PERBANDINGAN KETEPATAN KLASIFIKASI ANTARA METODE REGRESI LOGISTIK DAN KLASIFIKASI POHON PADA KASUS PROGRAM WAJARDIKDAS 9 TAHUN

Andhita Dessy Wulansari

Jurusan Tarbiyah STAIN Ponorogo

Abstract: *Wajardikdas 9 years is a compulsory government program that requires the Indonesian people to get a formal education certificate minimally at junior high school level/equivalent. In order to succeed the program, the availability of information relating to the classification of rural/village is absolutely important. Due to this information, the regency/city government can more easily prioritize which rural/village that needs a special attention. There are at least two statistical methods: 1) binary logistic regression and 2) logistic regression. This study compared the results of the classification accuracy between the logistic regression method and tree classification to get a more precise technique used in classifying the rural/village in case of 9 years Wajardikdas program. By using the data testing, it showed that the classification accuracy produced by the method of tree classification was higher than that of logistic regression. Therefore, it could be concluded that tree classification was more effective method to use. This was because the factors that affected the success of Wajardikdas 9 years like the ratio of student/teacher and student/school had been known.*

الملخص: كانت فرضية الدراسة لمدة ٩ سنوات برنامجا حكوميا وطنيا لجميع المواطنين، ليحصلوا على شهادة المرحلة المتوسطة. ولإنجاح هذا البرنامج فتوافر الإعلامات المتعلقة بتصنيف القرى مهمة جدّا، فبهذا الإعلام سهلت على الحكومات المديرية لوضع معيار الأولويات للقرى المحتاجة إلى الاهتمام البالغ. فعلى الأقلّ هناك طريقتان احصائيتان: (١) الانحدار اللوجستي الثنائي (٢) الانحدار اللوجستى. في هذا البحث تُعمل المقارنة بين نتيجة التصنيف بطريقة الانحدار اللوجستى والتصنيف الشجرى للحصول على أنسب الطرق المطبّقة في تصنيف القرى في البرنامج وجوب الدراسة لمدة ٩ سنوات. فعلى أساس البيانات بطريقة الاختبار، ظهرت النتيجة أن نتيجة التصنيف بطريقة التصنيف الشجرى أعلى من طريقة الانحدار اللوجستى. لذايقرّر بأن الطريقة المناسبة في هذا هي التصنيف الشجرى، لأن العوامل المؤثرة في حالة إنجاز وجوب الدراسة لمدة ٩ سنوات هي نسبة الطلاب/والمدرسين ونسبة الطلاب/والمدرسة وستُعرف هذه الحالة.

Keywords: *Klasifikasi, regresi logistic, klasifikasi pohon, wajardikdas*

PENDAHULUAN

Pendidikan merupakan salah satu modal bangsa yang kualitasnya harus terus ditingkatkan, sehingga dimasa datang diharapkan manusia-manusia Indonesia dapat menempatkan bangsanya pada tempat yang terhormat dalam pergaulan antar bangsa sedunia[1]. Dalam upaya mewujudkan tujuan tersebut, pada tahun 1994 pemerintah telah mencanangkan Program Wajib Belajar Pendidikan Dasar (Wajardikdas) 9 tahun yang mewajibkan seluruh penduduk Indonesia untuk memiliki ijazah minimal SMP/sederajat. Program Wajardikdas 9 tahun ini dapat dikatakan tuntas jika seluruh daerah yang ada dapat mencapai kategori Angka Partisipasi Kasar (APK) tuntas paripurna, yakni APK yang dicapai sekurang-kurangnya 95 persen atau dengan kata lain rasio antara jumlah siswa berapapun usianya yang sedang bersekolah ditingkat SMP/sederajat terhadap jumlah penduduk kelompok usia resmi yang berkaitan dengan jenjang pendidikan SMP/ sederajat (usia 13-15 tahun) sekurang-kurangnya 95 persen[2].

Dewasa ini otonomi daerah membawa masalah yang tidak sedikit bagi pemerintah kabupaten/kota, karena semua kebijakan yang semula tersentralisasi di pemerintah pusat kini harus ditangani sendiri oleh pemerintah kabupaten/ kota tanpa adanya campur tangan dari pemerintah pusat. Untuk menentukan kebijakan yang relevan dengan tujuan pembangunan nasional, pemerintah kabupaten/kota perlu memperhatikan banyak aspek terutama aspek pendidikan karena kualitas suatu kabupaten/kota tercermin dari kualitas pendidikannya. Beberapa hasil studi juga menunjukkan bahwa pendidikan tidak lagi dipandang sebagai pemborosan tetapi suatu investasi Sumber Daya Manusia (SDM) yang dapat mempengaruhi aspek-aspek lainnya. Hal ini terbukti dengan adanya penelitian yang dilakukan oleh Kementerian Pendidikan Nasional pada tahun 2003, menurut penelitian tersebut tingkat pendidikan penduduk selain berkorelasi positif terhadap status ekonomi penduduk yang diukur melalui *Purchasing Power Parity* (PPP) juga berkorelasi positif terhadap laju pertumbuhan penduduk dan derajat kesehatan penduduk[3].

Saat ini dikembangkan metode klasifikasi pohon yang lebih mudah, praktis dalam penggunaan dan interpretasinya jika dibandingkan dengan metode regresi logistik karena hasil klasifikasi yang diperoleh dapat dicari dengan cara menelusuri pohon klasifikasinya[4]. Metode ini juga digunakan untuk

---

[1] Santoso, I.S, *Pembinaan Watak Tugas Utama Pendidikan* (Jakarta: U-I Press, 1981), 8

[2] *Ibid*,

[3] Suryadi, A., dan Untung, "Gerakan Pemberantasan Buta Aksara Intensif", dalam *Aksara,* (Jakarta: Direktorat Pendidikan Masyarakat, 2005), 10-13

[4] *Ibid,*

menggambarkan hubungan antara variabel respon dengan satu set variabel prediktor baik itu berupa data kategori maupun kontinyu. Klasifikasi pohon juga dikenal sebagai metode pemilahan rekursif secara biner yang berarti sekelompok data yang terkumpul dalam satu ruang yang disebut simpul dapat dipilah menjadi dua simpul anak dan setiap simpul anak dapat dipilah lagi menjadi dua simpul anak, begitu seterusnya dan berhenti jika memenuhi kriteria tertentu. Model pohon yang dihasilkan bergantung pada skala variabel respon. Jika data variabel respon kontinyu maka model pohon yang dihasilkan adalah pohon regresi, apabila data variabel respon kategori, maka model pohon yang dihasilkan adalah pohon klasifikasi.

