TESIS – SS14 2501

SPLITTING RULE DAN PENERAPAN *BAGGING* PADA POHON KLASIFIKASI

Studi Kasus : Pekerja Anak di Provinsi Sulawesi Tengah

MOHAMMAD FAJRI
NRP. 1313 201 027

DOSEN PEMBIMBING
Dr. MUHAMMAD MASHURI, M.T

PROGRAM MAGISTER
JURUSAN STATISTIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
INSTITUT TEKNOLOGI SEPULUH NOPEMBER
SURABAYA
2015

THESIS – SS14 2501

SPLITTING RULE AND BAGGING APPLIED ON CLASSIFICATION TREE

Case Study : Child Labor in Central Sulawesi Province

MOHAMMAD FAJRI
NRP. 1313 201 027

SUPERVISOR
Dr. MUHAMMAD
MASHURI, M.T

PROGRAM OF MAGISTER
DEPARTMENT OF STATISTICS
FACULTY OF MATHEMATICS AND NATURAL SCIENCE
SEPULUH NOPEMBER INSTITUTE OF TECHNOLOGY
SURABAYA
2015

SPLITTING RULE DAN PENERAPAN _BAGGING_ PADA POHON KLASIFIKASI
(Studi Kasus : Pekerja Anak di Provinsi Sulawesi Tengah)

Tesis ini disusun untuk memenuhi salah satu syarat memperoleh gelar
Magister Sains (M.Si)
di
Institut Teknologi Sepuluh Nopember

Oleh :

MOHAMMAD FAJRI
NRP 1313201027

Tanggal Ujian : 11 Mei 2015
Periode Wisuda : September 2015

Disetujui Oleh :

1. **Dr. Muhammad Mashuri, M.T** **(Pembimbing)**
   NIP. 19620408 198701 1 001

2. **Dr. Bambang Widjanarko Otok, M.Si** **(Penguji)**
   NIP. 19681124 1994121 001

3. **Dr. Wahyu Wibowo, M.Si** **(Penguji)**
   NIP. 19740328 199802 1 001

Direktur Program Pascasarjana

Prof. Dr. Ir. Adi Soeprijanto, M.T
NIP. 19640405 199002 1 001

SPLITTING RULE DAN PENERAPAN _BAGGING_ PADA POHON KLASIFIKASI
Studi Kasus: Pekerja Anak di Provinsi Sulawesi Tengah

Nama Mahasiswa : Mohammad Fajri
NRP : 1313 201 027
Jurusan : Statistika FMIPA ITS
Dosen Pembimbing : Dr. Muhammad Mashuri, M.T

ABSTRAK

Pengklasifikasian merupakan metode statistika yang digunakan untuk mengelompokkan suatu pengamatan. Dalam statistika terdapat beberapa metode klasifikasi, salah satunya adalah pohon klasifikasi yang merupakan metode klasifikasi yang menghasilkan model pohon. Pembentukan pohon klasifikasi ditentukan oleh proses *splitting rule* yang didasarkan pada ukuran keheterogenan. Ukuran yang digunakan dalam tesis ini adalah kriteria indeks *Gini* dan indeks *Twoing*. Pohon klasifikasi memiliki beberapa keunggulan terkait model dan hasil klasifikasinya, tetapi juga memiliki kekurangan pada stabilitas model dan keakuratan prediksi. Guna mengatasi kelemahan tersebut, *bagging* (*bootstrap aggregating*) diterapkan pada pohon klasifikasi untuk meningkatkan stabilitas dan keakuratan prediksi. Metode ini diaplikasikan pada data pekerja anak di provinsi Sulawesi Tengah. Indeks *Gini* dan indeks *Twoing* memiliki konsep dan bentuk yang berbeda namun memiliki keuntungan masing-masing dalam penggunaannya. Pada kasus pekerja anak di Sulawesi Tengah, kedua indeks ini menghasilkan pohon klasifikasi optimal yang identik dengan enam variabel pembentuk, yaitu partisipasi sekolah anak, umur anak, jenis kelamin anak, umur kepala rumah tangga, jumlah anggota rumah tangga, pendapatan perkapita dan tingkat pendidikan kepala rumah tangga. Penerapan teknik *bagging* pada pohon klasifikasi dalam kasus ini mampu meningkatkan ketepatan klasifikasinya.

Kata Kunci : *bagging*, klasifikasi, pekerja anak, pohon klasifikasi, *splitting rule*.

(*Halaman Ini Sengaja Dikosongkan*)

SPLITTING RULE AND BAGGING APPLIED ON CLASSIFICATION TREE

Case Study: Child Labor in Central Sulawesi Province

Name	: Mohammad Fajri
NRP	: 1313 201 027
Department	: Statistics FMIPA ITS
Supervisor	: Dr. Muhammad Mashuri, M.T

ABSTRACT

Classification is statistical method that used to classify an observation. In statistics, there are several classification methods, one of the methods is classification tree that resulted classification trees model. Classification tree's formation determined by splitting rule process based on impurity measure. In this thesis, the used measure are Gini index and Twoing index criteria. Classification tree has some advantages related with the model and the classification result, although it has a weakness on stability model that resulted and classification accuracy. To solve this weakness, bagging (bootstrap aggregating) technique was applied on classification tree method to increase the stability and classification accuracy. This method applied on child labor's case in Central Sulawesi province. Gini index and Twoing index have different concept and form, however both index have each advantage on its use. On child labor's case in Central Sulawesi, both index resulted identic optimal classification tree with six formed variables, child's school participation, child's age, child's sex, number of household member, income per capita and head of household's education level. Bagging technique on classification tree resulted higher classification accuracy.

Keywords: bagging, child labor, classification, classification tree, splitting rule.

(*Halaman Ini Sengaja Dikosongkan*)

KATA PENGANTAR

Alhamdulillah penulis panjatkan kehadirat Allah SWT yang telah memberikan rahmat, taufiq, nikmat serta hidayah-Nya kepada penulis sehingga tesis ini dapat diselesaikan tepat pada waktunya. Tesis ini berjudul "*SPLITTING RULE* DAN PENERAPAN *BAGGING* PADA POHON KLASIFIKASI (Studi Kasus : Pekerja Anak di Provinsi Sulawesi Tengah)". Tesis ini disusun untuk memenuhi salah satu syarat dalam menyelesaikan pendidikan program Pascasarjana di Jurusan Statistika FMIPA ITS.

Penulis menyadari bahwa tesis ini dapat terselesaikan tidak lepas dari bantuan berbagai pihak. Oleh karena itu, penulis mengucapkan banyak terima kasih kepada :

1. Kedua orang tua tersayang, Bapak Anwar AR dan Ibu Haerani yang telah memberikan doa, dukungan dan kasih sayang yang tak terkira. Kepada adik-adik tersayang, Dian Muzizat, Dita Muthia Rahma, Dini Noviyandari dan Imam Rizaldi yang memberikan kasih sayang dan doa kepada penulis. Kepada Nindy Gustianti yang telah memberikan doa dan dukungan kepada penulis.
2. Bapak Dr. Muhammad Mashuri, M.T, selaku dosen pembimbing dan ketua Jurusan Statistika FMIPA ITS yang telah dengan sabar memberikan banyak bimbingan, nasehat, motivasi serta banyak ilmu pengetahuan demi terselesaikan tesis ini.
3. Bapak Dr. Bambang Widjanarko Otok, M.Si dan Bapak Dr. Wahyu Wibowo, M.Si selaku dosen penguji yang telah memberikan banyak masukan dan saran demi kesempurnaan tesis ini.
4. Bapak Dr. Suhartono, M.Sc, selaku dosen wali dan ketua Prodi Pascasarjana Statistika FMIPA ITS yang telah memberikan nasehat kepada penulis, baik akademik maupun nonakademik.
5. Bapak dan Ibu dosen pengajar Jurusan Statistika FMIPA ITS yang telah mengajarkan ilmu yang bermanfaat kepada penulis.

6. Bapak dan Ibu staf dan karyawan Jurusan Statistika FMIPA ITS yang telah membantu penulis selama masa perkuliahan.
7. Teman-teman seperjuangan dan seperantauan di Pascasarjana Statistika di kampus perjuangan khususnya angkatan 2013.
8. Semua pihak yang telah membantu penulis yang tidak dapat disebutkan satu-persatu, terimah kasih atas doa dan dukungannya.

Surabaya, Mei 2015

Penulis

DAFTAR ISI

DAFTAR TABEL

(*Halaman Ini Sengaja Dikosongkan*)

DAFTAR GAMBAR

(*Halaman Ini Sengaja Dikosongkan)*

DAFTAR LAMPIRAN

(*Halaman Ini Sengaja Dikosongkan*)

BAB 1
PENDAHULUAN

1.1. Latar Belakang

Pengklasifikasian merupakan salah satu metode statistika yang digunakan untuk mengelompokkan atau mengklasifikasi suatu pengamatan. Masalah klasifikasi amat sering dijumpai dalam kehidupan sehari-hari, baik itu pengklasifikasian pada data bidang pendidikan, sosial, pemerintahan, ekonomi, maupun pada bidang yang lainnya. Masalah klasifikasi ini muncul ketika terdapat sebuah pengamatan yang bersifat heterogen. Secara garis besar, pengklasifikasian dapat diartikan sebagai metode pengelompokkan atau pengalokasian suatu objek atau observasi ke dalam suatu grup tertentu (Johnson, R. and Wichern, D., 2007).

Dalam statistika, ada beberapa metode klasifikasi yang familiar, seperti analisis diskriminan dan regresi logistik yang dalam beberapa literatur klasifikasi sering disebut sebagai model klasik. Kedua metode ini mempunyai beberapa asumsi yang harus dipenuhi berkaitan dengan skala pengukuran prediktor, keterkaitan antara prediktor, maupun asumsi-asumsi klasik lainnya. Misalnya pada analisis diskriminan memerlukan pemenuhan asumsi normalitas, homogenitas dan skala prediktor yang metric (Johnson and Wichern, 2007). Sedangkan pada regresi logistik memerlukan asumsi tidak adanya multikoliniearitas (Schaefer, R.L., 1986). Asumsi-asumsi ini nyatanya merupakan suatu keterbatasan, karena banyak ditemui data yang tidak dapat memenuhi asumsi-asumsi tersebut.

Seiring dengan perkembangan ilmu pengetahuan, metode klasifikasi juga mengalami perkembangan untuk mengatasi keterbatasan yang terdapat pada metode klasik. Metode yang dikembangkan antara lain adalah *Classification and Regression Trees* (*CART*), *Multivariate Adaptive Regression Spline* (*MARS*), *Neural Network* (*NN*), *Genetic Algorithm* (*GA*), dan metode-metode lainnya yang umumnya merupakan metode klasifikasi yang lebih fleksibel karena sebagian besar tidak mengharuskan pemenuhan asumsi tertentu, namun tetap menghasilkan hasil klasifikasi yang optimal. Di antara beberapa metode klasifikasi tersebut,

CART merupakan metode yang unik, karena selain menghasilkan model klasifikasi visual, metode ini juga memiliki perhitungan komputasi yang lebih efisien dan lebih mudah dalam penginterpretasian modelnya (Breiman, L., Friedman, J., Olshen, R. and Stone, C., 1984).

CART diperkenalkan oleh Breiman *et al* (1984) merupakan sebuah metodologi inovatif untuk analisis data yang besar melalui prosedur pemilihan biner yang digunakan untuk menggambarkan hubungan antara variabel respon dengan variabel prediktor. Model yang dihasilkan oleh *CART* berupa model pohon. Model pohon tersebut akan menjadi pohon regresi jika variabel responnya berupa data kontinu. Sedangkan model pohon tersebut akan menjadi pohon klasifikasi jika variabel respon merupakan data kategori.

Dalam pohon klasifikasi, pembentukannya sangat ditentukan oleh *splitting rule*, yaitu proses pemilahan pemilah yang didasarkan pada ukuran keheterogenan. Ukuran keheterogenan yang digunakan adalah indeks *Gini* dan indeks *Twoing*. Indeks *Gini* dan indeks *Twoing* memiliki ukuran keheterogenan yang berbeda, sehingga dalam penerapannya, kedua metode ini memungkinkan akan menghasilkan model klasifikasi yang berbeda.

Pohon klasifikasi memiliki beberapa kelebihan dan kekurangan. Kelebihannya adalah tidak terikat oleh berbagai asumsi yang harus dipenuhi oleh variabel prediktor, memudahkan dalam eksplorasi dan pengambilan keputusan karena hasilnya merupakan model visual, dapat digunakan pada banyak variabel, hasil klasifikasi akhir sederhana, efisien dan mudah diinterpretasikan. Sedangkan kekurangannya yaitu rentan menghasilkan model yang tidak stabil, yaitu rentan menghasilkan model yang berubah-ubah sesuai dengan penentuan jumlah data *learning* dan *testing*, serta proses pemilihan yang kurang tepat jika digunakan pada struktur data yang kompleks (Lewis, R.J., 2000).

Dalam upaya mengatasi kekurangan tersebut, Breiman (1996) mengembangkan suatu metode *bagging* (*Bootstrap Aggregating*) sebagai tehnik untuk memperbaiki keakurasian prediksi klasifikasi dengan cara mereduksi variansi dari suatu prediktor. Ide sederhana dari *bagging* adalah menggunakan *bootstrap resampling* untuk membangkitkan prediktor dengan banyak versi, dimana ketika dikombinasikan seharusnya hasilnya lebih baik dibandingkan

dengan prediktor tunggal yang dibangun untuk menyelesaikan masalah yang sama.

Beberapa penelitian tentang penerapan *bagging* pada pohon klasifikasi antara lain Sumarmi (2009) mengklasifikasi anak putus sekolah di provinsi Jambi dan Prakosa (2011) mengklasifikasi kesejahteraan rumah tangga di provinsi Jawa Timur. Hasil dari kedua penelitian tersebut menghasilkan penerapan *bagging* pada pohon klasifikasi dapat meningkatkan ketepatan klasifikasi dibandingkan dengan menggunakan metode pohon klasifikasi biasa.

Metode klasifikasi ini juga dapat digunakan untuk kasus pengklasifikasian di berbagai sektor kehidupan, tak terkecuali sektor ketenagakerjaan. Tenaga kerja adalah setiap orang yang mampu melakukan pekerjaan guna menghasilkan barang atau jasa untuk memenuhi kebutuhan sendiri maupun masyarakat. Menurut Badan Pusat Statistik (BPS), orang yang dikategorikan sebagai tenaga kerja adalah orang dengan usia produktif (15-64 tahun). Tenaga kerja di Indonesia mempunyai peranan dan kedudukan yang sangat penting sebagai pelaku dan tujuan pembangunan nasional yang dilaksanakan dalam rangka pembangunan manusia Indonesia seutuhnya dan pembangunan masyarakat Indonesia keseluruhan. Sesuai dengan kedudukan dan peranan tersebut, diperlukan pembangunan ketenagakerjaan untuk meningkatkan kualitas tenaga kerja di Indonesia.

Namun di tengah kondisi ekonomi yang tidak stabil dan terkesan carut marut, muncul banyak masalah di bidang ketenagakerjaan. Mulai dari upah kerja yang sering menjadi masalah tarik ulur antara pemilik usaha dengan pekerja, kualitas sumber daya manusia, hingga masalah yang menjadi sorotan bukan hanya di Indonesia, tapi juga menjadi masalah internasional, yaitu permasalahan mengenai pekerja anak. Pekerja anak merupakan anak usia sekolah yang masuk dalam kegiatan ekonomi. Manik (2006) mengartikan pekerja anak sebagai anak yang harus melakukan pekerjaan yang menghalangi mereka bersekolah dan membahayakan kesehatan, fisik dan mentalnya.

Kasus pekerja anak di Indonesia pada awalnya banyak berkaitan dengan tradisi atau budaya membantu orang tua. Secara garis besar, tradisi tersebut banyak dianut oleh masyarakat yang beranggapan bahwa memberi pekerjaan pada

anak merupakan bagian dari proses belajar untuk menghargai kerja dan bertanggung jawab. Seiring berjalannya waktu, fenomena pekerja anak berkaitan erat dengan ekonomi keluarga dan kesempatan untuk mengenyam pendidikan. Biasanya pekerja anak banyak ditemui di sektor-sektor usaha yang tidak memerlukan keahlian atau keterampilan khusus.

Menurut laporan ILO, pada tahun 2012 jumlah pekerja anak di seluruh dunia menurun menjadi 168 juta dari 246 juta pada tahun 2000. Sedangkan di Indonesia, menurut data yang dilansir oleh Komisi Nasional Perlindungan Anak, pada tahun 2013 jumlah pekerja anak mencapai 4,7 juta anak. Hal ini merupakan permasalahan yang cukup serius karena idealnya anak pada usia tersebut bisa mendapatkan hak-haknya untuk memperoleh pendidikan yang memadai. Masalah pekerja anak ini merupakan masalah yang sangat kompleks karena melibatkan banyak pihak seperti Kemendikbud, Komnas Perlindungan Anak, Kemenakertrans, dan masih banyak lagi pihak lain yang terkait. Pada tahun 2002, Kemenakertrans menyampaikan bahwa pendidikan merupakan salah satu intervensi yang sangat penting untuk mengurangi pekerja anak selain peningkatan kesejahteraan keluarga secara umum.

Nachrowi (1997) menyebutkan bahwa faktor yang mempengaruhi munculnya pekerja anak meliputi sisi penawaran (*supply*) dan sisi permintaan (*demand*). Sekalipun masyarakat menyediakan tenaga kerja anak, tetapi jika tidak ada perusahaan yang mempekerjakannya, sudah pasti tidak akan muncul pekerja anak. Begitu pula sebaliknya, jika permintaan terhadap pekerja anak tinggi, tetapi masyarakat tidak menyediakan maka pekerja anak tidak akan muncul. Dari sisi permintaan, faktor yang menyebabkan perusahaan mempekerjakan anak adalah karena adanya kebutuhan dari perusahaan, disamping mempertimbangkan aspek lain yaitu pekerja anak mudah diatur, tidak banyak menuntut, mau dibayar murah dan sebagainya. Dari sisi penawaran, faktor yang menyebabkan anak untuk masuk ke dalam pasar kerja mencakup kemiskinan keluarga, putus sekolah, persepsi keluarga, dan sebagainya.

Sekalipun kemiskinan merupakan pendorong utama anak terjun ke dunia kerja, tetapi kenyataan menunjukkan bahwa tidak semua orang yang hidup di bawah garis kemiskinan membiarkan anaknya terjun ke dunia kerja. Artinya ada

faktor-faktor lain yang mempengaruhinya, baik faktor budaya, sosial, demografi maupun psikologi. Irwanto (1995) menyebutkan bahwa kemiskinan bukan satu-satunya faktor penyebab munculnya pekerja anak. Sedangkan Melkas, dkk (1996) menyatakan bahwa kekuatan ekonomi yang mendorong anak ke dalam pekerjaan di lingkungan yang membahayakan, namun adat dan pola sosial yang berakar juga mempunyai peranan yang penting.

Pekerja anak ironisnya dapat ditemui di seluruh Indonesia, tak terkecuali di Sulawesi Tengah. Sulawesi Tengah sebagai daerah yang berkembang, juga memiliki sektor ekonomi yang terus menggeliat di berbagai bidang dimana pertumbuhan ekonomi daerah ini sebesar 10,33% pada tahun 2013 dengan sumber utama adalah pertanian, pertambangan dan penggalian, jasa, konstruksi, perdagangan serta hotel dan restauran. Namun ironisnya, dari data Kemenakertrans, beberapa sumber utama perekonomian tersebut yaitu pertanian, pertambangan dan penggalian, jasa serta manufaktur merupakan tempat-tempat dimana pekerja anak ditemui di Sulawesi Tengah. Bahkan anak-anak juga bekerja sebagai pemecah batu yang notabene cukup berbahaya bagi anak, baik secara psikis maupun fisik.

Pernyataan ini juga diperkuat data yang dilansir oleh Komnas Perlindungan Anak pada tahun 2013, yang menyatakan bahwa provinsi-provinsi yang memiliki presentase pekerja anak paling tinggi sebagian besar berada di wilayah timur Indonesia, seperti Papua dan Sulawesi. Di Sulawesi Tengah sendiri jumlah pekerja anak ini mencapai 10% dari total pekerja. Hal ini tentunya berkaitan dengan berbagai macam faktor yang melatar belakangi munculnya kasus pekerja anak.

Jadi pada kasus ini, pengklasifikasian dengan menggunakan pohon klasifikasi akan dilakukan pada kasus pekerja anak di provinsi Sulawesi Tengah untuk mengetahui kecenderungan anak bekerja dan faktor yang mempengaruhi kasus pekerja anak tersebut serta akan menerapkan teknik bagging pada pohon klasifikasi. Konsep *splitting rule* pada pohon klasifikasi juga akan dibahas. Selain itu diharapkan nantinya model klasifikasi yang didapatkan akan bisa menjadi bahan masukan untuk memangkas jumlah pekerja anak di provinsi Sulawesi Tengah.

1.2. Rumusan Masalah

Dari latar belakang yang telah diuraikan sebelumnya, maka yang menjadi permasalahan dalam penelitian ini adalah sebagai berikut.

1. Bagaimana konsep *splitting rule* dalam pembentukan pohon klasifikasi.
2. Bagaimana perbandingan hasil pemodelan pohon klasifikasi menggunakan kriteria indeks *Gini* dan indeks *Twoing* pada kasus pekerja anak di provinsi Sulawesi Tengah.
3. Bagaimana penerapan *bagging* pada pohon klasifikasi dalam kasus pekerja anak di provinsi Sulawesi Tengah.

