengan "1". Sehingga hasil prediksi dari regresi logistik dan regresi multinomial ini mempunyai tipe skala yang sama yaitu berskala nominal atau ordinal.

Pada penelitian ini, untuk mengukur ketepatan hasil prediksi dengan model regresi logistik dan regresi multinomial dapat digunakan tabel Crosstab (tabulasi silang) [17] antara data hasil prediksi dan nilai data observasi sebenarnya seperti dapat dilihat pada Tabel I.

TABEL I
CROSSTAB ANTARA NILAI DATA OBSERVASI SEBENARNYA DAN NILAI DATA HASIL PREDIKSI

		Nilai Data Hasil Prediksi		**Total Baris**
		Kategori 1	Kategori 2	
Nilai Data Observasi Sebenarnya	Kategori 1	n_{11}	n_{12}	b_1
	Kategori 2	n_{21}	n_{22}	b_2
Total Kolom		k_1	k_2	n

Ketepatan hasil prediksi diukur dengan persamaan 6.

Presentasi Kecocokan = $(n_{11} + n_{22}) / n$ …(6)

Persamaan 6 tersebut yang merupakan ukuran akurasi untuk kedua model tersebut.

III. METODOLOGI PENELITIAN

Tahapan penelitian yang dilakukan pada penelitian ini adalah sebagai berikut :

1. Peneliti mengambil data mahasiswa UKDW dari PUSPINDIKA yang meliputi data mahasiswa Fakultas Teknologi Informasi yang terdiri dari Prodi SI dan Prodi TI. Sebagai batasan hanya akan diambil data mahasiswa SI dan TI sebanyak 8 tahun akademik yang berurutan yaitu dari angkatan 2008 sampai dengan angkatan 2015. Data yang dari Puspindika ini meliputi kategori SMA, status SMA, lokasi SMA dan Indeks Prestasi semester 1.
2. Setelah itu peneliti juga mengambil data level kemampuan bahasa Inggris calon mahasiswa FTI UKDW dari PUSPINDIKA (dulu dari PPBA UKDW).
3. Setelah itu peneliti memilih data mahasiswa FTI tersebut yang diterima melalui jalur prestasi. Kenyataan saat ini proses penerimaan mahasiswa baru di UKDW ada dua jalur yaitu jalur prestasi dan non jalur prestasi (reguler). Perbedaan antara dua jalur ini adalah pada proses penerimaannya, pada jalur prestasi tidak mengikuti tes kemampuan akademik (tes kemampuan Numerik, Verbal, Spasial dan Analogi) sedangkan pada non jalur prestasi harus mengikuti empat tes kemampuan akademik tersebut.
4. Kemudian setelah dibersihkan data tersebut diolah dengan menggunakan regresi logistik.

   Model regresi logistik sebagai fokus utama adalah z yang dapat dituliskan sebagai kombinasi linier dari x_1, x_2, x_3, …, x_k, yang dapat dilihat pada persamaan 2 dan 3, yaitu:

   $$z = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$$

   sehingga $\hat{f}(z) = \frac{1}{1+e^{-(\alpha+\sum \beta_i x_i)}}$

   Nilai α dan β_i merupakan parameter yang tidak diketahui dan akan diestimasi dari data sampel dengan menggunakan *Maksimum Likelihood Estimator* (MLE). Sehingga α dan β_i akan diestimasi oleh $\hat{\alpha}$ dan $\widehat{\beta_i}$.