Metode regresi logistik merupakan metode yang cukup robust untuk dapat diterapkan pada berbagai keadaan data. Metode ini seringkali dibandingkan dengan metode non parametrik seperti klasifikasi pohon yang tidak memerlukan asumsi-asumsi mengikat[5]. Oleh karena itu perbandingan analisa data dengan menggunakan kedua metode ini banyak diterapakan dalam berbagai bidang, terutama pada bidang kesehatan. Seperti yang dilakukan oleh Camdeviren, dkk pada tahun 2007 yang membandingkan metode klasifikasi pohon dengan regresi logistik pada kasus postpartum depression dan Kurt dkk pada tahun 2008 yang membandingkan metode klasifikasi pohon, regresi logistik dengan neural network pada kasus *coronary artery disease*. Dari kedua penelitian ini didapatkan hasil bahwa metode klasifikasi pohon mempunyai ketepatan klasifikasi yang lebih tinggi dibanding regresi logistik. Pada penelitian ini juga akan dibandingkan hasil ketepatan klasifikasi antara metode regresi logistik dan klasifikasi pohon tetapi pada kasus pendidikan yaitu klasifikasi desa/kelurahan berdasarkan variabel yang diduga berpengaruh terhadap kondisi ketuntasan Wajardikdas yang dicapai.

MODEL REGRESI LOGISTIK

Metode regresi logistik adalah prosedur pemodelan yang diterapkan untuk memodelkan variabel respon (y) yang bersifat kategori berdasarkan satu atau lebih variabel prediktor (x), baik itu yang bersifat kategori maupun kontinyu. Apabila variabel responnya (y) terdiri dari 2 kategori yaitu y=1 (sukses) dan y=0 (gagal) maka metode regresi logistik yang dapat diterapkan adalah regresi logistik biner. Secara umum model probabilitas regresi logistik dengan melibatkan beberapa variabel prediktor (x) dapat diformulasikan sebagai berikut:

---

[5] Feldesman, M.C. "Classification Trees as An Alternative to Linier Discriminant Analysis", *American Journal of Physical Anthropology,* Vol. 119, 2002, 257.

$$(1)\ \pi(x) = \frac{e^{(\beta_0+\beta_1 x_1+\beta_2 x_2+...+\beta_p x_p)}}{1+e^{(\beta_0+\beta_1 x_1+\beta_2 x_2+...+\beta_p x_p)}}$$

Fungsi π (x) merupakan fungsi non linier sehingga perlu dilakukan transformasi logit untuk memperoleh fungsi yang linier agar dapat dilihat hubungan antara variabel respon (y) dengan variabel prediktornya (x). Bentuk logit dari $\pi(x)$, adalah $g(x) = \ln\left(\frac{\pi(x)}{1-\pi(x)}\right)$ sehingga setelah persamaan (1) disubstitusikan pada $\pi)\theta)$ maka diperoleh,

$$(2)\quad g(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p$$

Nilai $\pi(x)$ pada persamaan (1) merupakan pedoman klasifikasi suatu obyek. Jika $\pi_1(x) = \pi(x)$ maka obyek diklasifikasikan ke kategori y=1 (sukses) atau y=0 (gagal) didasarkan pada kedua nilai tersebut yang terbesar[6].

METODE KLASIFIKASI POHON

Salah satu metode non parametrik yang dapat digunakan untuk menggambarkan hubungan antara variabel respon yang kategori dengan satu set variabel prediktor yang kategori dan kontinyu adalah klasifikasi pohon. Pohon klasifikasi terbentuk melalui pemilahan tiap simpul menjadi dua himpunan bagian turunan. Proses pemilahan dimulai dari simpul utama yang berisi data yang akan dipilah. Pemilahan dilakukan untuk memilah data menjadi dua kelompok yaitu, kelompok yang masuk simpul kiri dan simpul kanan. Pemilahan dilakukan pada tiap simpul sampai didapatkan simpul akhir. Variabel yang memilah pada simpul utama adalah variabel terpenting dalam menduga kelas amatan.

Ukuran pohon yang sangat besar akan dapat memberikan nilai Resubtitution Estimate R(T) yang sangat kecil. Sehingga pohon ukuran ini sering dipilih untuk menduga respon, tetapi ukuran pohon yang besar ini akan menyebabkan nilai cost complexity yang tinggi karena data yang digambarkan cenderung kompleks. Jika R(T) dipilih sebagai penduga terbaik, maka akan cenderung dipilih pohon ukuran terbesar, sebab semakin besar pohon akan semakin kecil

[6] Wibowo, W., "Perbandingan Hasil Ketepatan Klasifikasi Analisis Diskriminan dan Regresi Logistik pada Pengklasifikasian Data Respon Biner"

R(T). Resubtitution Estimate (penduga pengganti) adalah proporsi amatan yang mengalami kesalahan pengklasifikasian yaitu :

$$R(T) = \frac{1}{N}\sum_{n=1}^{N} X(d(x_n) \neq j_n) \quad (3)$$

dimana X(.) adalah fungsi indikator berbentuk,

$$X(.) = \begin{cases} 1, & \text{Jika pertanyaan dalam tanda kurung benar} \\ 0, & \text{Jika pertanyaan dalam tanda kurung salah} \end{cases}$$

Salah satu penduga yang dapat digunakan untuk mendapatkan pohon klasifikasi optimal (terbaik) adalah penduga sampel uji (*test sample estimate*). Pada penduga sample uji, sample L dibagi menjadi dua yaitu L1 (*Learning set*) dan L2 (*Testing set*). Amatan dalam L1 digunakan untuk membentuk pohon T, sedangkan amatan-amatan dalam L2 digunakan untuk menduga $R^s(T_t)$.