1.3. Tujuan Penelitian

Dari permasalahan yang ada, maka tujuan penelitian ini adalah sebagai berikut.

1. Membahas konsep *splitting rule* dalam pembentukan pohon klasifikasi.
2. Mendapatkan dan membandingkan model pohon klasifikasi dengan menggunakan kriteria indeks *Gini* dan indeks *Twoing* pada kasus pekerja anak di provinsi Sulawesi Tengah.
3. Mendapatkan pengaruh penerapan *bagging* pada pohon klasifikasi dalam kasus pekerja anak di provinsi Sulawesi Tengah.

1.4. Manfaat Penelitian

Manfaat yang diperoleh dari penelitian ini adalah sebagai berikut.

1. Bagi pihak-pihak yang terkait dengan pekerja anak, penelitian ini dapat digunakan sebagai bahan masukan pengambilan kebijakan untuk mengatasi kasus pekerja anak.
2. Bagi ruang keilmuan statistika, penelitian ini dapat meningkatkan wawasan keilmuan yang bersinggungan dengan pohon klasifikasi khususnya *splitting rule* dalam pembentukan pohon klasifikasinya dan penerapan *bagging* di dalamnya.
3. Bagi penulis, sebagai penerapan metode-metode statistik ke dalam masalah nyata.

1.5. Batasan Masalah

Batasan masalah dalam penelitian ini adalah sebagai berikut.

1. Data yang digunakan dalam penelitian ini adalah data dari Survey Sosial Ekonomi Nasional (Susenas) 2013.
2. Unit penelitian ini adalah rumah tangga di provinsi Sulawesi Tengah.

(*Halaman Ini Sengaja Dikosongkan*)

BAB 2
TINJAUAN PUSTAKA

2.1. Metode Klasifikasi

Metode klasifikasi merupakan bagian dari analisis statistika untuk mengelompokkan atau mengklasifikasi suatu pengamatan yang disusun secara sistematis. Masalah klasifikasi merupakan masalah yang sangat bersinggungan dengan kehidupan sehari-hari yang muncul ketika sejumlah ukuran yang terdiri dari satu atau beberapa kategori yang tidak dapat diidentifikasi secara langsung tetapi harus menggunakan suatu ukuran. Metode klasifikasi dapat dilakukan dengan pendekatan parametrik maupun pendekatan nonparametrik.

Dalam pendekatan parametrik, metode yang biasa digunakan adalah analisis diskriminan dan regresi logistik. Analisis diskriminan pertama kali dikembangkan pada tahun 1936 oleh Ronald A. Fisher. Dalam sebuah tulisannya, *"The Use of Multiple Measurements in Taxonomic Problems"*, Fisher menyatakan bahwa apabila dua atau lebih populasi diukur dalam beberapa karakter $X_1, X_2, \dots, X_p$, maka dapat dibangun fungsi linier tertentu dari pengukuran itu, dimana fungsi itu merupakan pembeda (pemisah) terbaik bagi populasi-populasi yang dipelajari. Fungsi linier yang dibangun inilah yang disebut sebagai fungsi diskriminan yang memaksimumkan jarak antar kelompok, sehingga memiliki kemampuan untuk membedakan antar kelompok, sebagai kriteria pengelompokkan. Oleh karena itu analisis diskriminan dapat dipergunakan sebagai metode pengklasifikasian.

Secara umum tujuan dari analisis diskriminan yaitu untuk mengetahui apakah ada perbedaan yang jelas antar grup pada variabel respon, untuk mengetahui variabel prediktor pada fungsi diskriminan yang membuat perbedaan pada variabel respon, untuk membuat fungsi diskriminan, dan melakukan klasifikasi terhadap objek. Dalam analisis diskriminan, terdapat empat langkah substansial, yaitu pengujian asumsi data berdistribusi multivariat normal, pengujian homogenitas matriks varian-kovarian, pembentukan fungsi diskriminan dan pengujian signifikansi fungsi diskriminan.

Senada dengan analisis diskriminan, regresi logistik merupakan salah satu metode klasik dalam pengklasifikasian. Regresi logistik merupakan pengembangan dari regresi linier, tapi dengan variabel respon yang bersifat kategorik. Secara umum, prosedur dalam regresi logistik sama dengan regresi linier. Regresi logistik akan menghasilkan model yang disebut model logit yang memuat model klasifikasi. Model logit atau persamaan regresi logistik yang dihasilkan merupakan model probabilitas. Karena variabel respon pada regresi logistik merupakan variabel kategorik, maka model yang didapatkan akan mengklasifikasikan variabel prediktor ke dalam variabel respon. Dalam regresi logistik terdapat pula nilai odd ratio yang bisa menggambarkan kecenderungan variabel prediktor terhadap variabel respon.

Kedua metode ini merupakan metode klasik dalam pengklasifikasian dan memiliki beberapa asumsi yang harus dipenuhi. Asumsi-asumsi tersebut berkaitan dengan skala pengukuran prediktor, keterkaitan antara prediktor, maupun asumsi-asumsi klasik lainnya. Sehingga kedua metode ini memiliki keterbatasan jika dihadapkan pada struktur data yang tidak dapat memenuhi asumsi-asumsi yang telah ditetapkan sebelumnya.

Dalam perkembangannya, muncul metode klasifikasi dengan pendekatan nonparametrik seperti *Classification and Regression Trees* (*CART*), *Multivariate Adaptive Regression Spline* (*MARS*), *Neural Network* (*NN*), *Genetic Algorithm* (*GA*), dan metode-metode lainnya yang lebih bebas dalam pemenuhan asumsi tertentu, namun tetap memiliki kemampuan klasifikasi yang optimal. Pohon klasifikasi dalam *CART* adalah salah satu metode klasifikasi yang memiliki beberapa keunggulan diantaranya memudahkan dalam eksplorasi dan pengambilan keputusan karena hasilnya merupakan model visual, dapat digunakan pada banyak variabel, hasil klasifikasi akhir sederhana, efisien dan mudah diinterpretasikan. Sedangkan kelemahan metode ini adalah pohon klasifikasi yang dihasilkan cenderung kurang stabil dan pemilahan data hanya didasarkan pada satu nilai variabel.

Untuk mengatasi kelemahan dalam metode tersebut, Breiman (1996) melakukan pendekatan *bagging* (*bootstrap aggregating*) untuk memperbaiki keakurasian prediksi klasifikasi dengan cara mereduksi variansi dari suatu

prediktor. Pemikiran dasar dari *bagging* adalah menggunakan *bootstrap resampling* untuk menghasilkan prediktor dengan banyak versi, dimana jika prediktor hasil *bootstrap* tersebut dikombinasikan seharusnya hasilnya lebih baik dibandingkan dengan prediktor tunggal yang dibangun untuk menyelesaikan masalah yang sama.

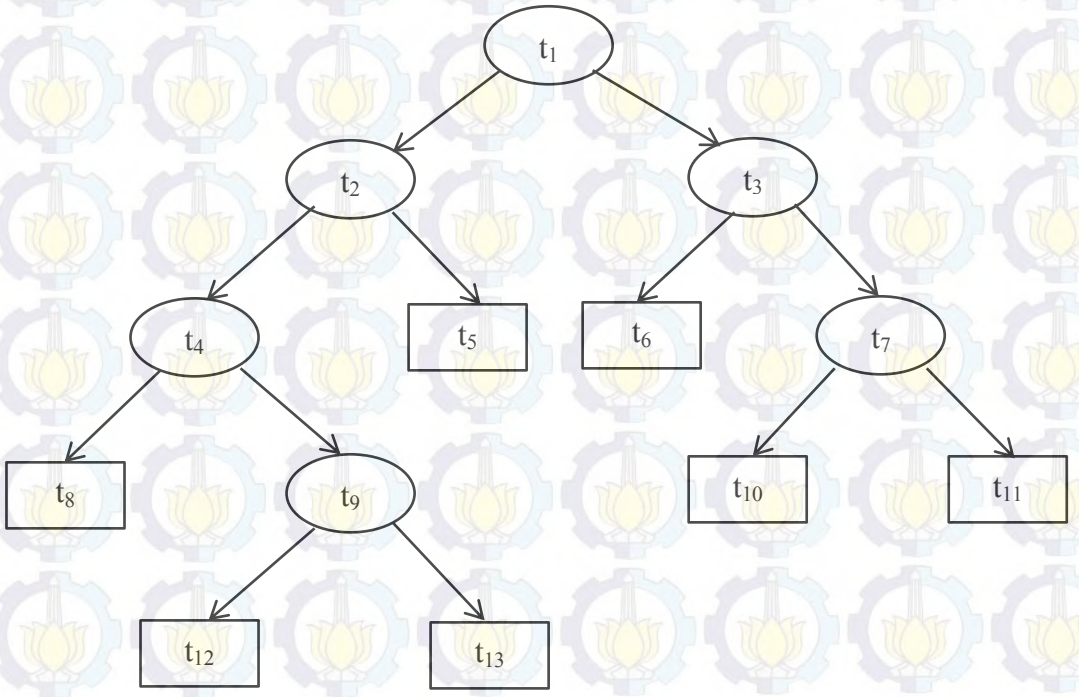
2.2. *Classification And Regression Trees* (*CART*)

Metode *CART* dikembangkan oleh Leo Breiman *et al* pada tahun 1984. *CART* adalah suatu metode statistik nonparametrik yang dapat menyeleksi variabel dan interaksinya yang terpenting. Dalam metode *CART*, pohon keputusan dibentuk dengan menggunakan algoritma penyekatan rekursif secara biner (*binary recursive partitioning*) (Lewis, 2000). *CART* merupakan cara pemilahan sekelompok data yang terkumpul dalam suatu ruang yang disebut simpul (*node*) menjadi dua simpul anak (*child nodes*) dimana simpul asli tersebut disebut simpul induk (*parent node*). Setiap simpul anak dapat dipilah lagi menjadi dua simpul anak berikutnya, begitu seterusnya dan berhenti jika telah mendapatkan sekelompok observasi atau objek yang relatif homogen. Istilah biner mengacu untuk simpul yang hanya dapat dipilah menjadi dua simpul berikutnya, sedangkan istilah rekursif mengacu pada proses penyekatan yang berulang-ulang. *CART* digunakan untuk menggambarkan hubungan antara variabel respon dengan satu atau lebih variabel prediktor. Tujuan utama *CART* adalah untuk mendapatkan suatu kelompok data yang akurat sebagai salah satu ciri dari suatu pengklasifikasian.

Metode ini adalah salah satu tehnik eksplorasi data yaitu suatu tehnik pohon keputusan (*decision tree*). Model pohon yang dihasilkan bergantung pada skala variabel respon, jika variabel respon data berbentuk kontinu maka model pohon yang dihasilkan adalah pohon regresi (*regression tree*) sedangkan jika variabel respon data bersifat kategori maka pohon yang dihasilkan adalah pohon klasifikasi (*classification tree*) (Breiman *et al*, 1984).

2.3. Pohon Klasifikasi

Pohon klasifikasi merupakan suatu metode klasifikasi yang menghasilkan model visual berupa model pohon. Menurut Lewis (2000), pohon klasifikasi memiliki beberapa kelebihan dan kekurangan. Kelebihannya adalah merupakan metode yang tidak terikat oleh berbagai macam asumsi yang harus dipenuhi oleh variabel prediktor, memudahkan dalam eksplorasi dan pengambilan keputusan karena struktur data yang diperoleh dapat dilihat secara visual menggunakan model klasifikasinya, dapat mengeksplorasi struktur data dengan banyak variabel, hasil klasifikasi akhir sederhana dan efisien, serta hasilnya lebih mudah untuk diinterpretasikan. Sedangkan kekurangannya yaitu memungkinkan mempunyai pohon keputusan yang tidak stabil yang akan mempengaruhi hasil akhir pengklasifikasian dan melakukan pemilahan hanya dengan satu variabel yang akan tidak tepat ketika digunakan pada struktur data yang kompleks.



Gambar 2.1. Struktur Pohon Klasifikasi

Simpul utama (*root node*) dinotasikan sebagai t_1, sedangkan simpul t_2, t_3, t_4, t_7 dan t_9 disebut simpul dalam (*internal nodes*). Simpul akhir yang juga disebut sebagai simpul terminal (*terminal nodes*) adalah t_5, t_6, t_8, t_{10}, t_{11}, t_{12} dan t_{13} dimana

pada simpul tersebut tidak terjadi lagi pemilahan. Kedalaman (*depth*) pohon klasifikasi dihitung dari simpul utama atau t_1 yang berada di kedalaman 1, t_2 dan t_3 berada di kedalaman 2, sampai pada t_{12} dan t_{13} yang berada pada kedalaman 5.

2.3.1. Proses Pembentukan Pohon Klasifikasi

Misalkan diambil N observasi sebagai data *learning* dan N_j adalah jumlah keseluruhan dari observasi yang termasuk dalam kelas ke-j, $j = 1, 2, \ldots, J$. Kemudian definisi dari probabilitas kelas adalah.

$$\{\pi(j)\}_{j=1}^{j=J} = \left\{\frac{N_j}{N}\right\}_{j=1}^{j=J} \tag{2.1}$$

Misalkan $N(t)$ banyaknya observasi pada simpul t dan $N_j(t)$ adalah banyaknya observasi dalam kelas ke-j pada simpul t yang sama, maka probabilitas gabungan suatu kejadian bahwa suatu observasi dari kelas ke-j berada dalam satu simpul t adalah.

$$P(j,t) = \pi(j)\frac{N_j(t)}{N_j} \tag{2.2}$$

Sehingga $P(t) = \sum_{j=1}^{J} P(j,t)$ dan probabilitas bersyarat suatu observasi termasuk dalam simpul t adalah.

$$P(j|t) = \frac{P(j,t)}{P(t)} = \frac{N_j(t)}{N(t)} \tag{2.3}$$

yaitu proporsi kelas j dalam simpul t dan $\sum_{j=1}^{J} P(j|t) = 1$.

Ukuran keheterogenan $i(t)$ yang digunakan sebagai indikator kehomogenan simpul pohon adalah.

$$i(t) = \varphi\big(P(1|t), P(2|t), \ldots, P(J|t)\big) \tag{2.4}$$

dimana $\varphi(t)$ adalah fungsi keheterogenan yang memenuhi sifat-sifat:

- $\varphi = \varphi(P_1, P_2, \ldots, P_J)$ adalah himpunan seluruh J pasangan terurut $(P_1, P_2, \ldots, P_J)$ yang memenuhi $\sum_{j=1}^{J} P_j = 1$ dan $0 \leq P_j \leq 1, j = 1,2, \ldots, J$.
- $\varphi \geq 0$
- φ hanya mencapai nilai maksimum pada $\left(\frac{1}{J}, \frac{1}{J}, \ldots, \frac{1}{J}\right)$, yaitu ketika memuat observasi-observasi dari seluruh kelas secara berimbang.

- φ hanya mencapai nilai minimum pada (1,0,0,…,0),(0,1,0,…,0),…,(0,0,0,…,1), yaitu ketika hanya memuat observasi-observasi yang berasal dari satu kelas.
- φ suatu fungsi simetrik dari $P_1, P_2, \dots, P_J$.

Dalam proses pembentukan ini, diperlukan data *learning L* yang terdiri dari *N* observasi dengan data (x_n, j_n), dimana $x_n \in X$ dan $j_n \in \{1,2,\dots,N\}, n = 1,2,\dots,N$. Jika *Y* sebagai variabel respon biner, maka metode pohon klasifikasi merupakan sebuah fungsi $d(x)$ yang memetakan *X* pada himpunan kelas *Y*. Jadi pengklasifikasi *d* mempartisi ruang *X* menjadi *J* himpunan bagian yang saling lepas $A_1, A_2, \dots, A_j$.

Proses pembentukan pohon klasifikasi terdiri dari tiga tahapan, yaitu *splitting rule* (pemilihan pemilah), penentuan simpul terminal dan penandaan label kelas.

1. *Splitting Rule* (*Pemilihan Pemilah*)

Breiman, *et al.* (1984) menyatakan bahwa aturan pemilah setiap simpul menjadi dua anak simpul adalah.

- Setiap pemilahan bergantung pada nilai yang berasal dari hanya satu variabel prediktor.
- Untuk variabel kontinu X_j, pemilahan ditentukan oleh apakah $X_j \leq c_i$, dimana nilai c_i adalah nilai tengah dari dua nilai observasi variabel X_j yang berurutan namun berbeda. Jika ruang sampel berukuran *n* dan terdapat *n* nilai observasi yang berbeda pada variabel X_j, maka akan terdapat sebanyak-banyaknya *n*-1 pemilahan yang berbeda.
- Untuk variabel yang bersifat kategori, pemilahan berasal dari semua kemungkinan pemilahan berdasarkan terbentuknya dua simpul yang saling lepas (*disjoint*). Jika variabel X_j merupakan kategori nominal bertaraf *L*, maka akan diperoleh sebanyak 2^{L-1}-1 pemilahan.

Splitting rule (Pemilihan Pemilah) pada data *learning* dilakukan dengan tujuan untuk mendapatkan setiap himpunan bagian turunan yang mempunyai sifat yang lebih homogen dibandingkan dengan himpunan bagian induknya. Hal ini dilakukan dengan cara menentukan fungsi keheterogenan (*impurity function*) pada

simpul *t*. Metode pemilihan pemilah yang digunakan adalah kriteria indeks *Gini* dan indeks *Twoing*.

a. Indeks *Gini*

Indeks *Gini* adalah aturan pemilahan yang paling sering digunakan. Bentuk fungsi keheterogenan Indeks *Gini* adalah.

$$i(t) = \sum_{j \neq i} p(j|t)p(i|t) \quad (2.5)$$

Bentuk lain dari indeks *Gini* adalah.

$$i(t) = 1 - \sum_{j \neq i} p^2(j|t) \quad (2.6)$$

Indeks *Gini* selalu memisahkan kelas yang anggotanya paling besar lebih dahulu atau yang merupakan kelas terpenting dalam simpul tersebut.

b. Indeks *Twoing*

Secara mendasar, cara kerja indeks *Twoing* akan membagi kelas menjadi dua kelas besar. Bentuk fungsi kehetogenan indeks *Twoing* adalah.

$$\phi(s,t) = \frac{p_L p_R}{4}\left[\sum_{j=1}^{J}|p(j|t_L) - p(j|t_R)|\right]^2 \quad (2.7)$$

Setelah itu, memilih pemilah terbaik dari masing-masing variabel prediktor dan memilah pemilah terbaik dari kumpulan pemilah terbaik. Pemilah terbaik adalah pemilah yang memaksimumkan ukuran kehomogenan masing-masing simpul anak relatif terhadap simpul induknya dan memaksimumkan ukuran pemisahan antara dua simpul anak tersebut.

Goodness of Split digunakan untuk mengevaluasi pemilahan yang dilakukan oleh pemilah *s* pada simpul *t*. Misalkan terdapat calon pemilah *s* yang memilah *t* menjadi simpul kiri t_L dengan proporsi p_L dan simpul kanan t_R dengan proporsi p_R, maka *Goodness of Split* $\phi(s,t)$ didefinisikan sebagai penurunan keheterogenan.

$$\phi(s,t) = \Delta i(s,t) = i(t) - p_L i(t_L) - p_R i(t_R) \quad (2.8)$$

Pengembangan pohon dilakukan dengan pencarian semua kemungkinan pemilahan pada simpul t_1, sehingga pada akhirnya diperoleh pemilah s^* yang memberikan nilai penurunan keheterogenan yang paling tinggi, yaitu.

$$\Delta i(s^*, t_1) = \max_{s \subset S} \Delta i(s, t_1) \quad (2.9)$$

Simpul t_1 akan dipilah menjadi t_2 dan t_3 dengan pemilah s^*. Cara yang sama dilakukan pada t_2 dan t_3 secara terpisah, begitu seterusnya.

2. *Penentuan Simpul Terminal*

Suatu simpul *t* akan menjadi simpul terminal jika:

a. Pada simpul *t* sudah tidak ada lagi penurunan keheterogenan secara berarti (devian simpul terminal < 1% devian simpul utama) sehingga simpul *t* tidak akan terpilah lagi.
b. Pada setiap simpul anak hanya memuat satu observasi atau adanya batasan minimum *n* yang telah terpenuhi. Pengembangan pohon berhenti apabila pada simpul *t* terdapat $n_i < 5$.
c. Adanya batasan jumlah tingkat kedalaman pohon maksimal, sehingga klasifikasi pohon berhenti.

3. *Penandaan Label Kelas*

Label kelas pada simpul terminal ditentukan berdasarkan aturan jumlah terbanyak, yaitu jika $p(j_0|t) = \max_j \frac{N_j(t)}{N(t)}$, maka label kelas untuk simpul terminal *t* adalah j_0 yang memberikan nilai dugaan kesalahan klasifikasi pada simpul *t* paling kecil sebesar $r(t) = 1 - \max_j p(j|t)$.