5. Peneliti kemudian menggunakan model regresi multinomial mempunyai bentuk model pada persamaan 4.
   $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$ dengan y adalah variabel dependen dan $x_1, x_2, x_3, x_4, \ldots x_k$ adalah variabel independen. Dengan menggunakan estimasi LSE (*Least Square Estimation*) maka akan didapat nilai $\hat{\beta}$ dalam bentuk persamaan normal pada persamaan 5.
6. Mencari 7 model Regresi Logistik dan 7 model Regresi Multinomial untuk kasus penerimaan mahasiswa baru melalui jalur prestasi adalah:
   - Model A adalah model Regresi Logistik dengan menggunakan data mahasiswa angkatan 2008
   - Model B adalah model Regresi Logistik dengan menggunakan data mahasiswa angkatan 2008 sampai dengan angkatan 2009
   - Model C adalah model Regresi Logistik dengan menggunakan data mahasiswa angkatan 2008 sampai dengan angkatan 2010
   - Model D adalah model Regresi Logistik dengan menggunakan data mahasiswa angkatan 2008 sampai dengan angkatan 2011
   - Model E adalah model Regresi Logistik dengan menggunakan data mahasiswa angkatan 2008 sampai dengan angkatan 2012
   - Model F adalah model Regresi Logistik dengan menggunakan data mahasisiwa angkatan 2008 sampai dengan angkatan 2013
   - Model G adalah model Regresi Logistik dengan menggunakan data mahasiswa angkatan 2008 sampai dengan angkatan 2014
   - Model H adalah model Regresi Logistik dengan menggunakan data mahasiswa angkatan 2008 sampai dengan angkatan 2015
7. Menggunakan 8 model Regresi Logistik untuk memprediksi kategori IPS 1 mahasiswa baru angkatan 2016 yang penerimaannya melalui jalur prestasi.
8. Menggunakan 8 model Regresi Multinomial untuk memprediksi kategori IPS 1 mahasiswa baru angkatan 2016 yang penerimaannya melalui jalur prestasi.
9. Mengukur akurasi hasil prediksi regresi logistik dan regresi multinomial dari ke tujuh model tersebut terhadap data observasi sebenarnya dengan menggunakan *Crosstab* (tabulasi silang) untuk penerimaan melalui jalur prestasi.

IV. HASIL DAN PEMBAHASAN

Setelah data dilakukan proses data *cleaning* maka dapat diperoleh jumlah kumulatif mahasiswa FTI yang akan digunakan untuk membuat regresi logistik disajikan pada Tabel II dan III.

TABEL III
JUMLAH MAHASISWA FAKULTAS TEKNOLOGI INFORMASI DARI TAHUN 2008 SAMPAI DENGAN TAHUN 2016

No	Angkatan	Total	Jalur Prestasi	Non Jalur Prestasi
1	2008	368	63	305
2	2009	260	11	249
3	2010	264	55	209
4	2011	251	144	107
5	2012	244	125	119
6	2013	205	125	80
7	2014	171	90	81
8	2015	254	193	61
9	2016	210	99	111

TABEL IIIII
JUMLAH KUMULATIF MAHASISWA FAKULTAS TEKNOLOGI INFORMASI DARI TAHUN 2008 SAMPAI DENGAN TAHUN 2015

No	Angkatan	Total	Jalur Prestasi	Non Jalur Prestasi
1	2008	368	63	305
2	2008 sampai dengan 2009	628	74	554
3	2008 sampai dengan 2010	892	129	763
4	2008 sampai dengan 2011	1143	273	870
5	2008 sampai dengan 2012	1387	398	989
6	2008 sampai dengan 2013	1592	523	1069
7	2008 sampai dengan 2014	1763	613	1150
8	2008 sampai dengan 2015	2017	806	1211
9	2016	210	99	111

Terdapat sedikit perubahan data, sehingga data yang digunakan agak berbeda dengan jumlah data pada tulisan sebelumnya [3]. Karena data variabel dependen harus data dikotomi D(0,1) maka Indeks Prestasi Semester 1 (IPS1) dijadikan data biner yaitu :
0 < IPS1 < rata-rata IPS1 dikodekan menjadi “0” dan (rata-rata +0,0001) < IPS1 < 4,000 dikodekan menjadi “1”, dalam hal ini rata-rata diambil berdasarkan seluruh sampel. Jadi kode “0” berarti IPS1 yang dibawah rata-rata dan kode “1” berarti IPS1 di atas rata-rata.
Tahap berikutnya adalah *Data Transformation* (data ditransformasikan atau diubah menjadi bentuk yang cocok untuk diolah atau dianalisis).

Data kategori SMA (“1” untuk SMA dan “2” untuk SMK) , status SMA (“1” untuk Negeri dan “2” untuk

Swasta) , lokasi SMA ("1" untuk Jawa dan "2" untuk Luar Jawa) dan level bahasa Inggris (level_1 dikodekan menjadi "1" , level_2 dikodekan menjadi "2" , level_3 dikodekan menjadi "3" dan ESP dikodekan menjadi "4") ditransformasi dari data alfa numerik menjadi data numerik karena data regresi logistik membutuhkan data bertipe numerik.