PROGRAM WAJIB BELAJAR

Program Wajib Belajar adalah program pendidikan minimal yang harus diikuti oleh warga negara Indonesia atas tanggung jawab pemerintah pusat dan daerah. Program Wajib Belajar dapat dikatakan berhasil jika seluruh daerah dapat mencapai Angka Partisipasi Kasar (APK) sekurang-kurangnya 95 persen. Besarnya APK disuatu daerah dapat dihitung dengan rumusan sebagai berikut:

$$(4)\ APK = \frac{\text{Jumlah murid pada jenjang pendidikan tertentu}}{\text{Jumlah penduduk kelompok usia yang sesuai}} x100\%$$

Hasil perhitungan APK ini digunakan untuk mengetahui banyaknya anak yang bersekolah di suatu jenjang pendidikan tertentu pada wilayah tertentu. Dalam perhitungan angka partisipasi terdapat empat kategori tuntas[7] yaitu :

Tuntas Paripurna : APK95≥ persen

Tuntas Utama : 90 persen ≤APK<94 persen

Tuntas Madya : 84 persen ≤APK<90 persen

Tuntas Pratama : 80 persen ≤APK<84 persen

[7] Sukriswandari, N, *Upaya Direktorat Pembinaan SMP dalam Penuntasan Wajar 9 Tahun*, (Jakarta: Direktorat Pembinaan Sekolah Menengah Pertama, 2008)

Untuk mengukur keberhasilan program Wajardikdas 9 tahun, ada beberapa indikator yang dapat digunakan, yaitu rasio murid/sekolah, rasio murid/kelas, rasio murid/guru, rasio kelas/ruang belajar, angka mengulang, angka putus sekolah dan angka lulusan[8].

KLASIFIKASI OBYEK PENGAMATAN

Proses awal sebelum dilakukannya klasifikasi obyek pengamatan baik dengan menggunakan metode regresi logistik dan klasifikasi pohon adalah membagi data menjadi 2 bagian secara random yaitu data learning dan data testing. Data learning adalah data yang digunakan untuk pembentukan model pohon sedangkan data testing adalah data yang digunakan untuk menguji ketepatan model pohon yang telah dibentuk oleh data learning. Tidak ada aturan khusus yang digunakan dalam pembagian data learning dan data testing, oleh karena itu disini akan dicoba 9 kondisi data yang berbeda, yang masing-masing data learning dan testing adalah 95% (137 pengamatan) dan 5% (7 pengamatan), 90% (130 pengamatan) dan 10% (14 pengamatan), 85% (122 pengamatan) dan 15% (22 pengamatan), 80% (115 pengamatan) dan 20% (29 pengamatan), 75% (108 pengamatan) dan 25% (36 pengamatan), 70% (101 pengamatan) dan 30% (43 pengamatan), 65% (94 pengamatan) dan 35% (50 pengamatan), 60% (86 pengamatan) dan 40% (58 pengamatan), 55% (79 pengamatan) dan 45% (65 pengamatan). Masing-masing 9 kondisi data yang berbeda ini diolah dengan 2 metode yaitu regresi logistik biner dan klasifikasi pohon kemudian dihitung ketepatan klasifikasi dari masing-masing model yang dihasilkan dan dibandingkan. Dari hasil perbandingan ini akan didapatkan metode yang lebih tepat digunakan untuk mengklasifikasikan desa/kelurahan yang ada di Kabupaten Gresik. Ketepatan klasifikasi pada data testing dijadikan sebagai dasar penentuan metode yang mempunyai ketepatan lebih tinggi. Hal ini dikarenakan data testing merupakan penguji model klasfikasi yang telah diperoleh dari data learning sehingga tinggi rendahnya ketepatan pada data testing menunjukkan ketepatan model klasifikasi yang dibentuk.

KLASIFIKASI DENGAN METODE REGRESI LOGISTIK

Regresi logistik merupakan salah satu metode statistika yang dapat digunakan dalam pengklasifikasian obyek pengamatan/observasi. Untuk mengetahui

---

[8] Utomo, I., *Laporan Keterangan Pertanggungjawaban Akhir Masa Jabatan Gubernur 2003-2008*, (Surabaya: Kantor Pemerintahan Daerah Tingkat Propinsi Jawa Timur, 2008).

kehandalan dari metode ini maka diberlakukan beberapa pengkondisian data seperti yang telah disebutkan sebelumnya.

Tabel 1: Model logit Regresi logistik Biner

NO	MODEL LOGIT	NILAI G
1	KTGR = 0.403 - 0.263MG + 0.007MS	7.504
2	KTGR = 0.361 - 0.247MG + 0.066MS	7.242
3	KTGR = 1.179 - 0.154MG + 0.004MS	3.737
4	KTGR = 1.159 - 0.553MG + 0.015MS	24.181
5	KTGR = 1.167 - 0.549MG + 0.015MS	24.743
6	KTGR = 0.896 - 0.303MG + 0.006MS	12.954
7	KTGR = 0.834 - 0.288MG + 0.006MS	12.288
8	KTGR = 0.771 + 0.278MG + 0.006MS	12.016
9	KTGR = 0.730 + 0.273MG + 0.006MS	12.036

Proses awal analisa data dengan menggunakan metode regresi logistik adalah pembentukan model regresi logistik multivariabel yang melibatkan tujuh variabel prediktor dengan dua kategori respon, dimana model yang terbentuk hanya satu. Setelah didapatkan model regresi logistik multivariabel, maka langkah selanjutnya adalah melakukan elimanisi Backward Wald pada model tersebut dengan cara membuang satu persatu variabel yang mempunyai tingkat signifikansi terendah. Setelah itu didapatkan model baru yang melibatkan variabel-variabel signifikan. Adapun model logit baru yang terbentuk dapat dilihat pada Tabel 1. Model tersebut dapat dibentuk menjadi $\pi(x_i)$ yang menyatakan probabilitas suatu pengamatan untuk diklasifikasikan ke kelompok desa/kelurahan yang tuntas Wajardikdas. Pada model logit baru yang terbentuk tersebut dilakukan uji signifikansi melalui uji serentak dan uji individu. Dalam pengujian serentak, digunakan nilai G dalam pengujian signifikansi modelnya. Hipotesisnya adalah sebagai berikut :

H0: $0 = 2\beta = 1\beta$

H1: minimal ada satu $\beta k0 \neq$, $k = 1,2$

Berdasarkan output software Minitab didapatkan nilai statistik uji G seperti pada Tabel 1, sedangkan nilai $\chi^2_{(2;0.6}$) adalah 5.991. Perbandingan kedua nilai tersebut menunjukkan bahwa nilai statistik uji G lebih besar dari $\chi^2_{(2;0.6}$) yang berarti didapatkan keputusan untuk tolak H0. Berdasarkan keputusan ini

diketahui bahwa secara serentak model logit pada Tabel 1 signifikan pada tingkat kepercayaan 95%. Setelah ditunjukkan bahwa secara serentak model signifikan, yang berarti minimal ada satu parameter yang signifikan secara individu maka dilakukan uji individu (Wald test). Uji ini digunakan untuk mengetahui variabel prediktor mana saja yang signifikan secara individu. Dengan digunakan hipotesis sebagai berikut :