Oleh karena penghentian pohon berdasarkan banyaknya observasi pada simpul akhir atau besarnya tingkat kehomogenan, maka pohon yang dibentuk dengan aturan pemilah dan kriteria *goodness of split* berukuran sangat besar. Semakin banyak pemilahan yang dilakukan, maka berakibat semakin kecilnya tingkat kesalahan prediksi. Ukuran pohon yang besar dapat menimbulkan adanya *overfitting*. Namun jika pengembangan pohon dibatasi dengan ketepatan batas tertentu, akan timbul kasus sebaliknya yaitu *underfitting*. Cara mengatasi hal ini adalah dengan mencari ukuran pohon yang layak yang akan dilakukan dengan:

a. Penentuan pohon awal yang besar.
b. Pohon dipangkas secara iteratif menjadi sekuen pohon yang makin kecil dan tersarang.
c. Pohon terbaik dipilih dari sekuen ini dengan menggunakan sampel uji (*test sample*) atau sampel validasi silang (*cross validation sample*).

2.3.2. Skor Variabel Penting (*Variable Importance Score*)

Skor variabel penting digunakan untuk mengidentifikasi urutan variabel yang berpengaruh dan masuk dalam model pohon klasifikasi. Urutan pertama menandakan variabel yang paling dominan dan menjadi pemilah utama dalam model. Skor ini dihitung berdasarkan penjumlahan semua kemungkinan *split* (pemilah) dalam satu variabel pada tiap simpul. Skor variabel penting x_m didefinisikan sebagai berikut.

$$M(x_m) = \sum_{t \in T} \Delta I(\tilde{s}_m, t) \tag{2.10}$$

Misalkan T adalah model pohon klasifikasi yang dihasilkan, jika aturan pemilahannya menggunakan kriteria indeks *Gini*, maka untuk tiap simpul $t \in T$ dihitung $\Delta I(\tilde{s}_m, t)$. Jika menggunakan kriteria indeks *Twoing*, maka yang digunakan adalah $\Delta I(\tilde{s}_m, t) = \max_{C_1} \Delta I(\tilde{s}_m, t, C_1)$. Skor variabel penting yang digunakan adalah nilai yang dinormalisasi berdasarkan ukuran $\frac{100M(x_m)}{\max_m M(x_m)}$, sehingga variabel yang paling penting akan memiliki skor 100, sedangkan variabel yang lain akan berada di rentang 0 sampai 100.

2.3.3. Pemangkasan Pohon Klasifikasi (*Prunning*)

Untuk mendapatkan pohon yang layak, maka diperlukan pemangkasan (*pruning*), yaitu suatu penilaian ukuran pohon tanpa mengurangi ketepatan dan kebaikannya dengan cara mengurangi simpul pohon, sehingga diperoleh ukuran pohon yang layak. Pemangkasan dilakukan jika pohon klasifikasi yang terbentuk berukuran sangat besar dan kompleks dalam menggambarkan struktur data. Pemangkasan dilakukan dengan cara memangkas bagian pohon yang kurang penting, sehingga diperoleh pohon optimal. Untuk melakukan ini, digunakan *cost complexity minimum* (Breiman, *et al*. 1984).

Untuk sembarang pohon T yang merupakan sub pohon dari pohon terbesar T_{max}($T < T_{max}$), ukuran *cost complexity* adalah.

$$R_\alpha(T) = R(T) + \alpha|\tilde{T}| \tag{2.11}$$

dimana

$R(T)$ = *Resubtitution estimate* (proporsi kesalahan pada sub pohon)

α = *Complexity parameter* (parameter kompleksitas)

$|\tilde{T}|$ = Ukuran banyaknya simpul terminal pohon *T*

$R_\alpha(T)$ merupakan kombinasi linier dari *resubtitution estimate* dan kompleksitasnya. Jika dilihat dari nilai kompleksitas tiap simpul terminal *T*, maka nilai $R_\alpha(T)$ ditunjukkan dengan menambahkan *cost penalty* bagi kompleksitas terhadap biaya kesalahan klasifikasi.

Breiman *et al*, (1984) menyatakan bahwa *cost complexity pruning* menentukan suatu pohon bagian $T(\alpha)$ yang meminimumkan $R_\alpha(T)$ pada seluruh pohon bagian, atau untuk setiap nilai α, dicari pohon bagian $T(\alpha) < T_{max}$ yang meminimumkan $R_\alpha(T)$, yaitu.

$$R_\alpha(T(\alpha)) = \min_{T<T_{max}} R_\alpha(T) \tag{2.12}$$

Apabila proses pemangkasan diulang sampai tidak ada lagi pemangkasan yang mungkin terjadi, maka akan didapatkan suatu barisan menurun dan tersarang dari pohon bagian, yaitu $\{T_1, T_2, \dots, \{t_1\}\}, T_1 > T_2 > \cdots > \{t_1\}$ dan suatu barisan menaik, yaitu $\{\alpha_k\}$, dengan $\alpha_1 = 0 < \alpha_2 < \alpha_3 < \cdots < \alpha_k, T(\alpha) = T(\alpha_k) = T_k$. Ketika *R(T)* digunakan sebagai kriteria penurunan pohon optimal, maka T_1 akan cenderung dipilih, karena semakin besar pohon akan semakin kecil nilai *R(T)* nya.

2.3.4. Pohon Klasifikasi Optimal

Pohon klasifikasi optimal yang dipilih adalah pohon optimal yang berukuran tepat dan mempunyai nilai penduga pengganti yang cukup kecil. Ukuran pohon klasifikasi yang sangat besar akan memberikan nilai penduga pengganti yang sangat kecil, sehingga pohon ini cenderung dipilih untuk menduga nilai respon. Yang perlu disoroti adalah ukuran pohon yang besar akan memiliki nilai kompleksitas yang tinggi karena struktur data yang digambarkan cenderung kompleks. Data sampel akan digunakan untuk mendapatkan nilai pengganti yang paling kecil dari pohon klasifikasi yang dipilih.

Ada dua macam penduga pengganti, yaitu penduga sampel uji (*test sample estimate*) dan penduga validasi silang lipat-*V* (*cross validation V-fold estimate*).

1. *Test Sample Estimate*

Setelah pohon klasifikasi T terbentuk, observasi-observasi dalam L melewati pohon tersebut. Proporsi observasi yang mengalami kesalahan pengklasifikasian merupakan penduga pengganti, yaitu.

$$R(T) = \frac{1}{N}\sum_{n=1}^{N} X(d(x_n) \neq j_n) \tag{2.13}$$

dimana $X(.)$ adalah suatu fungsi indikator yang berbentuk.

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

Pada penduga sampel, L dibagi menjadi dua himpunan secara acak, yaitu L_1 untuk membentuk pohon T dan L_2 untuk menduga $R(T)$. Jika N_2 adalah banyaknya observasi dalam L_2, maka penduga sampel uji adalah.

$$R^{ts}(T_k) = \frac{1}{N_2}\sum_{x_n, j_n \in L_2} X(d(x_n) \neq j_n) \tag{2.14}$$

Pohon optimal yang dipilih adalah T^* dengan kriteria $R^{ts}(T^*) = \min_k R^{ts}(T_k)$.

2. *Cross Validation V-Fold Estimate*

Observasi dalam L dibagi secara acak menjadi V bagian yang saling lepas dengan ukuran yang hampir sama besar untuk tiap kelasnya. Pohon $T^{(v)}$ dibentuk dari $L - L_v$ dengan $v = 1,2,\dots,V$. Misalkan $d^{(v)}(x)$ adalah hasil pengklasifikasian, maka penduga sampel uji untuk $R(T_t{}^{(v)})$ adalah.

$$R^{ts}\left(T_k^{(v)}\right) = \frac{1}{N_v}\sum_{x_n, j_n \in L_v} X(d^v(x_n) \neq j_n) \tag{2.15}$$

dimana $N_v \cong N/V$ adalah jumlah observasi dalam L_v.

Dengan observasi induk L, dibentuk deretan pohon $\{T_k\}$, maka penduga validasi silang lipat V untuk $T_k^{(v)}$ adalah.

$$R^{cv}(T_k) = \frac{1}{V}\sum_{v=1}^{V} R^{ts}(T^{(v)}) \tag{2.16}$$

Pohon klasifikasi optimum dipilih T^* dengan $R^{cv}(T^*) = \min_k R^{cv}(T_k)$.

Pemilihan ukuran V sangat menentukan kecepatan dalam memeperoleh pohon optimal yang tepat. Ukuran V = 10 merupakan ukuran V yang sering digunakan dalam penelitian. Breiman menyarankan penggunaan *cross validation V-fold estimate* untuk menghitung biaya pengganti relatif pada sampel yang jumlah yang kecil, sedangkan penggunaan pendekatan *test sample estimate* digunakan untuk jumlah sampel yang lebih besar.

2.4. *Bootstrap*

Metode *bootstrap* pertama kali diperkenalkan oleh *Efron* (1979). Metode *bootstrap* merupakan suatu metode penaksiran nonparametrik yang dapat menaksir parameter-parameter pada suatu distribusi, variansi sampel median, serta dapat menaksir tingkat kesalahan (*error*). Pada metode *bootstrap* dilakukan pengambilan sampel dengan pengembalian (*resampling with replacement*) pada sampel data.

Secara singkat algoritma bootstrap dapat dinyatakan sebagai berikut:

1. Ambil sampel berukuran *n*, yaitu $S: x_1, \dots, x_n$.
2. Ambil sampel kembali dari *S* dengan pengembalian berukuran *n* dan dapatkan nilai statistik $\hat{\theta}_i$ untuk sampel S_i.
3. Lakukan langkah 2 sebanyak *B* kali.
4. Tentukan nilai statistik dengan *bootstrap*:

$$\hat{\theta}_b = B^{-1} \sum_{i=1}^{B} \hat{\theta}_i \tag{2.17}$$

dan

$$\hat{\sigma}^2_{\theta_b} = \frac{\sum_{i=1}^{B} (\hat{\theta}_i - \hat{\theta}_b)^2}{(B-1)} \tag{2.18}$$

2.5. *Bootstrap Aggregating (Bagging)*

Bootstrap aggregating (Bagging) adalah tehnik yang diusulkan oleh Breiman (1996) sebagai alat untuk memperbaiki stabilitas dan kekuatan prediksi klasifikasi dan regresi pohon dengan cara mereduksi variansi dari suatu prediktor. Ide dasar dari *bagging* adalah menggunakan *bootstrap resampling* untuk membangkitkan prediktor dengan banyak versi. *Bagging* merupakan salah satu prosedur intensif perhitungan untuk memperbaiki estimator atau pengklasifikasi yang tidak stabil, khususnya masalah data dimensi tinggi.

Menurut Breiman (1996), *bagging* merupakan implementasi sederhana dari pembangkitan replikasi quasi *learning sample*. Didefinisikan peluang dari kasus ke-*n* dari suatu *learning sample* adalah $p(n) = {}^{1}/_{N}$. Kemudian ambil sampel sebanyak *N* kali dari distribusi $\{p(n)\}$, secara ekuivalen merupakan sampel dari *T* dengan pengembalian. Himpunan sampel dari *T* disampel kembali

menjadi himpunan *learning sample* T'. T' lebih dikenal dengan istilah sampel *bootstrap* dari *T*.

Penambahan tehnik *bagging* dilakukan dalam metode pohon klasifikasi. Prosedur *bagging* dijalankan pada saat melakukan pembentukan pohon klasifikasi. Penerapan tehnik *bagging* mempunyai tujuan untuk memberikan banyak versi prediktor dan menggunakannya untuk memperoleh aggregat prediktor. Dalam kasus klasifikasi, *B* sampel *bootstrap* diambil dari data *learning*, kemudian pada setiap sampel *bootstrap* tersebut diterapkan metode pohon klasifikasi untuk menghasilkan prediksi kelas dari suatu data input *X* (prediktor) yang sudah ditetapkan. Prediksi akhir merupakan kelas yang paling sering terjadi pada hasil prediksi *B* sampel *bootstrap*. Hasil ketepatan klasifikasi dari penerapan *bagging* tergantung pada berapa banyak replikasi sampel *bootstrap* yang dibuat sehingga penentuan banyaknya replikasi yang harus dibuat menjadi sangat krusial.

Ada beberapa peneliti yang memberikan rekomendasi tentang banyaknya replikasi *bootstrap* sampling yang harus dibuat, diantaranya Sutton (2004) merekomendasikan untuk melakukan replikasi sebanyak 25 atau 50 kali. Sedangkan Hastie *at al*. (2001) menyatakan bahwa peningkatan akurasi akan terjadi jika banyaknya replikasi ditingkatkan dari 50 ke 100 kali dan jika replikasinya ditingkatkan menjadi yang lebih dari 100 kali akan menghasilkan akurasi yang tidak lebih besar dibandingkan replikasi 100 kali.

Buhlmann dan Yu (2002), secara singkat menyatakan algoritma *bagging*.

1. Menyusun sampel *bootstrap* $\mathcal{L}_i^* = (Y_i^*, \mathbf{x}_i^*),\ i = 1, \dots, n,$ menurut distribusi empiris pada pasangan $\mathcal{L}_i = (Y_i, \mathbf{x}_i),\ i = 1, \dots, n.$
2. Hitung estimator *bootstrap* $\hat{\theta}_n^*(x)$ dengan prinsip *plug-in* yaitu:

$$\hat{\theta}_n^*(x) = h_n(\mathcal{L}_1^*, \dots, \mathcal{L}_n^*) \tag{2.19}$$

dengan

$$\hat{\theta}_n(x) = h_n(\mathcal{L}_1, \dots, \mathcal{L}_n) \tag{2.20}$$

3. Estimator bagging adalah sebagai berikut:

$$\hat{\theta}_{n_B}(x) = E^*\left[\hat{\theta}_{n_B}^*(x)\right] \approx B^{-1} \sum_{b=1}^{B} \hat{\theta}_{n_B}^*(x) \tag{2.21}$$

Menurut Buhlmann dan Yu (2002), secara heuristik kinerja variansi estimator *bagging* $\hat{\theta}_{n_B}(x)$ adalah sama dengan atau lebih kecil dibandingkan estimator orisinil $\hat{\theta}_n^*(x)$.

Breiman (1996) menjabarkan langkah-langkah untuk prosedur *bootstrap* untuk pohon klasifikasi adalah sebagai berikut.

1. Data dibagi menjadi dua bagian yaitu data *learning* *L* dan data *testing* *T*. Ukuran data *testing* diambil dengan jumlah yang lebih sedikit daripada ukuran data *learning*.
2. Pohon klasifikasi dibentuk dengan menggunakan data *learning*. Selanjutnya jalankan data *testing* pada pohon *L* dan didapatkan dugaan kesalahan pengklasifikasian $e_s(L,T)$.
3. Bangun sampel *bootstrap* yang dipilih secara acak dengan pengembalian dari *L* berukuran sama dengan data *L* dan menghasilkan data *learning* baru yang disebut sebagai data *learning bootstrap* L_B. *Bootstrap* dilakukan dengan replikasi sebanyak *M* kali sehingga menghasilkan *M* buah data *learning bootstrap*.
4. Bentuk pohon klasifikasi dengan menggunakan L_B. Jalankan data *T* pada pohon yang dibentuk dengan L_B dan akan menghasilkan berbagai macam pohon klasifikasi $\{\varphi_1(x), \varphi_2(x), \dots, \varphi_B(x)\}$.
5. Jika $(y_n, x_n) \in T$ maka estimasi kelas dari x_n akan memiliki bentuk jamak $\{\varphi_1(x), \varphi_2(x), \dots, \varphi_B(x)\}$. Proporsi estimasi kelas yang berbeda dari kelas sebenarnya merupakan angka kesalahan klasifikasi *bagging* $e_B(L,T)$.
6. Pembagian data secara random dilakukan dengan iterasi sebanyak *n* kali dan dihitung nilai $\bar{e}_S, \bar{e}_B$ yang merupakan rata-rata *n* iterasi. Menurut Breiman banyaknya iterasi yang dilakukan adalah 100 kali.

2.6. Ketepatan Klasifikasi

Untuk melihat ketepatan klasifikasi, dapat dilakukan dengan menghitung peluang kesalahan klasifikasi. Ukuran yang dapat digunakan adalah *Total Accuracy Rate, Total Error Rate, Sensitivity* dan *Specitivity*. *Total Accuracy Rate* merupakan proporsi observasi yang diprediksi secara benar oleh fungsi klasifikasi,

sedangkan *Total Error Rate* adalah fraksi (proporsi) pengamatan pada sampel yang salah diklasifikasikan oleh fungsi klasifikasi (Johnson dan Wichern, 1992). Secara umum tabel hasil klasifikasi diberikan oleh tabel berikut.

Tabel 2.1. Hasil Klasifikasi

Actual	Predicted		Total
	0	1	
0	n_{11}	n_{12}	N_1
1	n_{21}	n_{22}	N_2
Total	N_1	N_2	N

dimana

n_{11} = Jumlah observasi dari 0 yang tepat diprediksikan sebagai 0

n_{22} = Jumlah observasi dari 1 yang tepat diprediksikan sebagai 1

n_{12} = Jumlah observasi dari 0 yang salah diprediksikan sebagai 1

n_{21} = Jumlah observasi dari 1 yang salah diprediksikan sebagai 0

N_1 = Jumlah observasi dari 0

N_2 = Jumlah observasi dari 1

N = Jumlah observasi

$$\textit{Total Accuracy Rate} = \frac{n_{11}+n_{22}}{N} \tag{2.22}$$

$$\textit{Total Error Rate} = \frac{n_{12}+n_{21}}{N} \tag{2.23}$$

$$\textit{Sensitivity} = \frac{n_{11}}{N_1} \tag{2.24}$$

$$\textit{Specitivity} = \frac{n_{22}}{N_2} \tag{2.25}$$

2.7. Pekerja Anak

Anak adalah generasi yang akan menjadi penerus bangsa sehingga mereka harus dipersiapkan dan diarahkan sejak dini agar dapat tumbuh dan berkembang menjadi anak yang sehat jasmani dan rohani, maju, mandiri dan sejahtera menjadi sumber daya yang berkualitas serta dapat menghadapi tantangan di masa datang. Di Indonesia, setiap anak memiliki hak untuk tumbuh,

berkembang dan dilindungi sebagaimana diatur oleh Undang-Undang No. 23 Tahun 2002 tentang perlindungan anak. Segala jenis kegiatan pengeksploitasi anak merupakan tindakan yang dilarang, termasuk mempekerjakan anak di bawah umur, yang biasa disebut pekerja anak.

Pekerja anak diartikan sebagai anak yang harus melakukan pekerjaan yang menghalangi mereka bersekolah dan membahayakan kesehatan, fisik dan mentalnya (Manik, 2006). Pekerja anak adalah masalah sosial yang telah menjadi isu dan agenda global bangsa-bangsa di dunia, begitu pun halnya di Indonesia. Pekerja anak diyakini akan menjadi salah satu masalah terbesar yang akan dihadapi seiring perkembangan waktu, menyusul krisis perekonomian dunia dan tingginya tuntutan di era globalisasi.

Pekerja anak dimanapun mereka berada, dilihat secara umum kondisi dan situasinya diyakini akan mengancam kehidupan dan juga masa depannya. Dunia anak seharusnya dunia yang penuh kegembiraan, bermain, sekolah, perhatian dan kasih sayang orang tua. Suasana tersebut sebagai proses pendukung tumbuh dan berkembang seorang anak yang dapat memberikan landasan untuk kehidupan masa depannya mengingat anak adalah masa depan suatu bangsa. Berbagai studi tentang pekerja anak seringkali di temukan bahwa seorang pekerja anak selalu berada di kondisi yang tidak menguntungkan, rentan dalam bentuk eksploitasi dan minim dalam akses pengembangan diri secara fisik, mental, spritual, moral

Dalam sudut pandang ekonomi, anak merupakan penduduk non produktif karena masih tergolong usia belajar sehingga tugas utama anak adalah sekolah. Namun kemiskinan seringkali dijadikan alasan pembenaran terhadap praktek mempekerjakan anak dalam upaya untuk membantu memenuhi kebutuhan ekonomi keluarga. Dalam laporan ILO, bertambahnya jumlah keluarga miskin akan memunculkan lebih banyak anak yang putus sekolah dan dipaksa untuk bekerja. Menurut Direktur Jenderal UNESCO, Koichiro Matsuura, untuk mengatasi pekerja anak, pendekatan yang terintegrasi perlu dilakukan dengan mengurangi kemiskinan, menyediakan pendidikan bagi semua dan meningkatkan pengembangan ekonomi dan sosial. Harus ada jaminan pendidikan dasar gratis untuk keluarga miskin.

Kemiskinan bukan satu-satunya penyebab seorang anak bekerja di bawah umur, faktor lingkungan juga berperan sangat penting. Anak-anak yang hidup di lingkungan teman-teman yang cenderung menyukai bekerja daripada sekolah juga cenderung akan terjerumus dalam dunia kerja, meskipun orang tua mereka cukup mampu untuk membiayai sekolah mereka. Sebab lingkungan teman-teman sebaya mempunyai pengaruh yang cukup besar dalam menanamkan nilai-nilai tertentu yang mereka anggapan sesuai dengan dunia mereka.

Jadi jika dirunut satu per satu, faktor-faktor yang bisa menjerumuskan anak dalam dunia kerja mencakup faktor ekonomi, keadaan rumah tangga, pendidikan, baik pendidikan anak maupun pendidikan dalam ruang lingkup keluarga, keadaan psikologis anak, serta faktor budaya, sosial dan demografi. Tetapi tidak bisa ditampik bahwa faktor kemiskinan menjadi faktor utama yang menyebabkan terjadinya prektek mempekerjakan anak.