Pada penerimaan mahasiswa yang jalur prestasi melibatkan variabel status SMA (x_1), lokasi SMA (x_2) , kategori SMA (x_3), level bahasa Inggris (x_4). Sehingga model jalur prestasi hanya mempunyai empat variabel independen x_1, x_2, x_3 dan x_4. Seperti pada kasus regresi biasa model regresi yang sudah ada dapat disederhanakan dengan uji signifikansi koefisien beta. Berikut ini merupakan bentuk model regresi logistik dan regresi multinomial sebelum dan sesudah disederhanakan.

Model A adalah model Regresi Logistik dengan menggunakan data mahasiswa angkatan 2008 , untuk jalur prestasi adalah :
$P(\mathbf{x}) = 1/(1+\exp\{-[-21{,}331- 0{,}126(x_1) - 0{,}583(x_2) + 20{,}219(x_3) +1{,}403(x_4)]\})$
dan setelah disederhanakan modelnya menjadi :
$P(\mathbf{x}) = 1/(1+\exp\{-[-2{,}113 + 1{,}438(x_4)]\})$
untuk model Regresi Multinomial pada jalur prestasi adalah :
$Y = -0{,}940 - 0{,}329(x_1) - 0{,}278(x_2) + 0{,}357(x_3) + 0{,}532(x_4)$
dan setelah disederhanakan modelnya menjadi :
$Y= 1{,}534 + 0{,}502(x_4)$

Model B adalah model Regresi Logistik dengan menggunakan data mahasiswa angkatan 2008 – 2009 , untuk jalur prestasi adalah :
$P(\mathbf{x}) = 1/(1+\exp\{-[-22{,}137 +0{,}561(x_1) - 0{,}243(x_2) + 20{,}680(x_3) + 0{,}831(x_4)]\})$
dan setelah disederhanakan modelnya menjadi :
$P(\mathbf{x}) = 1/(1+\exp\{-[-0{,}700 - 0{,}760(x_4)]\})$
untuk model regresi Multinomial pada jalur prestasi adalah :
$Y = 1{,}116 + 0{,}386(x_1) - 0{,}469(x_2) + 0{,}480(x_3) + 0{,}389(x_4)$
dan setelah disederhanakan modelnya menjadi :
$Y= 1{,}663 + 0{,}370(x_4)$

Model C adalah model Regresi Logistik dengan menggunakan data mahasiswa angkatan 2008 – 2010 , untuk jalur prestasi adalah :
$P(\mathbf{x}) = 1/(1+\exp\{-[-1{,}033 - 0{,}031(x_1) - 1{,}116(x_2) + 1{,}321(x_3) + 0{,}869(x_4)]\})$
dan setelah disederhanakan modelnya menjadi :
$P(\mathbf{x}) = 1/(1+\exp\{-[2{,}519 - 1{,}378(x_2)]\})$
untuk model regresi Multinomial pada jalur prestasi adalah :
$Y = 1{,}811 + 0{,}124(x_1) - 0{,}655(x_2) + 0{,}667(x_3) +0{,}328(x_4)$
dan setelah disederhanakan modelnya menjadi :
$Y= 2{,}703 - 0{,}639(x_2) + 0{,}314(x_4)$

Model D adalah model Regresi Logistik dengan menggunakan data mahasiswa angkatan 2008 – 2011 , untuk jalur prestasi adalah :
$P(\mathbf{x}) = 1/(1+\exp\{-[0{,}573 - 0{,}134(x_1) - 1{,}363(x_2) + 0{,}066(x_3) + 0{,}812(x_4)]\})$
dan setelah disederhanakan modelnya menjadi :
$P(\mathbf{x}) = 1/(1+\exp\{-[0{,}390 - 1{,}337(x_2) + 0{,}813(x_4)]\})$
untuk model regresi Multinomial pada jalur prestasi adalah :
$Y = 2{,}761 + 0{,}097(x_1) - 0{,}979(x_2) + 0{,}216(x_3) +0{,}277(x_4)$
dan setelah disederhanakan modelnya menjadi :
$Y= 3{,}181 - 0{,}993(x_2) + 0{,}269(x_4)$