H0: β k=0

H1: β k0$\neq$, k=1,2

Berdasarkan hasil perhitungan dengan bantuan software Minitab, diketahui bahwa semua variabel signifikan. Hal ini diketahui dari nilai p value yang kurang dari $0.05=\alpha$ atau suatu nilai statistik uji $|W|$ yang lebih dari $1.96=Z_{\alpha/2}$ yang berarti tolak H0. Jadi kedua variabel tersebut signifikan menjadi variabel pembeda pada tingkat kepercayaan 95% dan fungsi klasifikasi dari model logit pada Tabel 1 merupakan yang terbaik.

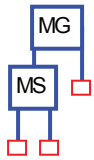
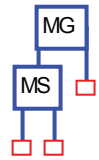
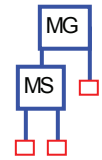
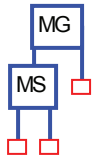
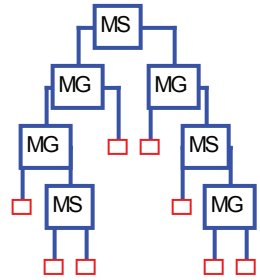
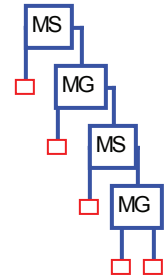
Dengan cara memasukkan nilai rasio murid/guru (MG) dan rasio murid/sekolah (MS) yang dimiliki oleh tiap pengamatan pada fungsi probabilitas $\pi(x_i)$ maka didapatkan π_2 yang merupakan probabilitas suatu pengamatan untuk diklasifikasikan ke kelompok desa/kelurahan yang tuntas Wajardikdas dan π_1 yang didapatkan dari 1-π_2 yang merupakan probabilitas suatu pengamatan untuk diklasifikasikan ke kelompok desa/kelurahan yang tidak tuntas Wajardikdas. Pengklasifikasian desa/kelurahan pada kategori tuntas Wajardikdas dan tidak tuntas Wajardikdas ini didasarkan pada kedua nilai tersebut (π_1 dan π_2) yang terbesar (maksimum).

KLASIFIKASI DENGAN METODE KLASIFIKASI POHON

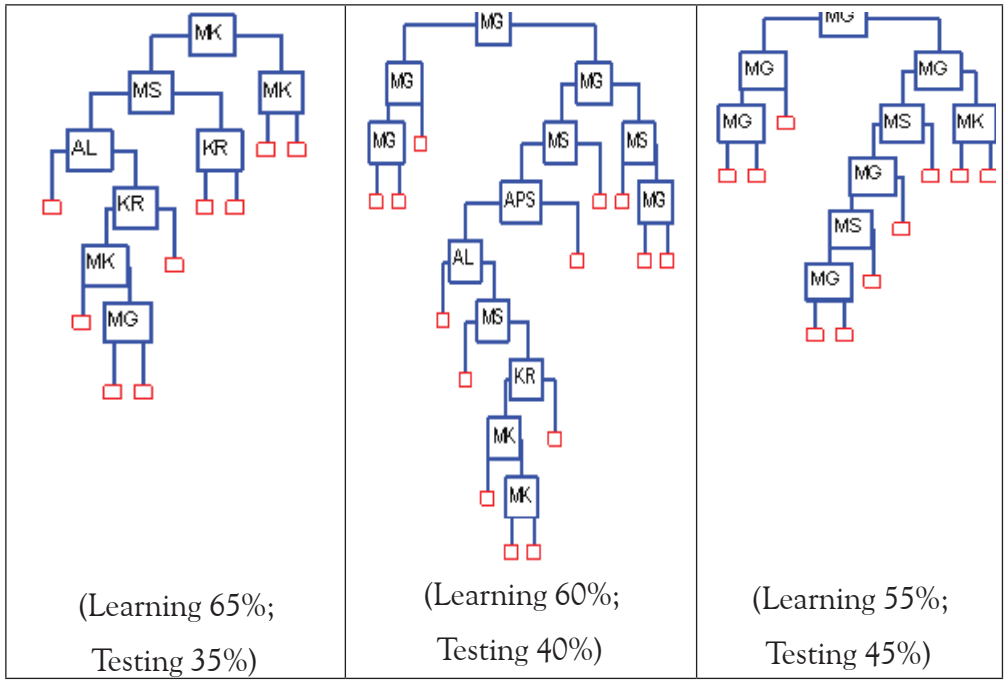
Selain metode regresi logistik, salah satu metode alternatif yang dapat digunakan dalam pengklasifikasian obyek pengamatan/observasi adalah klasifikasi pohon. Pada pengklasifikasian dengan menggunakan metode ini juga diberlakukan sembilan pengkondisian data.

Langkah awal yang perlu dilakukan sebelum mendapatkan pohon klasifikasi optimal adalah membentuk pohon klasifikasi maksimal. Dalam pembentukan pohon klasifikasi maksimal, proses awal yang dilakukan adalah mencari semua kemungkinan pemilahan yang mungkin dari variabel prediktor yang menjadi pemilah utama (primary spliter). Pemilah yang terpilih akan membentuk simpul akar (root node) yang kemudian akan disebut sebagai simpul 1. Setiap pemilahan akan menghasilkan 2 kelas yaitu simpul anak kiri dan simpul anak kanan yang kemudian disebut sebagai simpul 2 dan simpul 3. Pemilahan akan dilakukan

di tiap simpul sampai didapatkan simpul akhir. Tujuan pemilahan ini adalah untuk mengurangi keheterogenan pada simpul utama dan memaksimalkan ukuran kehomogenan dari masing-masing simpul anak relatif dari simpul induknya. Kriteria yang digunakan dalam dalam kasus ini adalah indeks Gini karena menurut Breiman[9], indeks Gini ini lebih mudah dan sesuai untuk diterapkan dalam berbagai kasus dan mempunyai perhitungan yang sederhana dan cepat. Indeks Gini berusaha untuk memisahkan kelas-kelas dengan cara memfokuskan pada satu kelas (simpul anak kiri atau kanan). Indeks Gini akan selalu memisahkan kelas yang anggotanya paling besar terlebih dahulu atau yang merupakan kelas terpenting dalam simpul tersebut. Pemilah yang memberikan penurunan keheterogenan tertinggi merupakan pemilah terbaik.