(*Halaman Ini Sengaja Dikosongkan*)

BAB 3
METODE PENELITIAN

3.1. Sumber Data

Data yang digunakan dalam penelitian ini adalah data sekunder yang yang berasal dari Survey Sosial Ekonomi Nasional (Susenas) provinsi Sulawesi Tengah tahun 2013. Data tersebut akan dibagi menjadi dua bagian, yaitu data *learning* dan data *testing*. Data *learning* berguna untuk verifikasi model, yaitu kemampuan model dalam mengakomodasi keragaman data. Sedangkan data *testing* berguna untuk validasi model, yaitu seberapa besar kemampuan model untuk memprediksi data baru. Besarnya pembagian data menjadi data *learning* dan data *testing* tidak ada batasan (Briman *et al*, 1984).

Susenas merupakan salah satu survei yang mengumpulkan jenis data sosial dan ekonomi secara tahunan dengan jumlah sampel yang relatif cukup besar yang terdiri dari data individu dan data rumah tangga. Data individu berisi tentang karakteristik individu, seperti usia, partisipasi sekolah, kesehatan, pekerjaan, dan lain-lain. Data rumah tangga berisi tentang karakteristik rumah tangga, seperti pengeluaran sebulan yang lalu untuk makanan dan non makanan, keterangan perumahan, dan lain-lain.

Survei Sosial Ekonomi Nasional (Susenas) Kor merupakan salah satu kegiatan Badan Pusat Statistik. Sejak tahun 1963 BPS menyelenggarakan Susenas, mulai tahun 1992 sampai dengan tahun 2010 Susenas Kor (Pokok) dan Modul (Rinci) dilaksanakan setiap tahun. Untuk modul dilakukan setiap tiga tahun secara bergilir. Mulai tahun 2011 Susenas Kor dan Modul Konsumsi dilaksanakan secara triwulanan setiap tahun. Tujuan Susenas KOR adalah mendapatkan data berkaitan dengan kesejahteraan rakyat. Bagi pemerintah, tersedianya data untuk perencanaan pembangunan sektoral maupun lintas sektoral.

Adapun *sampling frame* Susenas, terdiri dari tiga tahapan. Kerangka sampel pemilihan tahap pertama adalah daftar wilayah pencacahan (wilcah) SP2010 yang disertai dengan informasi banyaknya rumah tangga hasil listing SP2010 (Daftar RBL1), muatan blok sensus dominan (pemukiman biasa,

pemukiman mewah, pemukiman kumuh), informasi daerah sulit/tidak sulit, dan klasifikasi desa/kelurahan (rural/urban). Kerangka sampel pemilihan tahap kedua adalah daftar blok sensus pada setiap wilcah terpilih. Kerangka sampel pemilihan tahap ketiga adalah daftar rumah tangga biasa tidak termasuk *institutional household* (panti asuhan, barak polisi/militer, penjara, dsb) dalam setiap blok sensus sampel hasil pencacahan lengkap SP2010 (SP2010-C1) yang telah dimutahirkan pada setiap menjelang pelaksanaan survei.

Sedangkan untuk rancangan sampel Susenas terdiri dari tiga tahap. Tahap pertama, memilih nh wilcah dari *Nh* secara *PPS* (*Probability Proportional to Size*) dengan size banyaknya rumah tangga SP2010 (Mi). Kemudian wilcah tersebut dialokasikan secara acak ke dalam empat triwulan. Tahap kedua, memilih dua BS pada setiap wilcah terpilih Susenas Triwulan II, dan III, serta Triwulan I yang juga terpilih untuk Sakernas Triwulan I, yang selanjutnya dari blok-blok sensus terpilih dialokasikan secara acak satu untuk Susenas/SBH, dan satu Sakernas, atau memilih satu BS pada setiap wilcah terpilih Triwulan IV dan Triwulan I yang untuk Susenas saja secara *PPS* dengan *size* jumlah rumah tangga SP2010-RBL1. Tahap ketiga, dari setiap blok sensus terpilih untuk Susenas dipilih sejumlah rumah tangga biasa $(m = 10)$ secara sistematik berdasarkan hasil pemutakhiran *listing* rumah tangga SP2010-C1 dengan menggunakan daftar VSEN11-P. Daftar nama kepala rumah tangga disusun dari ekstrak SP2010-C1 untuk variabel nama KRT, alamat, dan tingkat pendidikan KRT, kemudian dilakukan pemutakhiran lapangan.

Selanjutnya pengumpulan data dari rumah tangga terpilih dilakukan dengan wawancara antara pencacah dan responden. Keterangan tentang rumah tangga dikumpulkan melalui wawancara dengan kepala rumah tangga, suami/istri kepala rumah tangga atau rumah tangga lain yang mengetahui karakteristik yang ditanyakan. Sedangkan untuk pertanyaan-pertanyaan dalam kuesioner Susenas yang ditujukan kepada individu diusahakan agar individu yang bersangkutan menjadi responden (BPS, 2013).

3.2. Variabel Penelitian

Berikut akan dijabarkan variabel-variabel yang digunakan dalam penelitian ini beserta definisi operasionalnya. Variabel respon (Y) adalah status bekerja anak dengan skala nominal yang dikategorikan menjadi:

$$Y = \begin{cases} 0, & \text{jika anak tidak bekerja} \\ 1, & \text{jika anak bekerja} \end{cases}$$

Pekerja anak adalah anggota rumah tangga umur 10-18 tahun yang bekerja minimal satu jam selama seminggu yang lalu.

Sedangkan variabel prediktor adalah sebagai berikut:

X_1 = Umur anak (tahun), dengan skala interval. Umur anak adalah lama hidup yang telah dijalani oleh anak dan dihitung berdasarkan ulang tahun terakhir.

X_2 = Jenis kelamin anak, dengan skala nominal. Jenis kelamin anak adalah keadaan biologis anak yang membedakannya untuk dikategorikan sebagai jenis kelamin laki-laki atau perempuan.

1 = Laki-laki, 2 = Perempuan.

X_3 = Partisipasi sekolah anak, dengan skala nominal. Partisipasi sekolah anak adalah keikutsertaan anak dalam kegiatan pendidikan.

0 = Tidak sekolah, 1 = Sekolah

X_4 = Umur kepala rumah tangga (Tahun), dengan skala interval. Umur kepala rumah tangga adalah lama hidup yang telah dijalani oleh kepala rumah tangga dan dihitung berdasarkan ulang tahun terakhir.

X_5 = Tingkat pendidikan kepala rumah tangga, dengan skala ordinal. Tingkat pendidikan adalah jenjang pendidikan tertinggi yang ditamatkan, yang biasanya ditandai dengan ijazah.

0 = tidak memiliki ijazah

1 = SD

2 = SMP

3 = SMA

4 = Di atas SMA

X_6 = Jenis kelamin kepala rumah tangga, dengan skala nominal. Jenis kelamin kepala rumah tangga adalah keadaan biologis kepala rumah tangga yang membedakannya untuk dikategorikan sebagai jenis kelamin laki-laki atau perempuan.

1 = Laki-laki, 2 = Perempuan

X_7 = Sektor pekerjaan kepala rumah tangga, dengan skala nominal. Sektor pekerjaan adalah bidang kegiatan dari pekerjaan/usaha/perusahaan.

1 = Pertanian

2 = Industri

3 = Perdagangan

4 = Jasa

5 = Lainnya

X_8 = Pendapatan perkapita rumah tangga (Rupiah), dengan skala interval. Pendapatan perkapita adalah imbalan atau penghasilan yang biasa diterima selama sebulan dari pekerjaan utama baik berupa uang maupun barang yang dihitung per individu dalam rumah tangga.

X_9 = Jumlah anggota rumah tangga (Orang), dengan skala interval. Anggota rumah tangga adalah semua orang yang biasanya bertempat tinggal di suatu rumah tangga, baik yang berada di rumah pada waktu pencacahan maupun sementara tidak ada.

3.3. Metode Analisis

Tahapan-tahapan analisis data untuk mencapai tujuan penelitian adalah sebagai berikut.

1. Membahas konsep *splitting rule* dalam pembentukan pohon klasifikasi.
2. Mendapatkan model pohon klasifikasi pada kasus pekerja anak di provinsi Sulawesi Tengah.

   Langkah-langkah membentuk model pohon klasifikasi adalah sebagai berikut.

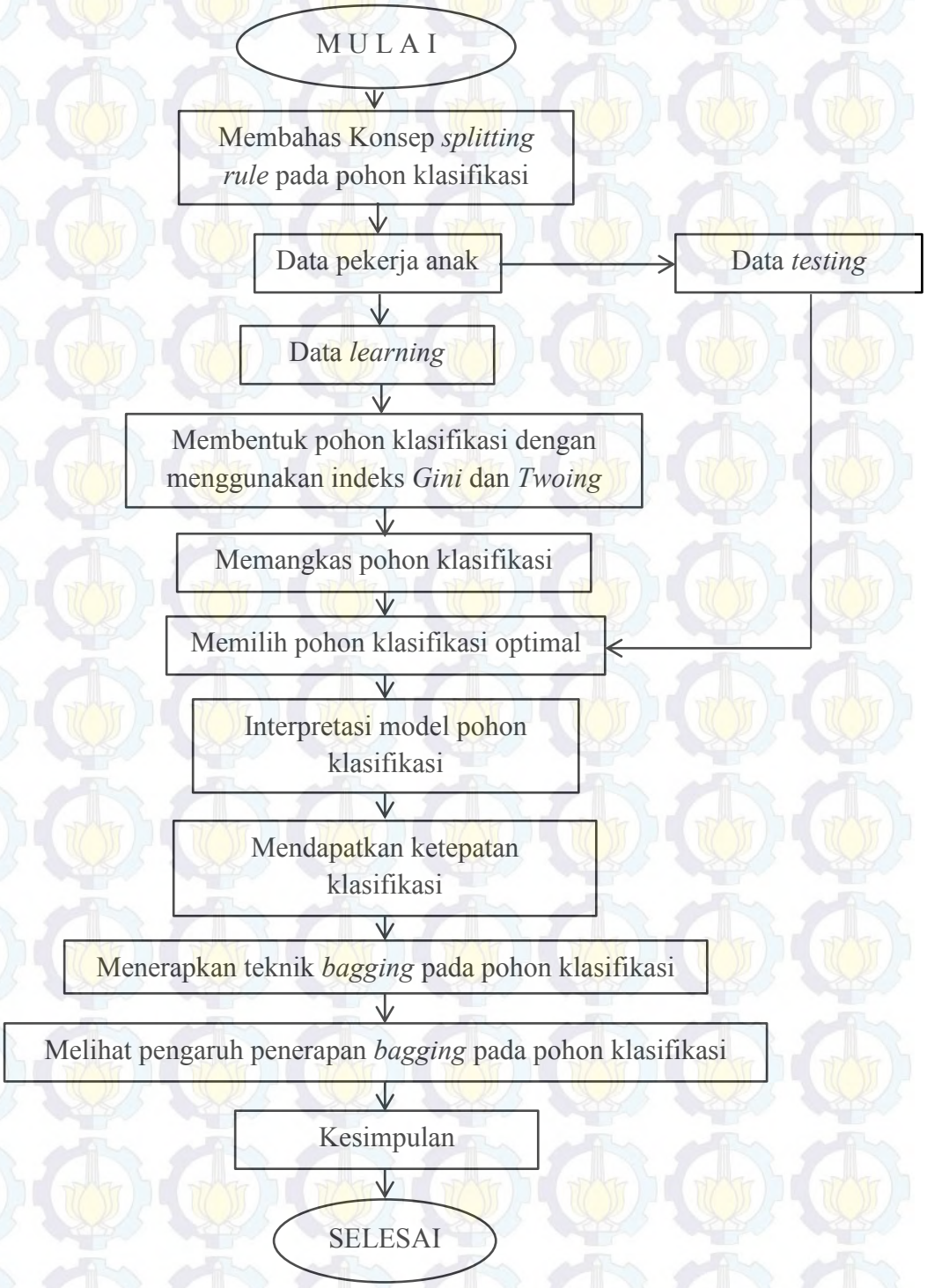
   - Membuat analisis deskriptif terhadap variabel respon dan variabel prediktor.
   - Membagi data menjadi dua bagian, yaitu data *learning* dan data *testing*. Selanjutnya, data yang akan dianalisis adalah data *learning*.

- Membentuk pohon klasifikasi dengan menggunakan indeks *Gini* dan indeks *Twoing*.
  - Menentukan pemilah dan pemilahan secara rekursif pada simpul, yaitu pemilahan yang memberikan penurunan keheterogenan tertinggi.
  - Menentukan simpul terminal, dilakukan jika suatu simpul *t* dicapai, sehingga tidak terdapat penurunan keheterogenan secara berarti.
  - Menentukan penandaan label kelas dari simpul terminal ditentukan berdasarkan aturan jumlah terbanyak.
- Menghentikan pembentukan pohon klasifikasi, dilakukan dengan menentukan minimum *n* pada simpul anak, kedalaman dalam pohon maksimal.
- Menghitung skor variabel penting pada model pohon klasifikasi yang terbentuk.
- Memangkas pohon klasifikasi, dilakukan dengan menggunakan kriteria ukuran *cost complexity minimum*.
- Memilih pohon klasifikasi optimal dengan *test sample estimation*.
- Memilih model pohon terbaik dengan *test set relative cost* yang minimum.
- Menguji keakuratan model pohon klasifikasi. Validasi model pohon klasifikasi dilakukan dengan menjalankan data *testing* ke dalam model tersebut sehingga dihasilkan angka ketepatan klasifikasi.

3. Menerapkan teknik *bagging* pada pohon klasifikasi dengan langkah-langkah sebagai berikut.
   - Membagi data menjadi dua bagian yaitu data *learning* dan data *testing*.
   - Membentuk pohon klasifikasi dengan data *learning* dengan penyeleksian menggunakan *test sample estimate*. Jalankan data pada pohon klasifikasi tersebut dan akan menghasilkan tingkat kesalahan klasifikasi (*misclassification rate*) $e_S(L,T)$.
   - Sampel *bootstrap* L_B diambil dari data *learning* dan pohon klasifikasi dibangun dari L_B dengan penyeleksian menggunakan *test sample estimate*. Ulangi langkah ini sebanyak *M* kali sehingga diperoleh pohon klasifikasi $\varphi_1(x), \varphi_2(x), \dots, \varphi_B(x)$. Mengadopsi rekomendasi dari para peneliti

terdahulu maka dalam penelitian ini replikasi sampel *bootstrap* yang dibuat adalah 25, 50, 75, 100, 150 dan 200 kali.

- Menjalankan data *testing* pada pohon klasifikasi yang terbentuk dari masing-masing sampel *bootstrap*. Kemudian hitung *bagging misclassification rate* $e_B(L_B, T)$.


M U L A I
Membahas Konsep *splitting rule* pada pohon klasifikasi
Data pekerja anak
Data *testing*
Data *learning*
Membentuk pohon klasifikasi dengan menggunakan indeks *Gini* dan *Twoing*
Memangkas pohon klasifikasi
Memilih pohon klasifikasi optimal
Interpretasi model pohon klasifikasi
Mendapatkan ketepatan klasifikasi
Menerapkan teknik *bagging* pada pohon klasifikasi
Melihat pengaruh penerapan *bagging* pada pohon klasifikasi
Kesimpulan
SELESAI

(*Halaman Ini Sengaja Dikosongkan*)

BAB 4
HASIL DAN PEMBAHASAN

4.1. Konsep *Splitting Rule*

Pada bagian ini diberikan beberapa definisi dan ulasan yang membahas tentang konsep dan dasar pembentukan pohon klasifikasi, khususnya dalam proses *splitting rule* dan ilustrasi pembentukan pohon klasifikasi tersebut.

4.1.1. Probabilitas Dalam Pohon Klasifikasi

Dalam sampel *learning L* dengan banyaknya kelas adalah *j*, diberikan probabilitas *prior*

$$(\pi_j) = \frac{N_j}{N} \tag{4.1}$$

dengan

N = banyaknya objek pada sampel *learning L*

N_j = banyaknya objek pada kelas *j*

Probabilitas *prior* merupakan informasi awal mengenai proporsi atau perbandingan banyaknya objek pada tiap-tiap kelas dalam *L*. nilai probabilitas *prior* ini diestimasi dari proporsi $\frac{N_j}{N}$ yang diperoleh dari data. Ada dua jenis dari probabilitas *prior* dalam pohon klasifikasi yaitu

- *Priors data*, mengasumsikan bahwa proporsi banyaknya objek dalam suatu kelas yang terdapat dalam sampel sama dengan yang terdapat dalam populasinya. *Prior data* diestimasi oleh $(\pi_j) = \frac{N_j}{N}$.
- *Priors equal*, mengasumsikan bahwa proporsi banyaknya objek tiap-tiap kelas adalah sama. Diestimasikan P(kelas 1) = P(kelas 2) = $\frac{1}{2}$.

Dalam sebuah simpul *t*, diberikan:

N_t = banyaknya objek dalam *L* yang mana $x_0 \in t$ (banyaknya objek dalam simpul *t*)

$N_j(t)$ = banyaknya objek kelas *j* yang berada dalam simpul *t*

$\frac{N_j(t)}{N_j}$ = proporsi objek-objek dalam kelas j yang berada di simpul *t*

$p(j,t)$ = probabilitas sebuah objek adalah anggota kelas *j* dan berada dalam simpul *t*

Sehingga

$$p(j,t) = \pi_j . \frac{N_j(t)}{N_j} = \frac{N_j}{N} . \frac{N_j(t)}{N_j} = \frac{N_j(t)}{N} \quad (4.2)$$

Jika $P(t)$ adalah probabilitas beberapa objek akan berada dalam simpul *t*, maka diperoleh

$$P(t) = \sum_j^J P(j,t)$$
$$= P(1,t) + P(2,t) + \cdots + P(J,t)$$
$$= \frac{N_1(t)}{N} + \frac{N_2(t)}{N} + \cdots + \frac{N_J(t)}{N}$$
$$P(t) = \frac{N_t}{N} \quad (4.3)$$

Jika $P(j|t)$ adalah probabilitas sebuah objek adalah anggota kelas *j* yang berada dalam simpul *t*, sehingga diperoleh

$$P(j|t) = \frac{P(j,t)}{P(t)} = \frac{\frac{N_j(t)}{N}}{\frac{N_t}{N}} = \frac{N_j(t)}{N_t} \quad (4.4)$$

dan $\sum_j P(j|t) = 1$.

4.1.2. *Splitting Rule*

Menurut Lewis, pada dasarnya proses pembentukan dan pembuatan pohon klasifikasi, terdiri dari proses *splitting nodes* yaitu proses pemecahan simpul induk menjadi dua buah simpul anak melalui aturan pemecahan tertentu dan dilakukan secara berulang-ulang serta pelabelan kelas tertentu melalui aturan pengidentifikasian. Proses pemecahan pada masing-masing simpul induk didasarkan pada kriteria *goodness of split*. *Goodness of split* ini dibentuk berdasarkan fungsi keragaman (*impurity function)*.

Definisi 4.1. (Breiman *et al*, 1984) Fungsi *impurity* adalah sebuah fungsi ϕ yang didefinisikan oleh $(P_1, P_2, \dots, P_j)$; $P_j \geq 0$ dan $\sum_j P_j = 1, j = 1,2, \dots, J$.

Fungsi *impurity* ϕ memenuhi kriteria:

(i) ϕ maksimum apabila nilai-nilai

$$(P_1, P_2, \dots, P_j) = \left(\frac{1}{j}, \frac{1}{j}, \dots, \frac{1}{j}\right) \quad (4.5)$$

(ii) ϕ minimum apabila nilai-nilai

$$(P_1, P_2, \dots, P_j) = (1,0, \dots ,0)(0,1, \dots ,0), \dots, (0,0, \dots ,1) \quad (4.6)$$

(iii) ϕ adalah fungsi simetris dari $P_1, P_2, \dots, P_j$.

Definisi 4.2. (Breiman *et al*, 1984) Diberikan fungsi *impurity*, maka *impurity measure* $i(t)$ dari beberapa simpul *t* adalah

$$i(t) = \phi\big(P(1|t), P(2|t), \dots, P(j|t)\big) \quad (4.7)$$

Jika sebuah *split s* dalam simpul *t* dibagi ke dalam t_R dengan proporsi banyaknya objek yang masuk dalam t_R adalah P_R, dan t_L dengan proporsi banyaknya objek yang masuk dalam t_L adalah P_L, maka didefinisikan *decrease impurity* (pengurangan keberagaman)

$$\Delta i(s,t) = i(t) - P_R i(t_R) - P_L i(t_L) \quad (4.8)$$

Nilai $\Delta i(s,t)$ digunakan sebagai uji kriteria *goodness of split*. Suatu *split s* akan digunakan untuk memecah simpul *t* menjadi dua buah simpul yaitu simpul t_R dan t_L jika *s* memaksimalkan nilai

$$\Delta i(s^*,t) = \max_s \Delta i(s,t) \quad (4.9)$$

Berdasarkan persamaan (4.8), $\Delta i(s,t)$ akan maksimum apabila diperoleh $P_R i(t_R)$ dan $P_L i(t_L)$ minimum. Hal ini berarti *splitting* dilakukan untuk membuat dua buah simpul baru yang keragamannya lebih kecil (homogen) apabila dibandingkan dengan simpul awalnya (simpul induk).