Model E adalah model Regresi Logistik dengan menggunakan data mahasiswa angkatan 2008 – 2012 , untuk jalur prestasi adalah :
$P(\mathbf{x}) = 1/(1+\exp\{-[0{,}896 - 0{,}230(x_1) - 1{,}105(x_2) + 0{,}145(x_3) + 0{,}640(x_4)]\})$
dan setelah disederhanakan modelnya menjadi :
$P(\mathbf{x}) = 1/(1+\exp\{-[0{,}579 - 1{,}029(x_2) + 0{,}639(x_4)]\})$
untuk model regresi Multinomial pada jalur prestasi adalah :
$Y = 2{,}631 + 0{,}027(x_1) - 0{,}815(x_2) + 0{,}408(x_3) +0{,}255(x_4)$
dan setelah disederhanakan modelnya menjadi :
$Y= 2{,}689 - 0{,}821(x_2) + 0{,}403(x_3) - 0{,}254(x_4)$

Model F adalah model Regresi Logistik dengan menggunakan data mahasiswa angkatan 2008 – 2013 , untuk jalur prestasi adalah :
$P(\mathbf{x}) = 1/(1+\exp\{-[0{,}662 - 0{,}183(x_1) - 1{,}096(x_2) - 0{,}108(x_3) + 0{,}799(x_4)]\})$
dan setelah disederhanakan modelnya menjadi :
$P(\mathbf{x}) = 1/(1+\exp\{-[0{,}187 - 1{,}063(x_2) + 0{,}808(x_4)]\})$
untuk model regresi Multinomial pada jalur prestasi adalah :
$Y = 2{,}882 + 0{,}006(x_1) - 0{,}792(x_2) + 0{,}085(x_3) +0{,}290(x_4)$
dan setelah disederhanakan modelnya menjadi :
$Y= 2{,}972 - 0{,}776(x_2) + 0{,}286(x_4)$

Model G adalah model Regresi Logistik dengan menggunakan data mahasiswa angkatan 2008 – 2014 , untuk jalur prestasi adalah :
$P(\mathbf{x}) = 1/(1+\exp\{-[0{,}514 - 0{,}196(x_1) - 1{,}009(x_2) - 0{,}197(x_3) + 0{,}894(x_4)]\})$
dan setelah disederhanakan modelnya menjadi :
$P(\mathbf{x}) = 1/(1+\exp\{-[-0{,}076 - 0{,}988(x_2) + 0{,}908(x_4)]\})$
untuk model regresi Multinomial pada jalur prestasi adalah :
$Y = 2{,}924 - 0{,}001(x_1) - 0{,}726(x_2) - 0{,}090(x_3) +0{,}329(x_4)$
dan setelah disederhanakan modelnya menjadi :
$Y= 2{,}839 - 0{,}747(x_2) + 0{,}334(x_4)$

Model H adalah model Regresi Logistik dengan menggunakan data mahasiswa angkatan 2008 – 2015 , untuk jalur prestasi adalah :
$P(\mathbf{x}) = 1/(1+\exp\{-[-0{,}381 + 0{,}001(x_1) - 0{,}189(x_2) - 0{,}509(x_3) + 0{,}934(x_4)]\})$
dan setelah disederhanakan modelnya menjadi :

$P(\mathbf{x}) = 1/(1+\exp\{-[-0{,}581 - 0{,}593(x_3) + 0{,}954(x_4)]\})$
untuk non jalur prestasi adalah :
$Y = 2{,}426 - 0{,}069(x_1) - 0{,}239(x_2) - 0{,}231(x_3) + 0{,}353(x_4)$
dan model regresi Multinomial pada jalur prestasi adalah :
$Y = 2{,}580 - 0{,}249(x_2) - 0{,}258(x_3) + 0{,}353(x_4)$

Hasil rangkuman koefisien regresi logistik untuk jalur prestasi setelah disederhanakan atau setelah uji signifikansi dan nilai kecocokan model dengan data sesungguhnya dapat diringkas dalam Tabel IV dan Tabel V.

TABEL IVV
TABEL RANGKUMAN KOEFISIEN REGRESI LOGISTIK UNTUK JALUR PRESTASI SETELAH UJI SIGNIFIKANSI

Tahun	α	β_1	β_2	β_3	β_4	% cocok
2008	-2,113	0	0	0	1,438	73,0%
2008 sd 2009	-0,700	0	0	0	0,76	71,6%
2008 sd 2010	2,519	0	-1,378	0	0	70,5%
2008 sd 2011	0,90	0	-1,337	0	0,813	69,6%
2008 sd 2012	0,579	0	-1,029	0	0,639	69,1%
2008 sd 2013	0,187	0	-1,063	0	0,808	70,9%
2008 sd 2014	-0,076	0	-0,988	0	0,908	71,6%
2008 sd 2015	-0,581	0	0	-0,593	0,954	71,2%
RATA-RATA						**70,94%**