 (Learning 95%; Testing 5%)	 (Learning 90%; Testing 10%)	 (Learning 85%; Testing 15%)
 (Learning 80%; Testing 20%)	 (Learning 75%; Testing 25%)	 (Learning 70%; Testing 30%)

[9] L, Breiman, et.al., *Classification and Regression Trees* (London-New York: Chapman and Hall, 1984), 47



Gambar 1: Pohon Klasifikasi Optimal untuk Semua Kondisi Data

Pohon klasifikasi maksimal yang diperoleh memang menghasilkan nilai resubtitution estimate/resubtitution relative cost terkecil. Tetapi apabila ukuran pohon ini dipilih sebagai pohon klsifikasi terbaik, maka pohon klasifikasi ini cenderung tidak dapat menduga data testing dengan baik. Karena dengan ukuran pohon yang sangat besar ini struktur data yang digambarkan cenderung kolmpleks. Oleh karena itu perlu dipilih ukuran pohon optimal yang berukuran sederhana tetapi memilliki nilai test sample estimate/test set relative cost terkecil. Setelah didapatkan pohon optimal seperti pada Gambar 1, maka langkah selanjutnya adalah menelusuri pohon klasifikasi terbaik tersebut dengan menggunakan data respon (learning set dan testing set).

PERBANDINGAN KETEPATAN KLASIFIKASI ANTARA KEDUA METODE

Berikut diberikan perbandingan ketepatan klasifikasi antara metode regresi logistik dan klasifikasi pohon secara ringkas pada Tabel 2.

Tabel 2: Ketepatan Klasifikasi Kedua Metode

KLASIFIKASI DATA	TOTAL ACCURACY RATE			
	Regresi Logistik		Klasifikasi Pohon	
	Learning (%)	Testing (%)	Learning (%)	Testing (%)
Learning 95%; Testing 5%	60.58	42.86	62.04	71.43
Learning 90%; Testing 10%	60.77	50.00	61.54	71.43
Learning 85%; Testing 15%	66.39	54.54	61.48	68.18
Learning 80%; Testing 20%	68.70	55.17	61.74	65.52
Learning 75%; Testing 25%	70.37	55.56	77.78	55.56
Learning 70%; Testing 30%	70.30	51.16	71.29	58.14
Learning 65%; Testing 35%	56.38	54.00	80.85	56.00
Learning 60%; Testing 40%	63.95	53.45	91.86	56.90
Learning 55%; Testing 45%	65.82	50.76	82.28	55.85

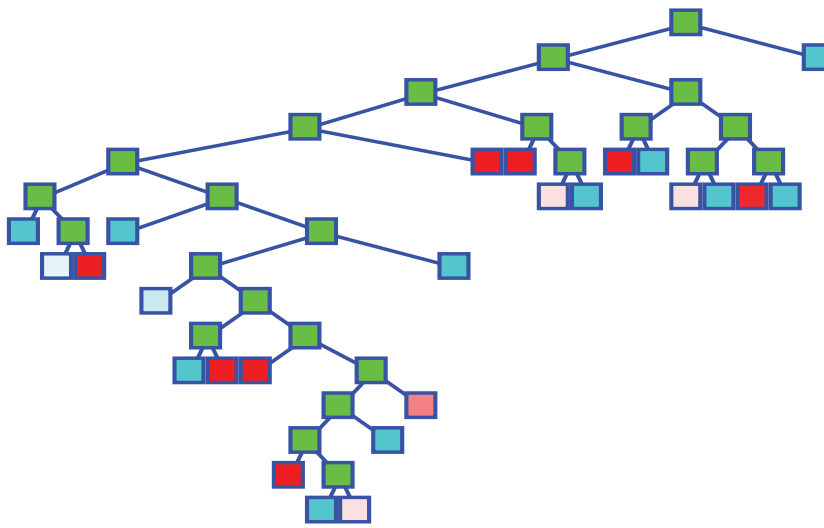
ANALISA DATA DENGAN KLASIFIKASI POHON

Setelah didapatkan hasil bahwa metode yang lebih tepat digunakan dalam pengklasifikasian desa/kelurahan di Kabupaten Gresik pada kasus Wajardikdas 9 tahun adalah klasifikasi pohon, maka metode ini akan digunakan dalam analisa data lebih lanjut. Melalui metode klasifikasi pohon akan didapatkan pohon optimal sebagai pohon klasifikasi terbaik bagi data dan juga akan didapatkan variabel-variabel prediktor yang secara signifikan berpengaruh terhadap kondisi ketuntasan Wajardikdas 9 tahun. Ada dua komponen penting yang harus disiapkan sebelum dilakukannya analisa data dengan menggunakan metode klasifikasi pohon, yaitu data *learning* dan data *testing*.

Pohon Klasifikasi Maksimal

Gambar 2 berikut, adalah pohon klasifikasi maksimal yang terbentuk dari tuju variabel prediktor yang diduga berpengaruh terhadap kondisi ketuntasan desa/kelurahan di Kabupaten Gresik pada program Wajardikdas 9 tahun. Berdasarkan Gambar tersebut, dapat diketahui bahwa pohon klasifikasi maksimal yang terbentuk mempunyai kedalaman 15. Kedalaman tersebut merupakan level atau tingkatan dalam pohon maksimal, setiap level terdiri dari beberapa