Misalkan sebuah pohon klasifikasi telah terbentuk dan memiliki sekumpulan atau himpunan simpul terminal $\tilde{T}$, dengan

$$I(t) = i(t)P(t) \quad (4.10)$$

didefinisikan pula *tree impurity* $I(T)$, dengan

$$I(T) = \sum_{t\in\tilde{T}} I(t) = \sum_{t\in\tilde{T}} i(t)P(t) \quad (4.11)$$

Pemilihan *split s* yang memaksimalkan $\Delta i(s,t)$ ekuivalen dengan pemilihan *split s* yang meminimalkan *tree impurity* $I(T)$. Ambil setiap simpul $t \in \tilde{T}$ dan gunakan

split s untuk membagi simpul menjadi t_L dan t_R. Pohon klasifikasi baru yang terbentuk T' memiliki keragaman

$$I(T') = \sum_{\tilde{T}-t} I(t) + I(t_L) + I(t_R) \quad (4.12)$$

dengan penurunan *tree impurity*

$$I(T) - I(T') = I(t) - I(t_L) - I(t_R) \quad (4.13)$$

Hal ini hanya bergantung pada pemilihan *split s* dan simpul *t*. Sehingga, memaksimalkan penurunan *tree impurity* dengan split pada *t* ekuivalen dengan memaksimalkan

$$\Delta i(s,t) = I(t) - I(t_R) - I(t_L) \quad (4.14)$$

Selanjutnya, kriteria *goodness of split* akan memilih *split* dengan nilai *resubtitution estimate* dari *misclassification cost* yang paling kecil. Kriteria ini digeneralisasi menjadi kriteria keberagaman *Gini* dan *Twoing* yang lebih dikenal dengan sebutan indeks *Gini* dan *Twoing*.

a. Indeks *Gini*

Konsep dari indeks *Gini* berawal dari *impurity measure* sesuai dengan yang tertera pada persamaan (4.7).

Definisi 4.3. (Breiman *et al*, 1984) Diberikan *impurity measure* $i(t)$, maka *Gini Diversity Index* (Indeks Keragaman *Gini*) adalah

$$i(t) = \sum_{j \neq i} p(j|t)p(i|t) \quad (4.15)$$

Dalam sebuah simpul *t*, andaikan terdapat $1,2,\ldots,j$ kelas. Untuk $i = 1$ dan *j* adalah kelas-kelas lainnya maka persamaan di atas dapat dituliskan

$$\begin{aligned}\sum_{j \neq i} p(j|t)p(i|t) &= P(1|t)P(2|t) + P(1|t)P(3|t) + \cdots + P(1|t)P(j|t) \\ &= P(1|t)[P(2|t) + P(3|t) + \cdots + P(j|t)] \quad (4.16)\end{aligned}$$

Karena $\sum_j p(j|t) = 1$, sehingga persamaan di atas menjadi

$$\begin{aligned}P(1|t)[P(2|t) + P(3|t) + \cdots + P(j|t)] &= P(1|t)\left(\sum_j P(j|t) - P(1|t)\right) \\ &= P(1|t)\left(1 - P(1|t)\right) \\ &= P(1|t) - P^2(1|t) \quad (4.17)\end{aligned}$$

Begitu pula untuk $i = 2$ dan *j* adalah kelas-kelas lainnya, maka indeks keragaman *Gini* dapat dituliskan

$$\sum_{j \neq i} p(j|t)p(i|t) = \sum_{j=1, j \neq i}^{2} P(j|t) - P^2(j|t) \quad (4.18)$$

Untuk $i = 3$ dan j adalah kelas-kelas lainnya, maka indeks keragaman *Gini* dapat dituliskan

$$\sum_{j\neq i} p(j|t)p(i|t) = \sum_{j=1,j\neq i}^{3} P(j|t) - P^2(j|t) \quad (4.19)$$

Sehingga secara umum, siperoleh

$$\begin{aligned}\sum_{j\neq i} p(j|t)p(i|t) &= \sum_j P(j|t) - P^2(j|t)\\ &= \sum_j P(j|t) - \sum_j P^2(j|t)\\ &= 1 - \sum_j P^2(j|t)\end{aligned} \quad (4.20)$$

Sehingga, indeks keragaman *Gini* (*Gini diversity Index*) dapat dituliskan

$$i(t) = 1 - \sum_j P^2(j|t) \quad (4.21)$$

b. Indeks *Twoing*

Ide dari indeks *Twoing* adalah membagi kelas menjadi dua kelas besar. Misalkan kelas dalam *C* dibagi menjadi dua kelas besar $C_1 = (j_1, j_2, \dots, j_n)$, $C_2 = C - C_1$.

Diketahui dari penjelasan sebelumnya bahwa akan dipilih *split s* pada simpul yang memaksimumkan $\Delta i(s,t)$, sehingga pada indeks ini akan dipilih *split s* dengan kelas besar C_1 yang memaksimumkan $\Delta i(s,t,C_1)$.

$$C_1(s) = \{j: P(j|t_L) \geq P(j|t_R)\} \quad (4.22)$$

dan

$$\max_{C_1} \Delta i(s,t,C_1) = \frac{P_L P_R}{4}\left[\sum_j |P(j|t_L) - P(j|t_R)|\right]^2 \quad (4.23)$$

Corrolary 4.1. Untuk setiap simpul *t* dan *split s* dari *t* dalam t_L dan t_R diberikan Indeks Keragaman *Twoing*

$$\phi(s,t) = \frac{P_L P_R}{4}\left[\sum_j |P(j|t_L) - P(j|t_R)|\right]^2 \quad (4.24)$$

Split Twoing terbaik $s^*(C_1^*)$ diberikan oleh *split* s^* yang memaksimumkan $\phi(s,t)$ dengan C_1^* sebagai berikut

$$C_1^* = \{j: P(j|t_L^*) \geq P(j|t_R^*)\} \quad (4.25)$$

dimana t_L^* dan t_R^* adalah simpul yang dihasilkan dari *split* s^*.

Bukti:

Dari persamaan (4.8)

$$\Delta i(s,t) = i(t) - P_L i(t_L) - P_R i(t_R)$$

Karena indeks Twoing membagi kelas menjadi dua kelas besar, maka

$$\Delta i(s,t) = P(1|t)P(2|t) - P_L P(1|t_L)P(2|t_L) - P_R P(1|t_R)P(2|t_R) \quad (4.26)$$

Dengan $P(2|.) = 1 - P(1|.)$, maka

$$\Delta i(s,t) = P(1|t) - P(1|t_L)P_L - P(1|t_R)P_R + P^2(1|t_L)P_L + P^2(1|t_R)P_R - P^2(1|t)$$

Dengan mengganti $P(1|t)$ dengan $P_L P(1|t_L) + P_R P(1|t_R)$ menghasilkan

$$\begin{aligned}\Delta i(s,t) &= P^2(1|t_L)P_L + P^2(1|t_R)P_R - \left(P_L P(1|t_L) + P_R P(1|t_R)\right)^2 \\ &= [P^2(1|t_L)P_L + P^2(1|t_R)P_R - P^2(1|t_L)P_L^2] - \\ &\quad [P^2(1|t_L)P_R^2 - 2P_L P_R P(1|t_L)P(2|t_L)] \\ &= [P^2(1|t_L)P_L - P^2(1|t_L)P_L^2 + P^2(1|t_R)P_R] - \\ &\quad [P^2(1|t_L)P_R^2 - 2P_L P_R P(1|t_L)P(2|t_L)] \\ &= [P^2(1|t_L)P_L(1-P_L) + P^2(1|t_L)P_R(1-P_R)] - \\ &\quad [P^2(1|t_L)P_R^2 - 2P_L P_R P(1|t_L)P(2|t_L)] \\ &= P_L P_R P^2(1|t_L) + P_L P_R P^2(1|t_R) - 2P_L P_R P(1|t_L)P(2|t_L) \\ &= P_L P_R\left(P(1|t_L) - P(1|t_R)\right)^2 \end{aligned} \quad (4.27)$$

Sehingga

$$\Delta i(s,t,C_1) = P_L P_R\left(P(C_1|t_L) - P(C_1|t_R)\right)^2 \quad (4.28)$$

Kemudian definisikan $z_j = P(j|t_L) - P(j|t_R)$, maka $\sum_j z_j = 0$. Untuk setiap bilangan real *z* misalkan z^+ dan z^- menjadi bagian positif dan negatif: $z^+ = z$ dan $z^- = 0$ jika ≥ 0 ; $z^+ = 0$ dan $z^- = -z$ jika $z \leq 0$. Dan $z = z^+ - z^-$, $|z| = z^+ + z^-$.

Dari persamaan (4.28), jika P_L, P_R hanya bergantung pada s dan tidak pada C_1, maka $C_1(s)$ akan memaksimumkan atau meminimumkan $s = \sum_{j \in C_1} z_j$. Nilai maksimum akan diperoleh dengan $C_1 = \{j; z_j \geq 0\}$ dan sama dengan $\sum_j {z_j}^+$. Nilai minimum akan diperoleh dengan $C_1 = \{j; z_j < 0\}$ dan sama dengan $-\sum_j {z_j}^-$.

Sehingga

$$\sum_j {z_j}^+ - \sum_j {z_j}^- = \sum_j z_j = 0 \quad (4.29)$$

Dapat dilihat bahwa nilai maksimum *s* sama dengan nilai absolut dari nilai minimum

$$\frac{(\sum_j z_j{}^{+} + \sum_j z_j{}^{-})}{2} = \frac{(\sum_j |z_j|)}{2} \quad (4.30)$$

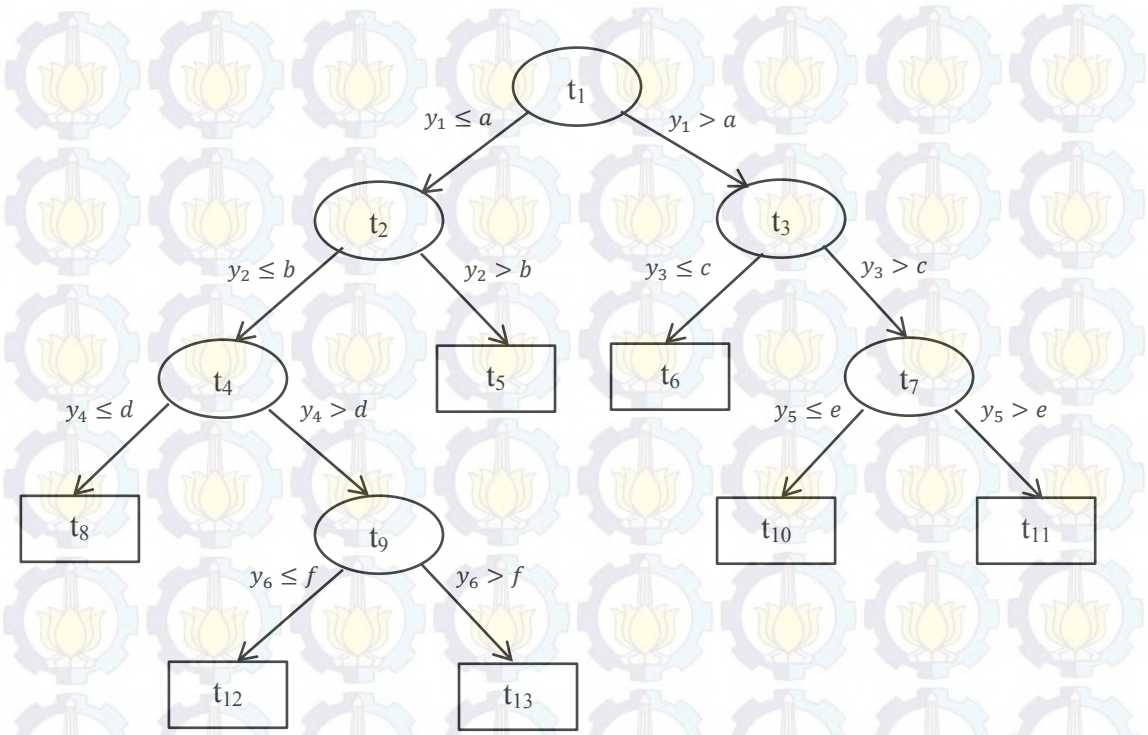
Kemudian menggunakan persamaan (4.28)

$$\max_{C_1} \Delta i(s,t,C_1) = \frac{P_L P_R}{4} \left(\left| \sum_j P(j|t_L) - P(j|t_R) \right| \right)^2 \quad (4.31)$$

4.1.3. Ilustrasi Pembentukan Pohon Klasifikasi

Teknik atau proses kerja dalam membuat sebuah pohon klasifikasi dikenal dengan istilah *binary recursive partitioning*. Proses ini disebut *binary* karena setiap simpul induk akan selalu mengalami pemecahan ke dalam tepat dua simpul anak. Sedangkan *recursive* berarti bahwa proses pemecahan tersebut akan diulang kembali pada setiap simpul anak hasil pemecahan terdahulu, sehingga simpul anak tersebut bisa jadi juga akan menjadi simpul induk. Proses pemecahan ini akan terus dilakukan sampai tidak ada lagi kesempatan untuk melakukan pemecahan berikutnya. Istilah *partitioning* sendiri merujuk pada pemecahan ke dalam bagian-bagian yang lebih kecil.

Kriteria pemecahan didasarkan pada nilai-nilai dari variabel prediktor yang dimiliki. Misalkan dimiliki variabel respon *y* yang bertipe kategorik dan variabel-variabel prediktor *x* ke dalam partisi-partisi yang berbentuk persegi panjang dan tidak saling tumpang tindih. Idenya adalah membagi ruang berdimensi *p* dari variabel-variabel prediktor tadi ke dalam beberapa partisi yang mana masing-masing partisi berisi objek-objek yang homogen dan seragam. Homogen dalam kasus ini mengindikasikan bahwa objek-objek tersebut merupakan anggota satu kelas yang sama. Proses *splitting* akan berlanjut sampai didapatkan pohon klasifikasi maksimal.



Gambar 4.1. Proses Partisi

Untuk memperjelas proses partisi, akan diberikan contoh pemilahan pada ilustrasi sebelumnya. Dengan menggunakan persamaan 4.21 untuk indeks *Gini* dan 4.24 untuk indeks *Twoing*, maka diperoleh *split* (pemilah) yang memilah tiap simpul berturut adalah y_1 dengan *split a*, y_2 dengan *split b*, y_3 dengan *split c*, y_4 dengan *split d*, y_5 dengan *split e* dan y_6 dengan *split f*. Secara rinci, proses partisi simpul t_1 dipilah dengan kriteria pemecahan $y_1 \leq a$ dan $y_1 > a$. Pemecahan yang dihasilkan adalah simpul t_2 akibat dari kriteria $y_1 \leq a$ dan simpul t_3 terbentuk akibat kriteria pemecahan $y_1 > a$. Kemudian proses partisi berlanjut pada simpul t_2, dengan kriteria pemecahan $y_2 \leq b$ dan $y_2 > b$. Simpul t_4 terbentuk karena memenuhi kriteria $y_2 \leq b$ dan simpul t_5 terbentuk karena kriteria $y_2 > b$. Begitu seterusnya sampai dengan simpul t_9 dipilah dengan kriteria pemecahan $y_6 \leq f$ dan $y_6 > f$.

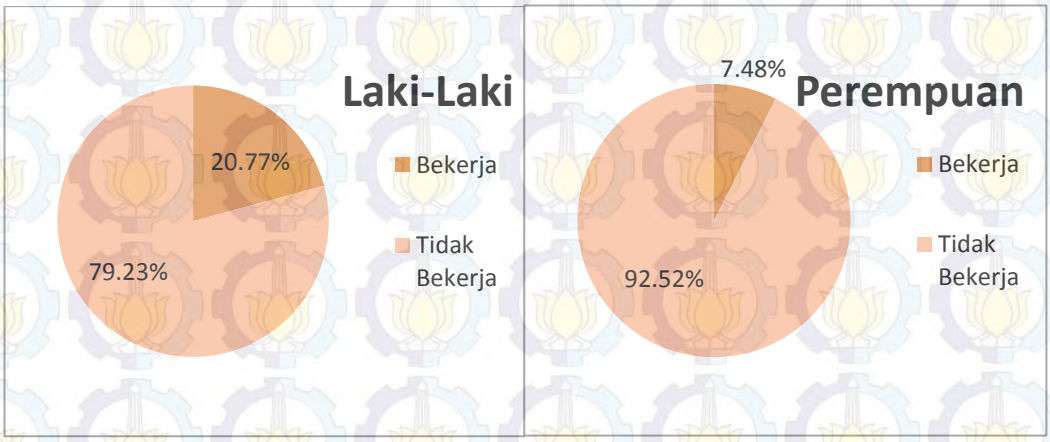
Dari model pohon klasifikasi, variabel yang membentuk model pohon klasifikasi dapat diketahui serta kontribusinya terhadap model yang terbentuk. Kontribusi variabel dalam model dapat dilihat berdasarkan ukuran *variable importance score*. Ukuran ini dihitung berdasarkan penjumlahan dari nilai-nilai

pemilah (*split*) yang mungkin untuk tiap variabel pada tiap simpul. Variabel yang memiliki jumlah nilai *split* terbesar akan menjadi pemilah utama dengan skor 100, sedangkan yang lainnya akan bernilai di bawah 100.

Dari ilustrasi sebelumnya, dengan menggunakan persamaan 2.10, y_1 merupakan pemilah utama dalam model dan variabel dengan skor 100. Artinya, variabel y_1 memiliki jumlah nilai-nilai pemilah (termasuk a) yang paling besar dibandingkan dengan variabel lainnya. Sedangkan untuk variabel lainnya, yaitu y_2, y_3, y_4, y_5 dan y_6 memiliki skor dengan rentang 0-100

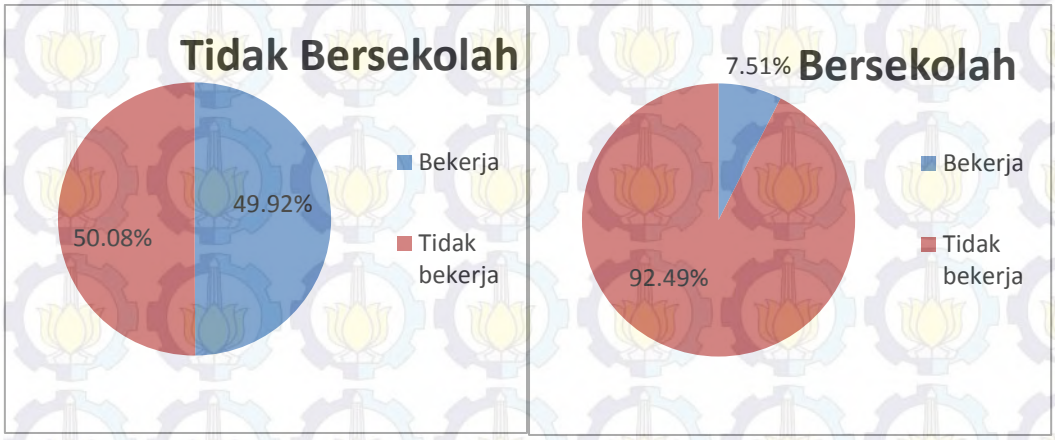
4.2. Analisis Jumlah Kasus Pekerja Anak di Sulawesi Tengah

Berdasarkan data Susenas provinsi Sulawesi Tengah tahun 2013, banyaknya sampel anak yang berusia 10-18 tahun berjumlah 3780 orang. Banyaknya anak yang masuk ke dalam dunia kerja sebagai pekerja dari jumlah tersebut mencapai 554 orang atau 14,66% dari jumlah keseluruhan. Jika ditinjau menurut jenis kelamin, dari 2041 orang anak usia 10-18 tahun yang berjenis kelamin laki-laki, terdapat 424 orang anak yang bekerja atau 20,77% dari total seluruhnya. Sedangkan untuk yang berjenis kelamin perempuan, terdapat 7,48% dari 1739 orang anak yang masuk ke dalam dunia kerja, atau sebanyak 130 orang anak.



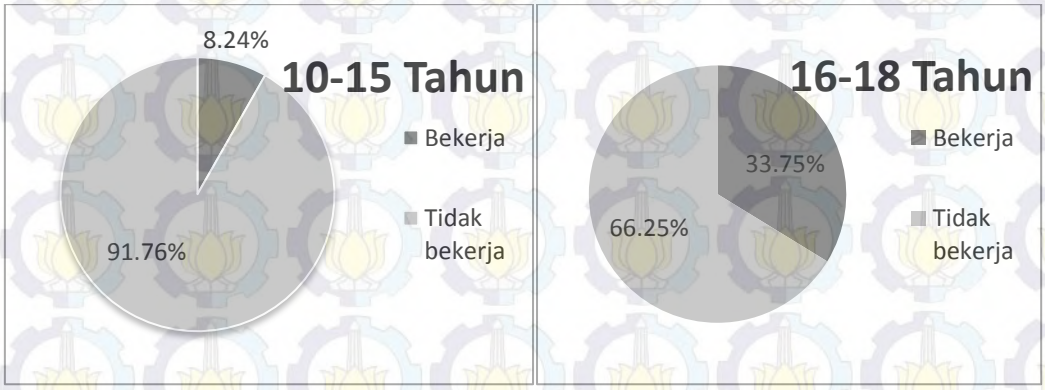
Gambar 4.2. Persentase Pekerja Anak Menurut Jenis Kelamin Anak

Berdasarkan partisipasi sekolah anak, dari total 3780 orang anak, ada 637 orang anak yang tidak bersekolah dan 3143 orang anak yang bersekolah. Dari 637 anak yang tidak bersekolah, terdapat 318 orang anak yang bekerja dan 319 orang anak yang tidak bekerja. Sedangkan untuk 3143 orang anak yang bersekolah terdapat 7,51% anak yang bekerja atau mencapai 236 orang anak. Hal ini menunjukkan bahwa proporsi anak yang bekerja pada kelompok anak yang tidak bersekolah jauh lebih besar daripada pada kelompok anak yang bersekolah.