Dari Tabel IV terlihat bahwa rata-rata presentase kecocokannya adalah 70,94%. Dari Tabel 4 juga dapat terlihat bahwa :

- Untuk model A (dengan menggunakan angkatan 2008) ternyata yang berpengaruh adalah level bahasa Inggris.
- Untuk model B (dengan menggunakan angkatan 2008-2009) ternyata yang berpengaruh adalah variabel level bahasa Inggris.
- Untuk model C (dengan menggunakan angkatan 2008-2010) ternyata yang berpengaruh adalah lokasi SMA.
- Untuk model D (dengan menggunakan angkatan 2008-2011) ternyata yang berpengaruh adalah lokasi SMA dan level bahasa Inggris.
- Untuk model E (dengan menggunakan angkatan 2008-2012) ternyata yang berpengaruh adalah lokasi SMA dan level bahasa Inggris.
- Untuk model F (dengan menggunakan angkatan 2008-2013) ternyata yang berpengaruh adalah lokasi SMA dan level bahasa Inggris.
- Untuk model G (dengan menggunakan angkatan 2008-2014) ternyata yang berpengaruh adalah lokasi SMA dan level bahasa Inggris.
- Untuk model H (dengan menggunakan angkatan 2008-2015) ternyata yang berpengaruh adalah kategori SMA dan level bahasa Inggris

TABEL V
TABEL RANGKUMAN KOEFISIEN REGRESI MULTINOMIAL UNTUK JALUR PRESTASI SETELAH UJI SIGNIFIKANSI

Tahun	α	β_1	β_2	β_3	β_4	% cocok
2008	1,534	0	0	0	0,502	61,9%
2008 sd 2009	1,663	0	0	0	0,370	56,8%
2008 sd 2010	2,703	0	-0,639	0	0,314	60,5%
2008 sd 2011	3,181	0	-0,993	0	0,269	68,1%
2008 sd 2012	2,689	0	-0,821	0,403	0,254	67,3%
2008 sd 2013	2,972	0	-0,776	0	0,286	70,0%
2008 sd 2014	2,839	0	-0,747	0	0,334	70,1%
2008 sd 2015	2,580	0	-0,249	-0,258	0,353	67,0%
RATA-RATA						**65,21%**

Dari Tabel V terlihat bahwa rata-rata presentase kecocokanya adalah 65,21%. Dari Tabel V juga dapat terlihat bahwa:

- Untuk model A (dengan menggunakan angkatan 2008) ternyata yang berpengaruh adalah level bahasa Inggris.
- Untuk model B (dengan menggunakan angkatan 2008-2009) ternyata yang berpengaruh adalah variabel level bahasa Inggris.
- Untuk model C (dengan menggunakan angkatan 2008-2010) ternyata yang berpengaruh adalah lokasi SMA dan level bahasa Inggris.
- Untuk model D (dengan menggunakan angkatan 2008-2011) ternyata yang berpengaruh adalah lokasi SMA dan level bahasa Inggris.
- Untuk model E (dengan menggunakan angkatan 2008-2012) ternyata yang berpengaruh adalah lokasi SMA, kategori SMA dan level bahasa Inggris.
- Untuk model F (dengan menggunakan angkatan 2008-2013) ternyata yang berpengaruh adalah lokasi SMA dan level bahasa Inggris.
- Untuk model G (dengan menggunakan angkatan 2008-2014) ternyata yang berpengaruh adalah lokasi SMA dan level bahasa Inggris.
- Untuk model H (dengan menggunakan angkatan 2008-2015) ternyata yang berpengaruh adalah lokasi SMA, kategori SMA dan level bahasa Inggris.

Sekarang terlebih dahulu akan digunakan ke delapan model pada Tabel II tersebut untuk memprediksi kategori Indeks Prestasi Semester 1 mahasiswa FTI pada penerimaan mahasiswa angkatan 2016 melalui jalur prestasi dan persentase kebenarannya dengan tabel Crosstab pada Tabel VI.