simpul. Kedalaman dihitung dari simpul utama (berwarna hijau) sampai simpul terminal, semakin besar kedalamannya maka ukuran pohon klasifikasi yang dihasilkan juga semakin besar. Kedalaman satu terletak pada simpul utama, kedalaman dua terletak dibawah simpul utama dan seterusnya sampai kedalaman 15, dimana kedalaman 15 terletak pada simpul akhir atau disebut juga simpul terminal. Sedangkan simpul terminal yang dihasilkan oleh pohon klasifikasi maksimal adalah 25 simpul. Masing-masing simpul terminal ditandai dengan label kelas dan dapat terjadi simpul terminal dengan label kelas yang sama. Pada Gambar 2 juga dapat dilihat adanya perbedaan warna pada masing-masing simpul terminal. Perbedaan warna ini menunjukkan adanya perbedaan label kelas pada masing-masing simpul tersebut. Pada simpul terminal, masing-masing ditandai dengan warna yang berbeda sesuai dengan label kelasnya. Untuk simpul terminal berwarna biru menunjukkan simpul dengan label kelas 1 dimana prosentase jumlah pengamatan yang tidak tuntas Wajardikdas pada simpul tersebut mendekati 100%. Warna biru perlahan jadi biru muda jika prosentase pengamatan yang tidak tuntas Wajardikdas pada simpul tersebut berkisar antara 50% sampai 75%. Sementara untuk simpul warna merah menunjukkan pada simpul tersebut ditandai dengan label kelas 2 dengan prosentase jumlah pengamatan yang tuntas Wajardikdas pada simpul tersebut mendekati 100%. Warna merah perlahan akan menjadi merah muda jika prosentase pengamatan yang tuntas Wajardikdas mendekati 50%.



Gambar 2: Topology pohon klasifikasi maksimal

Pohon klasifikasi maksimal memiliki biaya kesalahan relatif (*relative cost*) sebesar 1.417±0.348, biaya pengganti relatif (*resubtitution relative cost*) sebesar 0.248 dan kompleksitas parameter (*complexity parameter*) sebesar 0.000. Deretan

pohon (*tree squence*) dari pohon klasifikasi maksimal ini dapat dilihat pada Tabel 4.41. *Resubtitution relative cost* akan menurun seiring bertambahnya simpul. *Cost* ini seperti R-square pada regresi, dimana nilainya akan bertambah jika ada variabel yang ditambahkan kedalam model.[10] Pohon klasifikasi maksimal memang menghasilkan *resubtitution relative cost* (biaya kesalahan klasifikasi yang harus ditanggung karena telah menggunakan learning sample sebagai sample test) paling kecil, sehingga pohon ini cenderung dipilih untuk menduga nilai respon, tetapi struktur data yang digambarkan pada pohon klasifikasi maksimal cenderung lebih kompleks, sehingga pohon ini tidak dapat menduga data testing dengan baik seperti dengan menggunakan data learning. Untuk lebih jelasnya dapat dilihat pada kedua tabel berikut ini.

Tabel 3: Misklasifikasi Untuk Data Learning Pada Pohon Maksimal

CLASS	N CASES	N MISCLASSED	PCT ERROR	COST
Tuntas Wajardikdas	60	4	6.67	0.07
Tidak Tuntas Wajardikdas	77	14	18.18	0.18

Untuk data learning, hasil klasifikasi menunjukkan bahwa dari 60 pengamatan dengan label kelas tuntas Wajardikdas ada 4 pengamatan yang salah klasifikasi sehingga biaya kesalahan klasifikasi yang harus ditanggung adalah 0.07. Sedangkan untuk label kelas tidak tuntas Wajardikdas kesalahan klasifikasinya ada 14 pengamatan dari 77 pengamatan yang ada sehingga biaya kesalahan klasifikasi yang harus ditanggung adalah 0.18.

Tabel 4:Misklasifikasi Untuk Data Testing Pada Pohon Maksimal

CLASS	N CASES	N MISCLASSED	PCT ERROR	COST
Tuntas Wajardikdas	3	2	66.67	0.67
Tidak Tuntas Wajardikdas	4	3	75.00	0.75

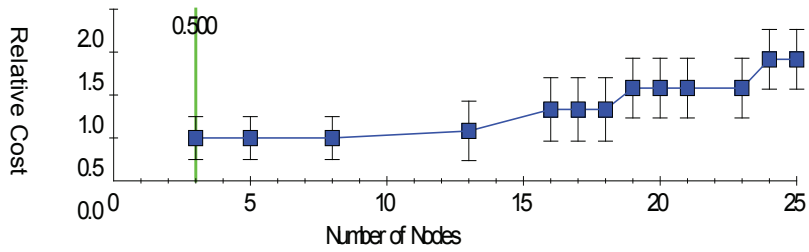
[10] Y, Yohannes, dan Huddinot, *Classification and Regression Trees: AnIntroduction,* Technical Guide, (Washington D.C.International Food Policy Research Institute, 1999), 65

Untuk data testing, hasil klasifikasi menunjukkan bahwa untuk label kelas tidak tuntas paripurna kesalahan klasifikasinya ada 2 pengamatan dari 3 pengamatan yang ada sehingga biaya kesalahan klasifikasi yang harus ditanggung adalah 0.67. Sedangkan dari 4 pengamatan dengan label kelas tuntas paripurna ada 3 pengamatan yang salah klasifikasi sehingga biaya kesalahan klasifikasi yang harus ditanggung adalah 0.75.

Dengan ukuran pohon yang sangat besar ini struktur data yang digambarkan cenderung kompleks, sehingga perlu dipilih pohon optimal yang berukuran sederhana tetapi memberikan biaya kesalahan relatif (relative cost) yang cukup kecil. Oleh karena itu tahap selanjutnya adalah pemangkasan (prunning) pohon maksimal sehingga diperoleh pohon optimal.

Pohon Klasifikasi Optimal

Berdasarkan Tabel 5 dan Gambar 3, didapatkan pohon optimal dengan jumlah simpul terminal kecil tetapi mempunyai relative cost (biaya kesalahan klasifikasi yang harus ditanggung karena telah menerapkan penduga ketika memangkas pohon) terkecil. Seiring dengan naiknya jumlah simpul, relative cost semakin menurun sampai menjangkau nilai minimum dan akhirnya naik lagi. Pohon dengan nilai relative cost minimum ini adalah pohon optimal.