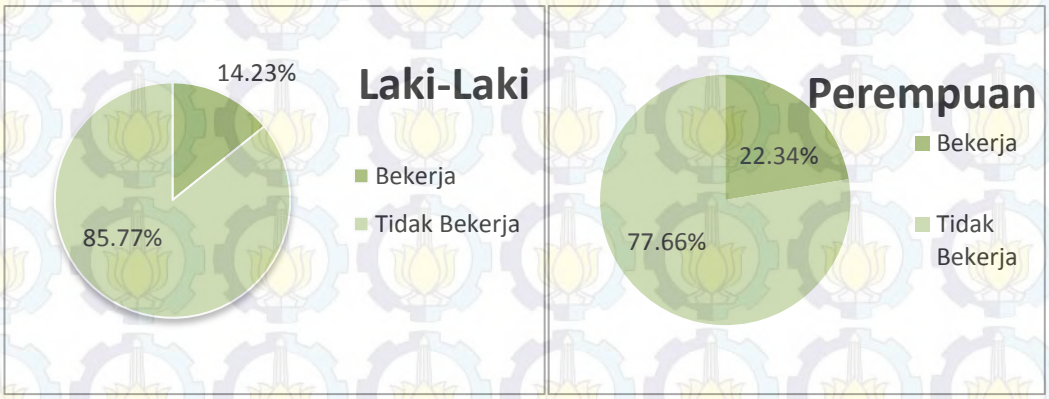
Gambar 4.3. Persentase Anak Bekerja Menurut Partisipasi Sekolah Anak

Jika ditelusuri berdasarkan kelompok umur, jumlah pekerja anak pada kelompok umur 10-15 tahun mencapai 8,24% dari total 2829 atau mencapai angka 233 orang. Sementara itu untuk kelompok umur yang lebih tinggi yaitu 16-18 tahun, jumlah anak yang bekerja sebanyak 321 dari 951 orang, atau sebanyak 33,75%. Jadi hal ini menunjukkan bahwa pekerja anak dengan kelompok umur 16-18 tahun lebih banyak dibandingkan dengan kelompok umur 10-15 tahun.



Gambar 4.4. Persentase Pekerja Anak Menurut Kelompok Umur Anak

Jumlah pekerja anak juga bisa ditinjau dari segi jenis kelamin kepala rumah tangga. Kepala rumah tangga dengan jenis kelamin laki-laki berjumlah 3583 dan perempuan berjumlah 197 kepala rumah tangga. Untuk kepala rumah tangga yang berjenis kelamin laki-laki, ada 510 anak yang masuk dalam dunia kerja atau sekitar 14,23%. Untuk kepala rumah tangga yang berjenis kelamin perempuan sendiri, terdapat 22,34% dari 197 orang anak yang masuk dalam dunia kerja atau sebanyak 44 orang anak. Dapat dilihat dari gambar berikut bahwa kepala rumah tangga dengan jenis kelamin perempuan memiliki persentase pekerja anak yang lebih besar dibandingkan dengan rumah tangga yang dikepalai oleh laki-laki.



Gambar 4.5. Persentase Pekerja Anak Berdasarkan Jenis Kelamin Kepala Rumah Tangga

Jika ditelusuri berdasarkan tingkat pendidikan kepala rumah tangga, dari 3780 orang anak, 1583 diantaranya memiliki kepala rumah tangga yang berpendidikan SD, 644 berpendidikan SMP, 678 berpendidikan SMA dan 875 tidak memiliki ijazah. Dari jumlah tersebut, 875 kepala rumah tangga yang tidak memiliki ijazah memiliki jumlah anak yang bekerja mencapai 21,37% atau sebanyak 187 orang anak. Sementara 1583 kepala rumah tangga yang berpendidikan SD memiliki 13,90% anak yang bekerja. Sedangkan untuk kepala rumah tangga dengan pendidikan SMP dan SMA masing-masing terdapat 13,35% dan 9% jumlah anak bekerja dari total keseluruhannya.

Tabel 4.1. Proporsi Pekerja Anak Menurut Tingkat Pendidikan Kepala Rumah Tangga

Tingkat Pendidikan Kepala Rumah Tangga	Pekerja Anak (%)		Total
	Bekerja	Tidak Bekerja	
Tidak Berijazah	21,37	78,63	875
SD	13,90	86,10	1583
SMP	13,35	86,65	644
SMA	9,00	91,00	678
Jumlah	14,66	85,34	3780

Jika ditelaah menurut sektor pekerjaan kepala rumah tangga, ditemukan 1963 orang anak dengan kepala rumah tangga berprofesi di sektor pertanian, 186 orang anak dengan kepala rumah tangga bekerja di sektor industri, 651 orang anak dengan kepala rumah tangga berprofesi di sektor perdagangan, 814 orang anak memiliki kepala rumah tangga yang bekerja di sektor jasa dan 166 orang anak dengan kepala rumah tangga yang menggeluti sektor pekerjaan lainnya. Untuk kepala rumah tangga yang bekerja di sektor pertanian, terdapat 17,07% anak yang bekerja dari jumlah keseluruhan, sedangkan untuk sektor industri, perdagangan, jasa dan lainnya berturut-turut adalah 17,20%, 13,52%, 8,48% dan 18,07% anak yang bekerja dari total keseluruhan.

Tabel 4.2. Proporsi Pekerja Anak Menurut Sektor Pekerjaan Kepala Rumah Tangga

Sektor Pekerjaan Kepala Rumah Tangga	Pekerja Anak (%)		Total
	Bekerja	Tidak Bekerja	
Pertanian	17,07	82,93	1963
Industri	17,20	82,80	186
Perdagangan	13,52	86,48	651
Jasa	8,48	91,52	814
Lainnya	18,07	81,93	166
Jumlah	14,66	85,34	3780

Kelompok umur kepala rumah tangga juga dapat digunakan sebagai acuan dalam mendeskripsikan jumlah pekerja anak. Untuk anak dengan kepala rumah tangga yang berumur ≤ 30 tahun berjumlah 55 orang dengan 7 diantaranya masuk sebagai pekerja anak atau setara dengan 12,73%. Untuk anak dengan kepala rumah tangga berumur 31-55 tahun berjumlah 3376 dengan 14,04% masuk ke dalam dunia kerja atau sebanyak 474 orang anak. Sedangkan untuk anak yang memiliki kepala rumah tangga berumur > 55 tahun berjumlah 349 orang anak dengan 73 diantaranya masuk ke dalam dunia kerja atau setara dengan 20,92%.

Tabel 4.3. Proporsi Pekerja Anak Menurut Kelompok Umur Kepala Rumah Tangga

Kelompok Umur Kepala Rumah Tangga	Pekerja Anak (%)		Total
	Bekerja	Tidak Bekerja	
≤ 30 tahun	12,73	87,27	55
31-55 tahun	14,04	85,96	3376
> 55 tahun	20,92	79,08	349
Jumlah	15,90	84,10	3780

Jika dilihat dari pendapatan perkapita rumah tangga, anak dengan pendapatan perkapita rumah tangga < Rp. 250000 berjumlah 2026 orang anak dengan 322 diantaranya masuk sebagai pekerja anak. Sedangkan untuk anak dengan pendapatan perkapita rumah tangga sebesar Rp. 250000 – Rp. 500000, Rp. 500001 – Rp. 1 juta dan > 1 juta masing-masing berjumlah 919, 513 dan 322 orang anak dengan anak yang masuk ke dalam dunia kerja berturut-turut adalah 124, 75 dan 33 orang anak.

Tabel 4.4. Proporsi Pekerja Anak Menurut Pendapatan Perkapita Rumah Tangga

Pendapatan Perkapita Rumah Tangga	Pekerja Anak (%)		Total
	Bekerja	Tidak Bekerja	
< Rp. 250000	15,89	84,11	2026
Rp. 250000 – Rp. 500000	13,49	86,51	919
Rp. 500001 – Rp. 1 juta	14,62	85,38	513
> 1 juta	10,25	89,75	322
Jumlah	14,25	85,75	3780

Jika dilihat dari jumlah anggota rumah tangga, dari 3780 orang anak terdapat 2381 orang anak dengan jumlah anggota rumah tangga sebanyak 2-5 orang, 1327 orang anak dengan jumlah anggota rumah tangga sebanyak 6-9 orang dan 72 orang anak dengan jumlah anggota rumah tangga sebanyak 10-14 orang. Dari 2381 orang anak tadi, 14,32% diantaranya atau sebanyak 341 anak masuk dalam dunia kerja, dari 1327 anak, terdapat 199 anak yang bekerja dan dari 72 anak terdapat 14 orang anak yang bekerja. Berdasarkan hasil ini dapat dilihat bahwa semakin banyak anggota rumah tangga maka semakin banyak pula persentase pekerja anak.

Tabel 4.5. Proporsi Pekerja Anak Menurut Jumlah Anggota Rumah Tangga

Jumlah Anggota Rumah Tangga	Pekerja Anak (%)		Total
	Bekerja	Tidak Bekerja	
2-5 orang	14,32	85,68	2381
6-9 orang	15,00	85,00	1327
10-14 orang	19,44	80,56	72
Jumlah	16,25	83,75	3780

4.3. Hubungan Antara Variabel Prediktor Dengan Variabel Respon

Pada bagian sebelumnya telah dipaparkan mengenai profil pekerja anak melalui analisis jumlah pekerja anak berdasarkan variabel prediktor yang mempengaruhinya. Untuk memperkuat dan memperjelas hubungan antara

variabel respon dengan variabel prediktor, maka perlu dilakukan pengujian. Adapun pengujian yang digunakan adalah tabulasi silang untuk variabel respon dengan variabel prediktor yang bersifat kategorik dan uji beda rata-rata untuk variabel prediktor yang berskala interval.

Selanjutnya, untuk mengetahui hubungan jenis kelamin anak, partisipasi sekolah anak, tingkat pendidikan kepala rumah tangga, jenis kelamin kepala rumah tangga dan sektor pekerjaan kepala rumah tangga terhadap munculnya kasus pekerja anak dilakukan pengujian tabulasi silang satu per satu. Masing-masing pasangan variabel respon dan prediktor memiliki hipotesis sebagai berikut

H_0 : Tidak ada hubungan antara variabel respon dengan variabel prediktor

H_1 : Ada hubungan antara variabel respon dengan variabel prediktor

Tabel 4.6. Nilai χ^2 dan *p-value* Dari Uji Tabulasi Silang

Variabel Prediktor	χ^2 Hitung	*p-value*	Kesimpulan
Jenis Kelamin Anak	132,761	0,000	Tolak H_0
Partisipasi Sekolah Anak	761,718	0,000	Tolak H_0
Tingkat Pendidikan Kepala Rumah Tangga	50,507	0,000	Tolak H_0
Jenis Kelamin Kepala Rumah Tangga	9,798	0,002	Tolak H_0
Sektor Pekerjaan Kepala Rumah Tangga	37,152	0,000	Tolak H_0

Dapat dilihat bahwa hasil yang tertera pada tabel di atas menghasilkan kesimpulan untuk menolak H_0 untuk setiap variabel. Hal ini mengindikasikan hubungan yang signifikan antara munculnya kasus pekerja anak dengan jenis kelamin anak, partisipasi sekolah anak, tingkat pendidikan kepala rumah tangga, jenis kelamin kepala rumah tangga dan sektor pekerjaan kepala rumah tangga.

Selanjutnya untuk melihat hubungan antara variabel respon dengan variabel prediktor yang berskala interval digunakan uji beda rata-rata. Hipotesis dalam pengujian ini adalah

H_0 : $\mu_1 = \mu_2$

Misalkan untuk variabel umur anak, maka hipotesis tersebut berarti rata-rata umur anak yang bekerja dengan anak yang tidak bekerja adalah sama. Begitu pula untuk variabel yang lain.

H_1 : $\mu_1 \neq \mu_2$

Misalkan untuk variabel umur anak, maka hipotesis tersebut berarti rata-rata umur anak yang bekerja dengan anak yang tidak bekerja adalah tidak sama. Begitu pula untuk variabel yang lain.

Tabel 4.7. Uji Beda Rata-Rata

Variabel Prediktor	F Hitung	*p-value*	t Hitung	*p-value*
Umur Anak	7,696	0,006	-21,579	0,000
Umur Kepala Rumah Tangga	1,256	0,263	-5,281	0,000
Pendapatan Perkapita Rumah Tangga	9,309	0,002	2,344	0,019
Jumlah Anggota Rumah Tangga	5,836	0,016	-0,985	0,325

Berdasarkan hasil tabel di atas dapat dilihat bahwa variabel umur anak, umur kepala rumah tangga dan pendapatan perkapita rumah tangga menghasilkan kesimpulan untuk menolak H_0 pada nilai t hitung yang berarti bahwa ketiga variabel tersebut tidak sama nilainya pada anak yang bekerja dengan anak yang tidak bekerja. Sedangkan untuk variabel jumlah anggota rumah tangga menghasilkan kesimpulan yang gagal menolak H_0, artinya jumlah anggota rumah tangga pada anak yang bekerja dengan anak yang tidak bekerja adalah sama. Sedangkan untuk homogenitas data pada nilai F hitung dapat dilihat bahwa hanya variabel umur kepala rumah tangga yang menghasilkan kesimpulan bahwa data pada variabel tersebut homogen.

4.4. Model Pohon Klasifikasi

Untuk membentuk suatu pohon klasifikasi, terlebih dahulu ditentukan pembagian besarnya jumlah data *learning* dan data *testing*. Breiman *et al* tidak menentukan secara khusus berapa persen porsi pembagian suatu data menjadi data *learning* dan data *testing*, namun data *learning* harus lebih besar dibandingkan dengan data *testing*, sehingga dalam penelitian ini pembentukan pohon klasifikasi dilakukan dengan membagi data total menjadi dua bagian, yaitu data *learning* sebesar 80% dan data *testing* sebesar 20%.

Selanjutnya, akan dibentuk suatu pohon klasifikasi dengan menggunakan data *learning* yang telah ditentukan tadi. Pembentukan pohon klasifikasi tersebut menggunakan dua kriteria pemilahan (*splitting rule*), yaitu dengan menggunakan kriteria indeks *Gini* dan indeks *Twoing*. Hasil dari pengembangan pohon klasifikasi tersebut adalah sebagai berikut.

4.4.1. Pohon Maksimal

Pohon maksimal adalah pohon klasifikasi dengan jumlah simpul terminal terbanyak. Pohon ini merupakan hasil pemilahan simpul utama maupun pemilahan simpul internal. Pohon klasifikasi maksimal yang terbentuk dengan menggunakan indeks *Gini* mempunyai 239 simpul terminal dengan pemilahan pertama terjadi pada variabel umur anak, sedangkan dengan menggunakan indeks *Twoing*, pohon maksimal yang terbentuk mempunyai 236 simpul terminal dengan pemilahan pertama terjadi pada variabel umur anak. Jadi umur anak merupakan variabel yang memiliki peranan utama terhadap pembentukan pohon klasifikasi dan merupakan variabel yang sangat dominan terhadap pengklasifikasian pekerja anak.

Tabel 4.8. Pembentukan Pohon Maksimal

Indeks	*Terminal Nodes*	*Test Set Relative Cost*	*Resubtitution Relative Cost*	*Complexity Parameter*
Gini	239	$0{,}537 \pm 0{,}045$	0,130	0,000
Twoing	236	$0{,}540 \pm 0{,}045$	0,128	0,000

Semua variabel prediktor masuk ke dalam model pohon maksimal yang terbentuk, baik yang dengan menggunakan kriteria indeks *Gini* maupun indeks *Twoing*. Adapun sembilan variabel tersebut berturut-turut adalah umur anak (X_1), partispasi sekolah anak (X_3), umur kepala rumah tangga (X_4), pendapatan perkapita (X_8), jumlah anggota rumah tangga (X_9), tingkat pendidikan kepala rumah tangga (X_5), sektor pekerjaan kepala rumah tangga (X_7), jenis kelamin anak (X_2), jenis kelamin kepala rumah tangga (X_6).

Tabel 4.9. Variabel-Variabel Yang Masuk ke Dalam Pohon Maksimal

Nama Variabel	Deskripsi Variabel	Score	
		Gini	*Twoing*
X_1	Umur anak	100	100
X_3	partispasi sekolah anak	86,36	86,95
X_4	umur kepala rumah tangga	63,34	63,44
X_8	pendapatan perkapita	56,46	57,52
X_9	jumlah anggota rumah tangga	40,18	40,43
X_5	tingkat pendidikan kepala rumah tangga	31,31	30,94
X_7	sektor pekerjaan kepala rumah tangga	28,03	27,84
X_2	jenis kelamin anak	17,44	17,56
X_6	jenis kelamin kepala rumah tangga	6,53	6,57

Tabel di atas menunjukkan bahwa variabel prediktor yang menjadi pemilah utama pada simpul induk adalah variabel umur anak (X_1). Hal ini menunjukkan bahwa variabel umur anak (X_1) merupakan pemilah yang memiliki peranan utama terhadap pembentukan model dan merupakan variabel yang sangat dominan dalam pengklasifikasian pekerja anak. Kontribusi variabel ini ditunjukkan dengan urutannya dalam ranking variabel penting untuk pohon maksimal seperti yang tampak pada tabel di atas.

Untuk hasil pengujian pohon klasifikasi maksimal dengan menggunakan data *learning* dan data *testing* dapat dilihat dalam tabel berikut.

Tabel 4.10. Hasil Klasifikasi Pohon Maksimal Untuk Data *Learning*

Indeks	Observasi	Prediksi		*Sensitivity*	*Specitivity*	Total Ketepatan
		0	1			
Gini	0	2280	299	88,4%	98,6%	89,9%
	1	6	423			
Twoing	0	2278	301	80,3%	98,8%	89,8%
	1	5	424			

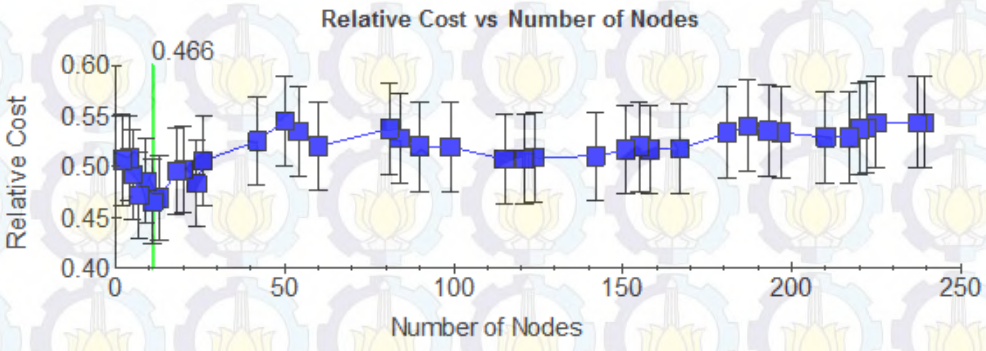
Tabel 4.11. Hasil Klasifikasi Pohon Maksimal Untuk Data *Testing*

Indeks	Observasi	Prediksi		*Sensitivity*	*Specitivity*	Total Ketepatan
		0	1			
Gini	0	522	125	80,7%	65,6%	78,2%
	1	43	82			
Twoing	0	520	127	80,4%	65,6%	78%
	1	43	82			

Seperti terlihat pada kedua tabel di atas, total ketepatan klasifikasi yang dihasilkan oleh pohon klasifikasi maksimal baik menggunakan kriteria indeks *Gini* maupun indeks *Twoing* tidak berbeda signifikan. Total ketepatan klasifikasi kedua indeks tersebut cukup berimbang, namun pohon klasifikasi maksimal dengan kriteria indeks *Gini* memiliki total ketepatan klasifikasi yang lebih tinggi.

4.4.2 Pemangkasan Pohon Maksimal

Meskipun pohon maksimal memiliki ketepatan klasifikasi yang tinggi, namun pohon maksimal memiliki stuktur yang sangat kompleks dengan simpul terminal sebanyak 239 dan 236 buah, sehingga diperlukan pemangkasan pohon klasifikasi secara iteratif menjadi bagian pohon yang makin kecil dan tersarang yang akan menghasilkan pohon optimal. Pemangkasan pohon klasifikasi dilakukan dengan menggunakan kaidah biaya pengganti relatif (*resubtitution relative cost*) terkecil yang dihitung dengan pendekatan penduga sampel uji (*test sample estimate*).



Gambar 4.6. Plot Biaya Relatif Dengan Jumlah Simpul Terminal

Perhitungan nilai biaya pengganti relatif (*resubtitution relative cost*) terkecil berdasarkan pendekatan penduga sampel uji (*test sample estimate*) menghasilkan biaya pengganti relatif minimum sebesar 0,466 yang ditandai oleh garis hijau (Gambar 4.1). Biaya pengujian relatif (*test set relative cost*) sebesar $0{,}466 \pm 0{,}042$ atau antara 0,424-0,508, biaya pengganti relatif sebesar 0,439 dan *complexity parameter* sebesar 0,002.