TABEL VI
TABEL CROSSTAB PREDIKSI KATEGORI IP SEMESTER 1 MAHASISWA ANGKATAN 2016 MELALUI JALUR PRESTASI DENGAN MENGGUNAKAN REGRESI LOGISTIK MODEL A SAMPAI MODEL H

Tahun angkatan	Model	Banyaknya Isi Sel				% cocok
		n_{11}	n_{12}	n_{21}	n_{22}	
2008	Model A	0	39	0	60	60,61
2008 sd 2009	Model B	0	39	0	60	60,61
2008 sd 2010	Model C	0	39	0	60	60,61
2008 sd 2011	Model D	32	7	20	40	72,72
2008 sd 2012	Model E	32	7	20	40	72,72
2008 sd 2013	Model F	22	17	9	51	73,73
2008 sd 2014	Model G	32	7	20	40	72,72
2008 sd 2015	Model H	23	16	15	45	68,69
Rata-rata						**67,80125**
Standar Deviasi						**6,1373**

Kemudian digunakan ke delapan model pada Tabel III untuk memprediksi kategori Indeks Prestasi Semester 1 mahasiswa FTI pada penerimaan mahasiswa angkatan 2016 melalui jalur prestasi dengan model Regresi Multinomial dan persentase kebenarannya dengan tabel Crosstab Tabel VII.

TABEL VII
TABEL CROSSTAB PREDIKSI KATEGORI IP SEMESTER 1 MAHASISWA ANGKATAN 2016 MELALUI JALUR PRESTASI DENGAN MENGGUNAKAN REGRESI MULTINOMIAL MODEL A SAMPAI MODEL H

Tahun angkatan	Model	Banyaknya Isi Sel				% cocok
		n_{11}	n_{12}	n_{21}	n_{22}	
2008	Model A	40	4	32	23	63,64
2008 sd 2009	Model B	37	4	35	23	60,61
2008 sd 2010	Model C	36	7	22	34	70,71
2008 sd 2011	Model D	36	7	25	31	67,68
2008 sd 2012	Model E	34	11	24	30	64,65
2008 sd 2013	Model F	36	7	25	31	67,68
2008 sd 2014	Model G	36	7	22	34	70,71
2008 sd 2015	Model H	35	8	20	36	71,72
Rata-rata						**67,175**
Standar Deviasi						**3,93**

TABEL VIII
TABEL PERSENTASE KEBENARAN PREDIKSI KATEGORI IP SEMESTER 1 MAHASISWA ANGKATAN 2016 JALUR PRESTASI DENGAN MENGGUNAKAN REGRESI LOGISTIK MODEL A SAMPAI DENGAN MODEL H.

Model Prediksi	Model A	Model B	Model C	Model D	Model E	Model F	Model G	Model H
Persentase Kebenaran	60,61 %	60,61%	60,61%	72,72%	72,72%	73,73%	72,72%	68,69%

Dari Tabel VIII terlihat bahwa rata-rata persentase kebenaran untuk prediksi Kategori IP Semester 1 Mahasiswa Angkatan 2016 melalui Jalur Prestasi dengan Menggunakan Model A sampai dengan Model H adalah 67,80125 %

TABEL IX
TABEL PERSENTASE KEBENARAN PREDIKSI KATEGORI IP SEMESTER 1 MAHASISWA ANGKATAN 2016 JALUR PRESTASI DENGAN MENGGUNAKAN REGRESI MULTINOMIAL MODEL A SAMPAI DENGAN MODEL H

Model Prediksi	Model A	Model B	Model C	Model D	Model E	Model F	Model G	Model H
Persentase Kebenaran	63,64%	60,61%	70,71%	67,68%	64,65%	67,68%	70,71%	71,72%

Dari Tabel IX terlihat bahwa rata-rata persentase kebenaran untuk prediksi kategori IP Semester 1 mahasiswa angkatan 2016 melalui non jalur prestasi dengan menggunakan Model A sampai dengan Model H adalah 67,175 %.

V. KESIMPULAN

Dari penelitian yang telah dilakukan, maka didapat beberapa kesimpulan sebagai berikut:

1. Hasil rata-rata akurasi pemodelan kategori IP semester 1 mahasiswa jalur prestasi dengan menggunakan Regresi Logistik adalah 70,94%, sedangkan bila menggunakan Regresi Multinomial adalah 65,21%.
2. Hasil rata-rata akurasi prediksi kategori IP semester 1 mahasiswa angkatan tahun 2016 jalur prestasi dengan menggunakan Regresi Logistik adalah 67.80125%, sedangkan bila menggunakan Regresi Multinomial adalah 67,175%.
3. Dapat disimpulkan bahwa untuk rata-rata akurasi pemodelan lebih tinggi menggunakan Regresi Logistik. Tetapi untuk kestabilan model prediksi terlihat bahwa model Regresi Multinomial lebih stabil daripada model Regresi Logistik, hal ini dapat dilihat dari standar deviasi model Regresi Multinomial sebesar 3,93 yang lebih rendah dari standar deviasi model Regresi Logistik yang sebesar 6,1373.
4. Dari dua model regresi tersebut apabila digunakan untuk memprediksi kategori IP semester 1 mahasiswa FTI angkatan 2016 mempunyai akurasi hampir sama yaitu sekitar 67%.