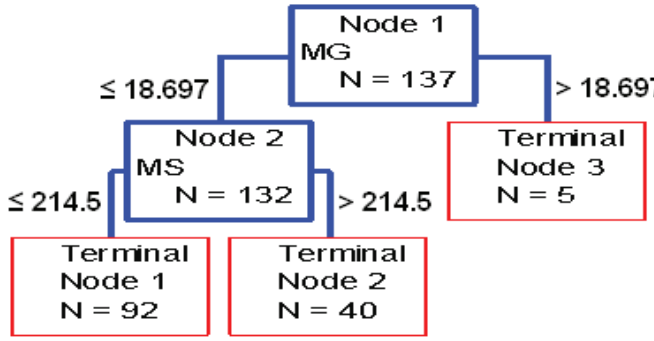
Gambar 3: Plot Relative Cost Dengan Jumlah Simpul Terminal

Pohon klasifikasi optimal terbentuk dengan jumlah simpul terminal sebanyak 3, relative cost 0.500±0.250, resubtitution relative cost sebesar 0.808 dan complexity parameter (kompleksitas parameter yang digunakan oleh CART pada saat pemangkasan pohon) sebesar 0.023 (dapat dilihat pada Tabel 5).

Tabel 5: Tree Squence

Terminal		Test Set	Resubstitution Complexity	
Tree	Nodes	Relative Cost	Relative Cost	Parameter
1	25	1.417 +/- 0.348	0.248	0.000
5	20	1.083 +/- 0.348	0.284	0.005
6	19	1.083 +/- 0.348	0.295	0.006
7	18	0.833 +/- 0.370	0.308	0.007
8	17	0.833 +/- 0.370	0.324	0.008
9	16	0.833 +/- 0.370	0.348	0.012
10	13	0.583 +/- 0.348	0.426	0.013
11	8	0.500 +/- 0.250	0.584	0.016
12	5	0.500 +/- 0.250	0.715	0.022
13**	3	0.500 +/- 0.250	0.808	0.023
14	1	1.000 +/- 0.000	1.000	0.048

Berdasarkan gambar 4, dapat diketahui bahwa pohon klasifikasi optimal yang terbentuk mempunyai kedalaman 3. Simpul terminal yang dihasilkan oleh pohon klasifikasi optimal adalah 3 simpul. Disini dapat dilihat bahwa variabel yang menjadi pemilah utama (primary spliter) adalah MG yaitu rasio murid/guru. Hal ini berarti variabel rasio murid/guru tersebut menjadi faktor utama yang mempengaruhi adanya perbedaan kondisi ketuntasan yang dicapai oleh desa/kelurahan di Kabupaten Gresik pada program Wajardikdas 9 tahun.



Gambar 4: Pohon Klasifikasi Optimal

Simpul utama terdiri dari 137 pengamatan, yang kemudian disebut sebagai simpul 1. Pada simpul ini ada sebanyak 136 pemilahan yang mungkin terjadi. Rasio murid/guru sebesar 18.697 merupakan nilai tengah dari 2 pengamatan yang terpilih karena dapat memberikan nilai penurunan keheterogenan tertinggi. Sebanyak 132 desa/kelurahan yang rasio murid/gurunya 18.697≤ dipilah ke simpul kiri, simpul kiri ini nantinya akan menjadi simpul 2 sedangkan 5 desa/kelurahan yang rasio murid/gurunya ¯18.697 dipilah ke simpul kanan yaitu simpul terminal 3 dan kemudian 5 desa/kelurahan tersebut diberi label kelas 1 atau dengan kata lain diprediksikan masuk dalam kategori tidak tuntas Wajardikdas. Tingkat kesalahan pengklasifikasian pada simpul terminal 3 adalah 0%, artinya semua desa/kelurahan dapat tepat diklasifikasikan kedalam kategori tidak tuntas Wajardikdas.

Simpul 2 terdiri dari 132 pengamatan sehingga ada sebanyak 131 pemilahan yang mungkin terjadi. Dari 131 nilai tengah yang ada, rasio murid/sekolah sebesar 214.5 terpilih karena dapat memberikan nilai penurunan keheterogenan tertinggi. Simpul 2 dipilah berdasarkan variabel MS yaitu rasio murid/sekolah dengan banyak pengamatan adalah 132, desa/kelurahan rasio murid/sekolahnya 214.5≤ dipilah ke simpul kiri yang kemudian menjadi simpul terminal 1 dengan banyak amatan adalah 92. Desa/kelurahan yang masuk kedalam simpul terminal 1 ini kemudian diberi label kelas 1 atau dengan kata lain diprediksikan masuk dalam kategori tidak tuntas Wajardikdas. Tingkat kesalahan pengklasifikasian pada simpul terminal 1 adalah 39.13%, artinya terdapat 36 desa/kelurahan yang diklasifikasikan dalam kategori tidak tuntas Wajardikdas padahal seharusnya desa/kelurahan tersebut masuk kedalam kategori tuntas Wajardikdas. Sedangkan rasio murid/sekolah ¯214.5 dengan banyak amatan 40 akan dipilah menjadi simpul kanan yaitu simpul terminal 2 dan kemudian sebanyak 40 desa/kelurahan ini diberi label kelas 2 atau dengan kata lain diprediksikan masuk dalam kategori tuntas Wajardikdas. Tingkat kesalahan pengklasifikasian pada simpul terminal 2 adalah 40%, artinya terdapat 16 desa/kelurahan yang diklasifikasikan dalam kategori tuntas Wajardikdas padahal seharusnya desa/kelurahan tersebut masuk kedalam kategori tidak tuntas Wajardikdas.

Berdasarkan pohon klasifikasi optimal diatas (Gambar 4), didapatkan hasil bahwa variabel rasio murid/guru rasio dan rasio murid/sekolah adalah faktor-faktor yang mempengaruhi adanya perbedaan kondisi ketuntasan yang dicapai oleh desa/kelurahan di Kabupaten Gresik pada program Wajardikdas 9 tahun. Berdasarkan Gambar 4 juga dapat dijelaskan bahwa desa/kelurahan yang mempunyai rasio murid/guru ≤18.697 dan rasio murid/sekolah ≤ 214.5 akan diprediksikan sebagai desa/kelurahan yang tidak tuntas Wajardikdas, sebaliknya desa/kelurahan yang mempunyai rasio murid/guru ≤ 18.697 dan rasio murid/sekolah ¯214.5 akan diprediksikan sebagai desa/kelurahan yang tuntas

Wajardikdas. Sedangkan desa/kelurahan yang rasio murid/gurunya 18.697 akan diprediksikan sebagai desa/kelurahan yang tidak tuntas paripurna.

Setelah didapatkan pohon klasifikasi optimal, langkah selanjutnya adalah menelusuri pohon klasifikasi menggunakan data respon (learning dan testing). Hasil pengklasifikasian dengan data learning dapat dlihat pada Tabel 6.