4.4.3. Pohon Klasifikasi Optimal

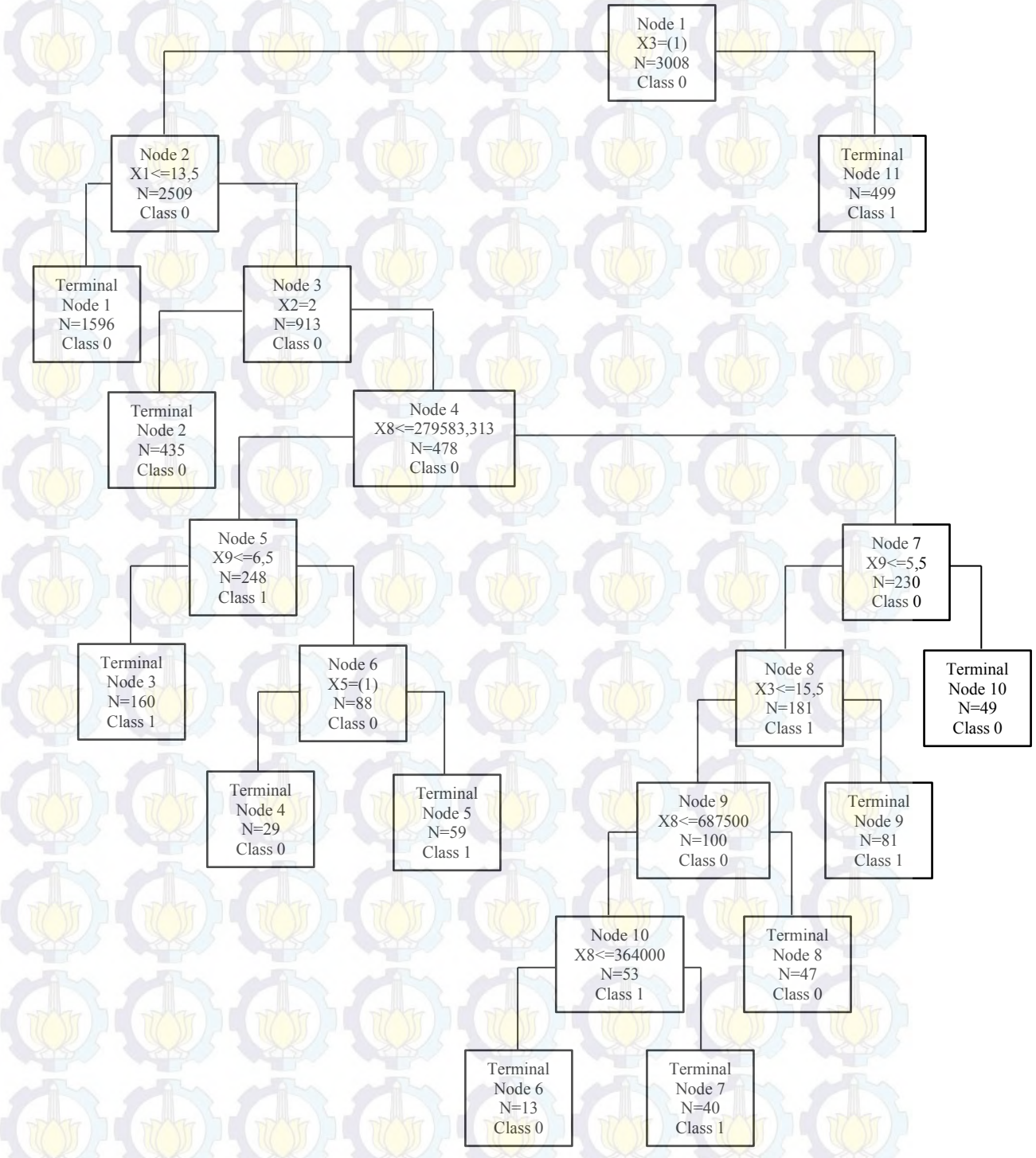
Pohon klasifikasi optimal yang dihasilkan dari proses pemangkasan sebelumnya tidak dibangun oleh sembilan variabel prediktor yang ada, hanya enam variabel saja yang masuk menjadi pemilah dalam model pohon optimal, baik yang dengan menggunakan kriteria indeks *Gini* maupun indeks *Twoing*, yaitu partisipasi sekolah anak (X_3), umur anak (X_1), jenis kelamin anak (X_2), jumlah anggota rumah tangga (X_9), pendapatan perkapita (X_8), dan tingkat pendidikan kepala rumah tangga (X_5).

Tabel 4.12. Variabel-Variabel Yang Masuk Dalam Model Pohon Optimal

Variabel	Deskripsi Variabel	Score
X_3	Partisipasi sekolah anak	100,000
X_1	Umur anak	82,334
X_2	Jenis kelamin anak	7,558
X_9	Jumlah anggota rumah tangga	5,631
X_8	Pendapatan perkapita	5,244
X_5	Tingkat pendidikan kepala rumah tangga	1,591

Berdasarkan tabel 4.12, terlihat bahwa variabel prediktor yang menjadi pemilah utama pada pohon optimal adalah variabel partisipasi sekolah anak (X_3) yang ditunjukkan oleh skor variabel penting. Skor variabel penting (*variabel importance*) merupakan suatu nilai yang menunjukkan besarnya kontribusi variabel sebagai pemilah utama maupun pemilah pengganti pada pohon klasifikasi yang terbentuk. Skor ini dihitung berdasarkan jumlah dari *impurity* yang dihasilkan oleh pemilah pada variabel dalam tiap simpul. Hal ini mengindikasikan

bahwa variabel partisipasi sekolah anak merupakan variabel yang paling dominan dalam pembentukan model pohon klasifikasi.



Gambar 4.7. Pohon Klasifikasi Optimal

Pohon klasifikasi optimal yang terbentuk dari hasil pemangkasan memiliki 11 simpul terminal dengan kedalaman 9. Pohon tersebut terdiri dari simpul utama, simpul internal dan simpul terminal. Adapun penjabaran pemilahan pada pohon optimal adalah sebagai berikut.

a. Simpul Internal

Simpul internal adalah simpul yang masih akan dipilah menjadi simpul internal maupun simpul terminal. Pada pohon klasifikasi optimal yang dihasilkan dari klasifikasi pekerja anak memiliki sebuah simpul utama, 9 simpul internal dan 11 simpul terminal.

- Simpul pertama merupakan simpul utama. Banyaknya sampel pada simpul utama adalah 3008 observasi. Variabel yang menjadi pemilah utama pohon optimal adalah variabel partisipasi sekolah anak (X_3) dengan nilai pemilah $X_3 = 1$ (anak bersekolah). Observasi yang termasuk dalam kelompok anak bersekolah $(X_3 = 1)$ masuk ke simpul kiri (simpul 2), sedangkan kelompok anak tidak bersekolah $(X_3 = 2)$ masuk ke simpul kanan (simpul terminal 11).
- Simpul 2 merupakan simpul internal yang berisi amatan untuk anak bersekolah dengan jumlah amatan sebanyak 2509. Selanjutnya simpul 2 dipilah berdasarkan variabel umur anak dengan nilai pemilah $X_1 \leq 13{,}5$ (umur dibawah 13,5 tahun). Anak bersekolah yang berumur dibawah 13,5 tahun akan masuk ke simpul kiri (simpul terminal 1), sedangkan anak bersekolah dengan umur antara 13,5 sampai dengan 15,5 tahun akan menjadi anggota simpul kanan (simpul 3).
- Simpul 3 merupakan simpul internal yang berisi 913 amatan dimana anggotanya adalah anak bersekolah usia 13,5 sampai 15,5 tahun. Simpul 2 dipilah berdasarkan variabel jenis kelamin anak dengan nilai pemilah $X_2 = 2$ (perempuan). Anak dengan jenis kelamin perempuan akan masuk ke simpul kiri (simpul terminal 2) sedangkan anak dengan jenis kelamin laki-laki akan masuk ke dalam simpul kanan (simpul 4).
- Simpul 4 mengandung 478 amatan yang beranggotakan anak laki-laki bersekolah usia 13,5 sampai 15,5 tahun. Selanjutnya simpul ini dipilah

berdasarkan variabel pendapatan perkapita dengan nilai pemilah $X_8 \leq 279583{,}313$. Amatan dengan pendapatan perkapita tidak lebih dari Rp. 279.583,313 akan menjadi anggota simpul kiri (simpul 5), sedangkan amatan dengan pendapatan perkapita antara Rp.279.583,313 sampai dengan Rp. 687.500 akan menjadi anggota simpul kanan (simpul 7).

- Simpul 5 mengandung 248 amatan dimana anggotanya adalah anak laki-laki bersekolah umur 13,5 sampai 15,5 tahun dengan pendapatan perkapita kurang dari Rp.279.583,313. Simpul ini akan dipilah berdasarkan variabel jumlah anggota rumah tangga dengan nilai pemilah $X_9 \leq 6{,}5$. Amatan dengan jumlah anggota rumah tangga 5 sampai 6 orang akan masuk ke dalam simpul kiri (simpul terminal 3), sedangkan amatan dengan jumlah anggota rumah tangga lebih dari 6 akan masuk ke dalam simpul kanan (simpul 6).
- Simpul 6 mempunyai anggota sebanyak 88 amatan, dimana anggotanya merupakan anak laki-laki bersekolah usia 13,5 sampai 15,5 tahun dengan pendapatan perkapita Rp.279,583,313 dan jumlah anggota rumah tangga lebih dari 6 orang. Selanjutnya simpul 6 akan dipilah berdasarkan variabel tingkat pendidikan kepala rumah tangga dengan nilai pemilah $X_5 = 1$ (Sekolah Dasar). Anak dengan orang tua yang berpendidikan setara sekolah dasar akan masuk ke dalam simpul kiri (simpul terminal 4), sedangkan anak dengan orang tua berpendidikan selain sekolah dasar akan masuk ke dalam simpul kanan (simpul terminal 5)
- Simpul 7 mempunyai anggota sebanyak 230 amatan, dimana anggotanya adalah anak laki-laki bersekolah usia 13,5 sampai 15,5 tahun dengan pendapatan perkapita antara Rp. 279.583,313 sampai Rp. 687.500. Simpul ini akan dipilah berdasarkan variabel jumlah anggota rumah tangga dengan nilai pemilah $X_9 \leq 5{,}5$. Anak dengan jumlah anggota rumah tangga paling banyak 5 orang akan masuk ke simpul kiri (simpul 8), sedangkan anak dengan jumlah anggota rumah tangga antara 5 sampai 6 orang akan masuk ke dalam simpul kanan (simpul terminal 10).

- Simpul 8 memiliki anggota sebanyak 181 amatan, dimana anggotanya merupakan anak laki-laki bersekolah usia 13,5 sampai 15,5 tahun dengan pendapatan perkapita antara Rp. 279.583,313 sampai Rp. 687.500 dan jumlah anggota rumah tangga paling banyak 5 orang. Simpul 8 akan dipilah berdasarkan variabel umur anak dengan nilai pemilah $X_1 \leq 15{,}5$. Anak dengan umur 13,5 sampai 15,5 tahun akan masuk ke simpul kiri (simpul 9), sedangkan anak dengan umur lebih dari 15,5 tahun akan masuk ke simpul kanan (simpul terminal 9).
- Simpul 9 merupakan simpul internal yang berisi 100 amatan. Simpul ini beranggotakan anak laki-laki bersekolah usia 13,5 sampai 15,5 tahun dengan pendapatan perkapita antara Rp. 279.583,313 sampai Rp. 687.500 dan jumlah anggota rumah tangga paling banyak 5 orang. Simpul 9 akan dipilah berdasarkan variabel pendapatan perkapita dengan nilai pemilah $X_8 \leq 687500$. Amatan dengan pendapatan perkapita antara Rp. 279.583,313 sampai Rp. 687.500 akan masuk ke simpul kiri (simpul 10), sedangkan amatan dengan pendapatan perkapita lebih dari Rp. 687.500 akan masuk ke simpul kanan (simpul terminal 8).
- Simpul 10 merupakan simpul internal berisikan 53 amatan yang anggotanya adalah anak laki-laki bersekolah usia 13,5 sampai 15,5 tahun dengan pendapatan perkapita antara Rp. 279.583,313 sampai Rp. 687.500 dan jumlah anggota rumah tangga paling banyak 5 orang. Selanjutnya simpul ini akan dipilah berdasarkan variabel pendapatan perkapita dengan nilai pemilah pemilah $X_8 \leq 364000$. Amatan dengan pendapatan perkapita antara Rp.279.583,313 sampai Rp. 364.000 akan masuk ke simpul kiri (simpul terminal 6), sedangkan amatan dengan pendapatan perkapita antara Rp. 364.000 sampai Rp. 687.500 akan masuk ke simpul kanan (simpul terminal 7).

b. Simpul Terminal

Simpul terminal adalah titik akhir dari suatu pemilahan berstruktur pada pohon klasifikasi. Simpul ini tidak dapat dipilah kembali menjadi simpul lain. Artinya, simpul terminal merupakan simpul yang mengandung amatan-amatan

yang homogen dan akhirnya akan dimasukkan sebagai suatu kelas tertentu. Berdasarkan pohon optimal yang dihasilkan dari klasifikasi pekerja anak, terdapat 10 simpul terminal. Adapun rincian simpul-simpul tersebut adalah sebagai berikut.

- Simpul terminal 1 merupakan simpul yang diberi label kelas 0, yang berarti amatan-amatan dalam simpul ini diidentifikasi sebagai kelompok anak yang tidak bekerja. Struktur sekuensial dari simpul ini mengindikasikan bahwa anggota simpul terminal ini merupakan anak bersekolah usia dibawah 13,5 tahun. Banyaknya amatan dalam kelompok ini adalah 1596. Amatan dalam simpul terminal 1 diprediksi sebagai anak yang tidak bekerja dengan probabilitas sebesar 0,376.
- Simpul terminal 2 terdiri atas 435 amatan yang diidentifikasi sebagai anak yang tidak bekerja dengan probabilitas sebesar 0,113. Struktur sekuensial dari simpul ini menggambarkan anak perempuan bersekolah usia 13,5 sampai 15,5 tahun.
- Simpul terminal 3 terdiri dari 160 amatan yang diidentifikasi sebagai anak yang bekerja dengan probabilitas sebesar 0,076. Menurut struktur sekuensialnya, amatan-amatan dalam simpul ini merupakan anak laki-laki bersekolah usia 13,5 sampai 15,5 tahun dengan pendapatan perkapita tidak lebih dari Rp. 279.583,313 dan memiliki anggota rumah tangga 5 sampai 6 orang.
- Simpul terminal 4 beranggotakan anak laki-laki bersekolah usia 13,5 sampai 15,5 tahun dengan pendapatan perkapita tidak lebih dari Rp. 279.583,313 dan memiliki anggota rumah tangga 5 sampai 6 orang serta pendidikan kepala rumah tangganya setara sekolah dasar. Simpul ini memiliki amatan sebanyak 29 dan diidentifikasi sebagai anak yang tidak bekerja.
- Simpul terminal 5 beranggotakan anak laki-laki bersekolah usia 13,5 sampai 15,5 tahun dengan pendapatan perkapita tidak lebih dari Rp. 279.583,313 dan memiliki anggota rumah tangga 5 sampai 6 orang serta pendidikan kepala rumah tangganya selain sekolah dasar. Simpul ini memiliki amatan sebanyak 59 dan diidentifikasi sebagai anak yang bekerja.

- Simpul terminal 6 merupakan simpul yang anggotanya diidentifikasi sebagai anak yang tidak bekerja dan memiliki amatan sebanyak 13. Simpul ini berdasarkan struktur sekuensialnya mengindikasikan anak laki-laki bersekolah usia 13,5 sampai 15,5 tahun dengan pendapatan perkapita antara Rp. 279.583,313 sampai Rp. 364.000 dan jumlah anggota rumah tangga paling banyak 5 orang.
- Simpul terminal 7 merupakan simpul dengan jumlah amatan sebanyak 40 dan diidentifikasi sebagai anak yang bekerja. Menurut struktur sekuensialnya, amatan-amatan dalam simpul ini adalah anak laki-laki bersekolah usia 13,5 sampai 15,5 tahun dengan pendapatan perkapita antara Rp. 364.000 sampai Rp. 687.500 dan jumlah anggota rumah tangga paling banyak 5 orang.
- Simpul terminal 8 merupakan simpul yang struktur sekuensialnya menggambarkan anak laki-laki bersekolah usia 13,5 sampai 15,5 tahun dengan pendapatan perkapita paling sedikit Rp. 687.500 dan jumlah anggota rumah tangga paling banyak 5 orang. Simpul ini terdiri dari 47 amatan dan diidentifikasi sebagai anak yang tidak bekerja.
- Simpul terminal 9 terdiri atas 81 amatan dan diidentifikasi sebagai anak yang bekerja. Menurut struktur sekuensialnya, simpul ini mengindikasikan anak laki-laki bersekolah usia di atas 15,5 tahun dengan pendapatan perkapita antara Rp. 279.583,313 sampai Rp. 687.500 dan jumlah anggota rumah tangga paling banyak 5 orang.
- Simpul terminal 10 terdiri dari 49 amatan dan diidentifikasi sebagai anak yang tidak bekerja. Struktur sekuensial simpul ini menggambarkan anak laki-laki bersekolah usia 13,5 sampai 15,5 tahun dengan pendapatan perkapita antara Rp. 279.583,313 sampai Rp. 687.500 dan jumlah anggota keluarga lebih dari 5 orang.
- Simpul terminal 11 terdiri dari 499 amatan dan diidentifikasi sebagai anak yang bekerja. Struktur sekuensialnya menggambarkan anak yang tidak bersekolah.

Berdasarkan penelusuran pohon optimal pada gambar 4.7, terlihat bahwa anak yang berada dalam simpul terminal 3, 5, 7, 9 dan 11 merupakan anak yang diindikasikan masuk sebagai anak yang bekerja. Selain itu, anak yang menjadi anggota simpul terminal 1, 2, 4, 6, 8 dan 10 teridentifikasi sebagai anak yang tidak bekerja.

Selanjutnya, setelah dilakukan pembentukan pohon klasifikasi optimal, dilakukan pengujian ketepatan klasifikasi pohon optimal yang terbentuk dengan melihat hasil klasifikasi untuk data *learning* dan data *testing*. Hasil klasifikasi data *learning* pada pohon klasifikasi optimal adalah sebagai berikut.

Tabel 4.13. Hasil Klasifikasi Data *Learning* Pada Pohon Klasifikasi Optimal

Observasi	Prediksi		*Sensitivity*	*Specitivity*	Ketepatan Klasifikasi
	0	1			
0	2066	513	80,1%	76,0%	79,5%
1	103	326			

Berdasarkan tabel di atas terlihat bahwa dari 3008 pengamatan data *learning*, terdapat 616 pengamatan yang salah diklasifikasi, dimana 513 pengamatan diklasifikasikan sebagai kategori 1 (anak yang bekerja) padahal sebenarnya merupakan kategori 0 (anak tidak bekerja) dan 103 pengamatan diklasifikasikan sebagai kategori 0 (anak tidak bekerja) padahal sebenarnya merupakan kategori 1 (anak yang bekerja).

Setelah diperoleh ketepatan klasifikasi dengan menggunakan data *learning*, selanjutnya menguji ketepatan klasifikasi dengan menggunakan data *testing*. Hasil klasifikasi data *testing* pada pohon klasifikasi optimal adalah sebagai berikut.

Tabel 4.14. Hasil Klasifikasi Data *Testing* Pada Pohon Klasifikasi Optimal

Observasi	Prediksi		*Sensitivity*	*Specitivity*	Ketepatan Klasifikasi
	0	1			
0	501	146	77,4%	76,0%	77,2%
1	30	95			

Berdasarkan tabel di atas terlihat bahwa dari 772 pengamatan data *testing*, terdapat 176 pengamatan yang salah diklasifikasi, dimana 146 pengamatan diklasifikasikan sebagai kategori 1 (anak yang bekerja) padahal sebenarnya merupakan kategori 0 (anak tidak bekerja) dan 30 pengamatan diklasifikasikan sebagai kategori 0 (anak tidak bekerja) padahal sebenarnya merupakan kategori 1 (anak yang bekerja).

Dari uraian yang telah dijabarkan, diperoleh bahwa pembentukan pohon klasifikasi dengan menggunakan kriteria indeks *Gini* dan indeks *Twoing* pada kasus pekerja anak di Sulawesi Tengah menghasilkan pohon maksimal yang berbeda dengan menghasilkan pohon dengan simpul terminal berturut-turut adalah 239 dan 236 buah. Namun pada proses pemangkasan dan pembentukan pohon klasifikasi optimal, kedua indeks ini menghasilkan pohon dan detail yang serupa, yaitu pohon optimal dengan 11 simpul terminal.

4.5. Hasil Klasifikasi Pekerja Anak Dengan Menerapkan *Bagging* Pada Pohon Klasifikasi

Penerapan teknik *bagging* dimaksudkan untuk meningkatkan ketepatan klasifikasi dari pohon klasifikasi yang dihasilkan. Proses *bagging* dilakukan untuk menghasilkan suatu himpunan data baru yang akan digunakan untuk membangun pohon klasifikasi. Mengadopsi rekomendasi dari para peneliti terdahulu maka dalam penelitian ini replikasi sampel *bootstrap* yang dibuat adalah 25, 50, 75, 100, 150 dan 200 kali. Hasil pengujian ketepatan klasifikasi dari pohon klasifikasi yang dihasilkan dari replikasi sampel bootstrap tersebut tersaji sebagai berikut.