UCAPAN TERIMA KASIH

Peneliti mengucapkan terima kasih kepada Fakultas Teknologi Informasi Universitas Kristen Duta Wacana Yogyakarta atas bantuan dana dan infrastruktur sehingga penelitian ini dapat berjalan dengan baik dan lancar.

REFERENSI

[1] R. G. Santosa and H. Setiadi, “Analisis Faktorial untuk Uji Pengaruh Beberapa Faktor Terhadap Indeks Prestasi Mahasiswa Fakultas Teknologi Informasi UKDW,” Unpublished, Yogyakarta, 2015.

[2] R. G. Santosa and A. R. Chrismanto, “Regresi Logistik untuk Prediksi Kategori IP Mahasiswa Fakultas Teknologi Informasi UKDW,” Unpublishes, Yogyakarta, 2016.

[3] R. G. Santosa and A. R. Chrismanto, “Logistic Regression Model for Predicting First Semester Students GPA category Based on High School Academic Achievement,” *Researchersworld Journal of Arts, Science & Commerce Volume-VIII Issue-2(1) April 2017*, pp. 58-66, 2017.

[4] M. H. Dunham, Data Mining : Introductory and Advanced Topics, New York USA: Prentice Hall PTR Upper Saddle River , 2002.

[5] A. Berson, S. Smith and K. Thearling, Building Data Mining Applications for CRM, MacGraw Hill Company, 2000.

[6] J. Han, M. Kamber and J. Pei, Data Mining : Concepts and Techniques, Morgan Kaufmann Publishers, 2011.

[7] M. Kantardzic, Data Mining Concepts, Models, Methods and Algorithms, IEEE Press and Wiley-Interscience, 2003.

[8] O. P. Rud, Data Mining Cook Book Modeling Data for Marketing, Risk and Customer Relationship Management, John Wiley and Sons Inc., 2001.

[9] C. Romero and S. Ventura, “Data Mining in Education,” *WIREs Data Mining Knowledge Discovery 2013, 3*, pp. 12-27, 2013.

[10] B. K. Baradwaj and S. Pal, "Mining Educational Data to Analyze Students' Performance," *International Journal of Advanced Computer Science and Applications (IJACSA),* vol. 2, no. 6, pp. 63-69, 2011.

[11] R. Asif, A. Merceron, S. A. Ali and N. G. Haider, "Analyzing Undergraduate Students' Performance Using Educational Data Mining," *Elsevier Computers & Education,* vol. 113, no. 2017, pp. 177-194, 2017.

[12] S. L. Prabha and A. Shanavas, "Educational Data Mining Applications," *Operations Research and Applications: An International Journal (ORAJ),* vol. 1, no. 1, pp. 23 - 29, 2014.

[13] D. G. Kleinbaum and M. Klein, Logistic regression : A self-Learning Text Second edition, New York: Springer Verlag Inc., 2002.

[14] D. Alverina, A. R. Chrismanto and R. G. Santosa, "Perbandingan Akurasi Algoritma C4.5 dan CART dalam Memprediksi Kategori Indeks Prestasi Mahasiswa," *Jurnal Teknologi dan Sistem Komputer Volume 6 , No. 2 April 2018,* pp. 76-83, 2018.

[15] R. G. Santosa and A. R. Chrismanto, "Perbandingan Akurasi Model Regresi Logistik untuk Prediksi Kategori IP Mahasiswa Jalur Prestasi dengan Non Jalur Prestasi," *Jurnal Teknik dan ilmu Komputer Volume 07 No 25 Januari - Maret 2018,* pp. 107-121, 2018.

[16] D. W. Hosmer and S. Lemeshow, Applied Logistic Regression, John Wiley and Sons, Inc., 2000.

[17] G. K. Bhattachryya and R. A. Johnson, Statistical Principles and Methods 6th edition, John Wiley & Sons, Inc, 2010.