Tabel 6: Prediksi Sukses Untuk Data Learning Pada Pohon Optimal

ACTUAL CLASS	TOTAL CASES	PERCENT CORRECT	PREDICTED CLASS	
			Tidak Tuntas Wajardikdas	Tuntas Wajardikdas
Tidak Tuntas Wajardikdas	77	79.22	61	16
Tuntas Wajardikdas	60	40	36	24

Berdasarkan Tabel 6, dapat diketahui bahwa data respon dengan kategori tidak tuntas Wajardikdas yang tepat di klasifikasikan ke kategori tidak tuntas Wajardikdas ada 61 pengamatan (79.221%). Sedangkan data kategori tuntas Wajardikdas yang tepat diklasifikasikan ada 24 pengamatan (40%). Sehingga keseluruhan data learning yang tepat diklasifikasikan ada 108 pengamatan. dan yang tidak tepat diklasifikasikan ada sebanyak 52 pengamatan.

Setelah didapatkan ketepatan klasifikasi dengan menggunakan data learning, maka langkah selanjutnya adalah menelusuri pohon optimal yang terbentuk dengan menggunakan data testing. Penelusuran pohon klasifikasi ini untuk menguji ketepatan model pohon klasifikasi yang terbentuk dari data learning. Hasil ketepatan model pohon klasifikasi dengan menggunakan data testing ditunjukkan pada Gambar 7.

Tabel 7: Prediksi Sukses Untuk Data Testing Pada Pohon Optimal

ACTUAL CLASS	TOTAL CASES	PERCENT CORRECT	PREDICTED CLASS	
			Tidak Tuntas Wajardikdas	Tuntas Wajardikdas
Tidak Tuntas Wajardikdas	3	100	3	0
Tuntas Wajardikdas	4	50	2	2

Hasil pengklasifikasian pada Tabel 7 diatas menunjukkan bahwa untuk kategori tidak tuntas Wajardikdas yang tepat dalam pengklasifikasiannya ada 3 pengamatan (100%), sedangkan pada kategori tuntas Wajardikdas hanya ada 2 pengamatan (50%) yang tepat diklasifikasikan ke kategori tuntas Wajardikdas. Jadi keseluruhan data testing yang tepat diklasifikasikan ada 5 pengamatan dan yang tidak tepat diklasifikasikan ada sebanyak 2 pengamatan.

KESIMPULAN

Berdasarkan hasil penelitian dan pembahasan maka didapatkan kesimpulan bahwa pada kasus program Wajardikdas 9 tahun di Kabupaten Gresik dengan menggunakan data learning dan data testing, ketepatan klasifikasi yang dihasilkan oleh metode regresi logistik terlihat lebih konstan daripada klasifikasi pohon, karena ketepatan klasifikasi yang dihasilkan oleh metode klasifikasi pohon cenderung tergantung pada kondisi pembagian data. Semakin tinggi proporsi data learning, maka ketepatan klasifikasi yang dihasilkan oleh data testing semakin tinggi. Walaupun demikian, ketepatan klasifikasi yang dihasilkan oleh metode klasifikasi pohon secara keseluruhan (dengan menggunakan 9 kondisi data) dapat dikatakan lebih tinggi dari metode regresi logistik. Metode yang lebih tepat dipergunakan untuk pengklasifikasian desa/kelurahan di kabupaten Gresik pada program Wajardikdas 9 tahun adalah klasifikasi pohon, karena ketepatan klasifikasi data testing yang dihasilkan lebih tinggi daripada regresi logistik. Dengan metode klasifikasi pohon ini didapatkan hasil bahwa, variabel yang berpengaruh terhadap kondisi kentuntasan Wajardikdas 9 tahun adalah variabel rasio murid/guru dan rasio murid/sekolah.

DAFTAR PUSTAKA

Agresti, A., Categorical Data Analysis, John Wiley & Sons, Inc., New York, 1990.

Breiman, L,. et.al., *Classification and Regression Trees*, New York- London: Chapman and Hall, 1984.

Camdeviren, H.A. et.al. "Comparison of Logistic Regression Model and Classification Tree: An Aplication to Postpartum Depression Data", Expert System with Application. Vol. 32, 2007.

Feldesman, M.C. "Classification Trees as An Alternative to Linier Discriminant Analysis", *American Journal of Physical Anthropology*, Vol. 119, 2002.

Hosmer, D. W. dan Lemeshow, S., *Applied Logistic Regression*, New York John Wiley & Sons, Inc., 2000.

Kurt, I., et. al., "Comparing Performances of Logistic Regression, Classification and Regression Tree, and Neural Networks forPredicting Coronary Artery Disease", Expert Systems with Application, Vol. 34, 2008.

Lewis, R. J. "An Introduction to Classification and Regression Tree (CART) Analysis", *Annual Meeting of The Society for Academic Emergency Medicine in San Fransisco*, California: Harbor-UCLA Medical Center, 2000.

Phelps, M.C., dan Merkle, E.C. "Classification and Regression Trees as alternatives to Regression", *Proceeding of The 4th Annual GRASP Symposium*, Wichita:Wichita State University, 2008.

Santoso, I.S., *Pembinaan Watak Tugas Utama Pendidikan*, Jakarta: U-I Press, 1981.

Steiberg, D., dan Philps, C., *CART-Classification and Regression Trees*, San Diego: Salford System, 2005.

Sukriswandari, N., *Upaya Direktorat Pembinaan SMP dalam Penuntasan Wajar 9 Tahun*, Jakarta: Direktorat Pembinaan Sekolah Menengah Pertama, 2008.

Suryadi, A., dan Untung, "Gerakan Pemberantasan Buta Aksara Intensif", dalam *Aksara*, Jakarta: Direktorat Pendidikan Masyarakat, 2005.

Utomo, I., *Laporan Keterangan Pertanggungjawaban Akhir Masa Jabatan Gubernur 2003-2008*, Surabaya: Kantor Pemerintahan Daerah Tingkat I Propinsi Jawa Timur, 2008.

Wibowo, W., "Perbandingan Hasil Ketepatan Klasifikasi Analisis Diskriminan dan Regresi Logistik pada Pengklasifikasian Data Respon Biner", Surabaya: ITS, 2002.

Yohannes, Y dan Huddinot, J., *Classification and Regression Trees: An Introduction*, Washington D.C: Technical Guide, International Food Policy Research Intitut, 1999