Tabel 4.15. Tingkat Ketepatan Klasifikasi Dari Penerapan *Bagging*

Replikasi Sampel *Bootstrap*	*Sensitivity*	*Specitivity*	Total Akurasi
25 kali	93,9%	50%	86,4%
50 kali	94,4%	48,6%	86,7%
75 kali	94,2%	47,1%	86,2%
100 kali	94,4%	48,6%	86,7%
150 kali	94,7%	50%	87,1%
200 kali	95%	50%	87,4%

Berdasarkan tabel 4.15 terlihat bahwa ketepatan klasifikasi yang paling tinggi dicapai pada replikasi sebanyak 200 kali yaitu sebesar 87,4% dengan *sensitivity* sebesar 95% dan *specitivity* sebesar 50%. Banyaknya replikasi *bootstrap* juga menunjukkan beberapa penemuan, antara lain replikasi *bootstrap* 50 kali dan 100 kali menghasilkan ketepatan klasifikasi yang sama, namun peningkatan jumlah *bootstrap* dari 50 kali menjadi 75 kali justru menurunkan ketepatan klasifikasi sebelumnya. Peningkatan replikasi dari 100 kali hingga 200 kali menghasilkan peningkatan klasifikasi.

Pohon klasifikasi yang dihasilkan melalui penerapan *bagging* merupakan pohon klasifikasi yang kompleks karena dibentuk oleh semua variabel prediktor dan jumlahnya sesuai dengan banyaknya replikasi sampel bootstrap yang digunakan. Pengujian ketepatan pohon klasifikasinya dilakukan dengan menjalankan data *testing* ke dalam masing-masing pohon yang terbentuk. Prediksi yang dihasilkan kemudian divoting, dimana prediksi kelas yang paling banyak muncul merupakan prediksi akhir kasus tersebut.

Pengujian ketepatan klasifikasi dari pohon klasifikasi *bagging* dilakukan berdasarkan nilai *bagging misclassification rate*. Sedangkan untuk hasil klasifikasi sebelum penerapan *bagging* dilihat berdasarkan data *testing* yang dijalankan pada model pohon yang terbentuk. Adapun perbandingan hasil klasifikasi sebelum menerapkan teknik *bagging* dan sesudah menerapkan teknik *bagging* adalah sebagai berikut.

Tabel 4.16. Perbandingan Ketepatan Klasifikasi Sebelum dan Sesudah Penerapan Teknik *Bagging*

Uraian	*Sensitivity* (%)	*Specitivity* (%)	*Accuracy* (%)
Pohon Klasifikasi Tanpa *Bagging*	77,4	76	77,2
Pohon Klasifikasi Dengan *Bagging* (Replikasi 200 kali)	95	50	87,4
Perbedaan (%)	17,6	26	10,2

Tabel 4.16 menunjukkan bahwa penerapan teknik *bagging* pada pohon klasifikasi meningkatkan ketepatan klasifikasi dari 77,2% menjadi 87,4%. Artinya pada klasifikasi pekerja anak di provinsi Sulawesi Tengah penerapan *bagging* pada pohon klasifikasi dapat meningkatkan ketepatan klasifikasi sebesar 10,2%.

BAB 5
KESIMPULAN DAN SARAN

5.1. Kesimpulan

Berdasarkan hasil dan pembahasan yang telah dilakukan sebelumnya, maka dapat diperoleh beberapa kesimpulan sebagai berikut.

1. Indeks *Gini* dan indeks *Twoing* digunakan sebagai kriteria untuk membentuk pohon klasifikasi serta memiliki konsep dan bentuk fungsi yang berbeda. Dalam penggunaannya, kedua metode ini memiliki keuntungan masing-masing.
2. Pohon klasifikasi optimal yang dihasilkan dengan menggunakan indeks *Gini* dan indeks *Twoing* pada kasus pekerja anak di provinsi Sulawesi Tengah menghasilkan pohon yang identik. Pohon ini dibentuk oleh variabel partisipasi sekolah anak, umur anak, jenis kelamin anak, jumlah anggota rumah tangga, pendapatan perkapita dan tingkat pendidikan kepala rumah tangga. Berdasarkan pohon klasifikasi optimal tersebut dihasilkan enam kelompok yang teridentifikasi sebagai anak yang tidak bekerja dan lima kelompok yang diprediksi sebagai anak yang bekerja. Ke lima kelompok tersebut mempunyai karakteristik yang berbeda-beda, yaitu:
   - Kelompok 1 terdiri dari 160 amatan yang diidentifikasi sebagai anak yang bekerja dengan probabilitas sebesar 0,076. Dengan karakteristik anak laki-laki bersekolah usia 13,5 sampai 15,5 tahun dengan pendapatan perkapita tidak lebih dari Rp. 279.583,313 dan memiliki anggota rumah tangga tidak lebih dari 6 orang.
   - Kelompok 2 beranggotakan anak laki-laki bersekolah usia 13,5 sampai 15,5 tahun dengan pendapatan perkapita tidak lebih dari Rp. 279.583,313 dan memiliki anggota rumah tangga tidak lebih dari 6 orang serta pendidikan kepala rumah tangganya selain sekolah dasar. Simpul ini memiliki amatan sebanyak 59 dengan probabilitas sebesar 0,021 dan diidentifikasi sebagai anak yang bekerja.

- Kelompok 3 merupakan simpul dengan jumlah amatan sebanyak 40 dengan probabilitas sebesar 0,016 dan diidentifikasi sebagai anak yang bekerja. Menurut struktur sekuensialnya, amatan-amatan dalam simpul ini adalah anak laki-laki bersekolah usia 13,5 sampai 15,5 tahun dengan pendapatan perkapita antara Rp. 364.000 sampai Rp. 687.500 dan jumlah anggota rumah tangga paling banyak 5 orang.
- Kelompok 4 terdiri atas 81 amatan dan diidentifikasi sebagai anak yang bekerja dengan probabilitas sebesar 0,032. Menurut struktur sekuensialnya, simpul ini mengindikasikan anak laki-laki bersekolah usia di atas 15,5 tahun dengan pendapatan perkapita antara Rp. 279.583,313 sampai Rp. 687.500 dan jumlah anggota rumah tangga paling banyak 5 orang.
- Kelompok 5 terdiri atas 499 amatan dan diidentifikasi sebagai anak yang bekerja dengan probabilitas sebesar 0,334. Struktur sekuensialnya menggambarkan anak yang tidak bersekolah.

3. Penerapan teknik *bagging* pada pohon klasifikasi meningkatkan ketepatan klasifikasi dari 77,2% menjadi 87,4%. Dengan kata lain, pada klasifikasi pekerja anak di provinsi Sulawesi Tengah penerapan *bagging* pada pohon klasifikasi dapat meningkatkan ketepatan klasifikasi sebesar 10,2%.

5.2. Saran

Penelitian ini membahas konsep *splitting rule* dalam pembentukan pohon klasifikasi serta aplikasinya pada data kasus pekerja anak di Sulawesi Tengah, sehingga direkomendasikan pada penelitian selanjutnya dapat mengkaji bagian-bagian lain pada langkah-langkah pohon klasifikasi, seperti pemangkasan pohon klasifikasi (*pruning*), pemilihan pohon klasifikasi optimal, maupun bagian-bagian lainnya. Metode ini juga dapat diaplikasikan pada kasus yang sama maupun kasus-kasus lainnya yang masih masuk dalam ruang lingkup klasifikasi.

DAFTAR PUSTAKA

Aeni, E.Q. 2009. *Pendekatan CART Arcing Untuk Klasifikasi Kesejahteraan Rumah Tangga Di Provinsi Jawa Tengah*. Tesis: Institut Teknologi Sepuluh Nopember.

Badan Pusat Statistik. 2010. *Katalog Publikasi*. Jakarta: BPS Pusat.

Badan Pusat Statistik. 2014. *Survey Sosial Ekonomi Nasional 2013 Kor Gabungan*. Jakarta: BPS Pusat.

Buhlmann, P. and Bin, Y. 2002. *Analyzing Bagging*. The Annals Of Statistics. Vol. 30, No. 4, 927-961.

Breiman, L., Friedman, J., Olshen, R. and Stone, C. 1984. *Classification and Regression Trees*. New York – London: Chapman & Hall.

Breiman, L. 1996. *Bagging Predictors*. Berkeley: Statistics Department University of California.

Breiman, L. 1994. *Bagging Predictors*. Berkeley: University of California.

Damayanti, L.K. 2011. *Aplikasi Algoritma CART untuk mengklasifikasi Data Nasabah Asuransi Jiwa Bersama Bumiputera 1912 Surakarta*. Skripsi: Universitas Sebelas Maret.

Efron, B. and Tibshirani, R.J. 1993. *An Introduction To The Bootstrap*. London – New York – Washington DC: Chapman & Hall.

Hartati, A. 2012. *Analisis CART Pada Faktor-Faktor Yang Mempengaruhi Kepala Rumah Tangga Di Jawa Timur Melakukan Urbanisasi*. Tugas Akhir: Institut Teknologi Sepuluh Nopember.

Hastie, T., Tibshirani, R. and Friedman, J. 2001. The Elements of Statistical Learning, Data Mining, Inference and Prediction. New York. Springer-Verlag.

Husnaini, Z. 2011. *Pekerja Anak Di Bawah Umur (Studi Kasus Keluarga Pekerja Anak Di Kota Padang)*. Skripsi: Universitas Andalas.

International Labour Organization. 2013. *Press Releasess*. ILO.

Irwanto. 1995. *Child Labor in Three Metropolitan City, Jakarta, Surabaya and Medan*. UNICEF and Atma Jaya Research Centre Series.

Johnson, R. and Wichern, D. 2007. *Applied Multivariate Statistical Analysis Sixth Edition*. New Jersey: Pearson Prentice Hall.

Lewis, R.J. 2000. *An Introduction To Classification And Regression Tress (CART) Analysis*. California: Department of Emergency Medicine Harbor.

Manik. 2006. *Konsep Pekerja Anak*. Banda Aceh: PKPA.

Melkas., A. 1996. *Economics Incentives for Children and Families to Eliminate or Reduce Child Labour*. International Labour Office.

Morgan, J. 2014. *Classification and Regression Tree Analysis*. Boston: Boston University School of Public Health.

Nachrowi, D. 1997. *Pembangunan Keluarga Sejahtera dan Masalah Pekerja Anak*. Seminar Ilmiah Dies Natalis UI ke 47, 18 Maret 1997, Depok.

Nandi. 2006. *Pekerja Anak Dan Permasalahannya*. Jurnal GEA Jurusan Pendidikan Geografi. Vol 6.

Petersen, M., Molinaro, A., Sinisi., S. and Van Der Laan, M. 2005. *Cross-Validated Bagged Learning*. Journal of Multivariate Analysis, 98, 1693-1704.

Prakosa, H.M. 2011. *Klasifikasi Kesejahteraan Rumah Tangga di Provinsi Jawa Timur Dengan Pendekatan Bootstrap Aggregating Classification and Regression Trees*. Tugas Akhir: Institut Teknologi Sepuluh Nopember.

Schaefer, R.L. 1986. *Alternative estimators in logistic regression when the data are collinear*. Stat Comput Simul, 25, 79-91.

Siswadi. 2009. *Analisis Regresi Logistik Biner Bivariat Pada Partisipasi Anak Dalam kegiatan Ekonomi Dan Sekolah Di Jawa Timur*. Tesis: Institut Teknologi Sepuluh Nopember.

Sumarmi. 2009. *Bagging CART Pada Klasifikasi Anak Putus Sekolah Di Jambi*. Tesis: Institut Teknologi Sepuluh Nopember.

Sutton, C.D. 2005. *Classification and Regression Trees, Bagging and Boosting*. Handbook of Statistics, Vol 24.

UNICEF Indonesia. 2012. *Ringkasan Kajian Perlindungan Anak*. Jakarta: UNICEF.

Wibowo, A. 2011. *Penerapan Bagging Untuk Memperbaiki Hasil Prediksi Nasabah Perusahaan Asuransi X*. Tugas Akhir: Jurusan Teknik Informatika Politeknik Negeri Batam.

Xie, Z., Xu, Y., Hu, Q. and Zhu, P. 2012. *Margin Distribution Based Bagging Pruning*. Neurocomputing, 85, 11-19.

Yohannes, Y. and Hoddinot, J. 1999. *Classification and Regression Trees: An Introduction*. Washington DC: International Food Policy Research Intitute.

Zhang, D., Zhou, X., Leung, S. and Zheng, J. 2010. *Vertical Bagging Decision Trees Model For Credit Scoring*. Expert Systems With Application, 37, 7838-7843.

(*Halaman Ini Sengaja Dikosongkan*)

Lampiran 1. Tabulasi Silang

```
CROSSTABS
TABLES=Status_Bekerja BY JK_Anak Bersekolah Tk_Pendidikan JK_KRT
Pekerjaan
[DataSet0]
```

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Status_Bekerja * JK_Anak	3780	100.0%	0	0.0%	3780	100.0%
Status_Bekerja * Bersekolah	3780	100.0%	0	0.0%	3780	100.0%
Status_Bekerja * Tk_Pendidikan	3780	100.0%	0	0.0%	3780	100.0%
Status_Bekerja * JK_KRT	3780	100.0%	0	0.0%	3780	100.0%
Status_Bekerja * Pekerjaan	3780	100.0%	0	0.0%	3780	100.0%

Status_Bekerja * JK_Anak

Crosstab

Count

		JK_Anak		Total
		Laki-Laki	Perempuan	
Status_Bekerja	Tdk Bekerja	1617	1609	3226
	Bekerja	424	130	554
Total		2041	1739	3780

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	132.761[a]	1	.000		
Continuity Correction[b]	131.700	1	.000		
Likelihood Ratio	140.188	1	.000		
Fisher's Exact Test				.000	.000
Linear-by-Linear Association	132.726	1	.000		
N of Valid Cases	3780				

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 254.87.

b. Computed only for a 2x2 table

Lampiran 1 (Lanjutan)

Status_Bekerja * Bersekolah

Crosstab

Count

		Bersekolah		Total
		Tdk Sekolah	Sekolah	
Status_Bekerja	Tdk Bekerja	319	2907	3226
	Bekerja	318	236	554
Total		637	3143	3780

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	761.718[a]	1	.000		
Continuity Correction[b]	758.331	1	.000		
Likelihood Ratio	591.286	1	.000		
Fisher's Exact Test				.000	.000
Linear-by-Linear Association	761.517	1	.000		
N of Valid Cases	3780				

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 93.36.

b. Computed only for a 2x2 table

Status_Bekerja * Tk_Pendidikan

Crosstab

Count

		Tk_Pendidikan				Total
		Tdk Berijasah	SD	SMP	SMA	
Status_Bekerja	Tdk Bekerja	688	1363	558	617	3226
	Bekerja	187	220	86	61	554
Total		875	1583	644	678	3780

Lampiran 1 (Lanjutan)

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	50.507[a]	3	.000
Likelihood Ratio	49.633	3	.000
Linear-by-Linear Association	42.468	1	.000
N of Valid Cases	3780		

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 94.39.

Status_Bekerja * JK_KRT

Crosstab

Count

		JK_KRT		Total
		Laki-Laki	Perempuan	
Status_Bekerja	Tdk Bekerja	3073	153	3226
	Bekerja	510	44	554
Total		3583	197	3780

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	9.798[a]	1	.002		
Continuity Correction[b]	9.161	1	.002		
Likelihood Ratio	8.738	1	.003		
Fisher's Exact Test				.004	.002
Linear-by-Linear Association	9.795	1	.002		
N of Valid Cases	3780				

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 28.87.

b. Computed only for a 2x2 table

Lampiran 1 (Lanjutan)

Status_Bekerja * Pekerjaan

Crosstab

Count

		Pekerjaan					Total
		Pertanian	Industri	Perdagangan	Jasa	Lainnya	
Status_Bekerja	Tdk Bekerja	1628	154	563	745	136	3226
	Bekerja	335	32	88	69	30	554
Total		1963	186	651	814	166	3780

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	37.152[a]	4	.000
Likelihood Ratio	40.401	4	.000
Linear-by-Linear Association	21.266	1	.000
N of Valid Cases	3780		

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 24.33.

Lampiran 2. Uji Beda Rata-Rata

[DataSet1]

Group Statistics

	Status_Bekerja	N	Mean	Std. Deviation	Std. Error Mean
Umur_Anak	Tdk Bekerja	3226	13.0849	2.39150	.04211
	Bekerja	554	15.4332	2.21291	.09402
Umur_KRT	Tdk Bekerja	3226	44.2365	7.90550	.13919
	Bekerja	554	46.1679	8.22287	.34936
Pendapatan	Tdk Bekerja	3226	363863.9833	498475.25854	8776.29930
	Bekerja	554	311118.0956	432557.94446	18377.62606
Jml_ART	Tdk Bekerja	3226	5.2514	1.64539	.02897
	Bekerja	554	5.3267	1.76354	.07493

Lampiran 2 (Lanjutan)

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
Umur_Anak	Equal variances assumed	7.696	.006	-21.579	3778	.000	-2.34828	.10882	-2.56163	-2.13493
	Equal variances not assumed			-22.795	791.612	.000	-2.34828	.10302	-2.55049	-2.14606
Umur_KRT	Equal variances assumed	1.256	.263	-5.281	3778	.000	-1.93135	.36574	-2.64843	-1.21428
	Equal variances not assumed			-5.136	739.294	.000	-1.93135	.37606	-2.66963	-1.19308
Pendapatan	Equal variances assumed	9.309	.002	2.344	3778	.019	52745.88771	22506.42745	8619.96388	96871.81155
	Equal variances not assumed			2.590	826.621	.010	52745.88771	20365.67134	12771.37490	92720.40052
Jml_ART	Equal variances assumed	5.836	.016	-.985	3778	.325	-.07532	.07649	-.22529	.07465
	Equal variances not assumed			-.938	727.907	.349	-.07532	.08033	-.23303	.08239

Lampiran 3. Output Pengolahan Data Pohon Maksimal Dengan Kriteria Indeks *Gini*

```
CART VERSION 4.0.0.20

 RECORDS READ: 3780
 RECORDS WRITTEN IN LEARNING SAMPLE: 3008
 RECORDS WRITTEN IN TEST SAMPLE: 772

 =============
 TREE SEQUENCE
 =============

 Dependent variable: Y

  Terminal          Test Set          Resubstitution   Complexity
 Tree Nodes      Relative Cost       Relative Cost    Parameter
 ------------------------------------------------------------------
    1    239    0.537 +/- 0.045          0.130          0.000
   30     20    0.497 +/- 0.043          0.408          0.001
   31     18    0.496 +/- 0.042          0.415          0.002
   32     13    0.469 +/- 0.042          0.431          0.002
   33**   11    0.466 +/- 0.042          0.439          0.002
   34      9    0.486 +/- 0.042          0.448          0.002
   35      7    0.472 +/- 0.042          0.456          0.002
   36      5    0.493 +/- 0.044          0.471          0.004
   37      4    0.509 +/- 0.042          0.482          0.005
   38      2    0.507 +/- 0.045          0.530          0.012
   39      1    1.000 +/- .305176E-04    1.000          0.235

 Initial misclassification cost = 0.500
 Initial class assignment = 0

 ================================
 TEST SAMPLE CLASSIFICATION TABLE
 ================================

 Actual       Predicted Class              Actual
 Class              0          1           Total
 ------------------------------------------------
 0               522.00     125.00       647.00
 1                43.00      82.00       125.00
 ------------------------------------------------
 PRED. TOT.      565.00     207.00       772.00
 CORRECT          0.807      0.656
 SUCCESS IND.    -0.031      0.494
 TOT. CORRECT     0.782

     SENSITIVITY:        0.807     SPECIFICITY:        0.656
 FALSE REFERENCE:        0.076  FALSE RESPONSE:        0.604
 REFERENCE = "0", RESPONSE = "1"
 ------------------------------------------------------------
```

Lampiran 3 (Lanjutan)

```
==========================================
TEST SAMPLE CLASSIFICATION PROBABILITY TABLE
==========================================

Actual       Predicted Class                Actual
Class                 0          1          Total
-----------------------------------------------------
0                 0.807      0.193          1.000
1                 0.344      0.656          1.000
-----------------------------------------------------
====================================
LEARNING SAMPLE CLASSIFICATION TABLE
====================================

Actual       Predicted Class                Actual
Class                 0          1          Total
-----------------------------------------------------
0               2280.00     299.00       2579.00
1                  6.00     423.00        429.00
-----------------------------------------------------
PRED. TOT.      2286.00     722.00       3008.00
CORRECT            0.884      0.986
SUCCESS IND.       0.027      0.843
TOT. CORRECT       0.899

    SENSITIVITY:       0.884     SPECIFICITY:       0.986
FALSE REFERENCE:       0.003  FALSE RESPONSE:       0.414
REFERENCE = "0", RESPONSE = "1"
-------------------------------------------------------------

================================================
LEARNING SAMPLE CLASSIFICATION PROBABILITY TABLE
================================================

Actual       Predicted Class                Actual
Class                 0          1          Total
-----------------------------------------------------
0                 0.884      0.116          1.000
1                 0.014      0.986          1.000
-----------------------------------------------------
====================
VARIABLE IMPORTANCE
====================

              Relative     Number Of   Minimum
            Importance    Categories  Category
-----------------------------------------------
X1             100.000
X3              86.359          2         0
X4              63.344
X8              56.459
X9              40.175
X5              31.307          4         0
X7              28.031          5         1
X2              17.441          2         1
X6               6.531          2         1
```

Lampiran 4. Output Pengolahan Data Pohon Optimal Dengan Kriteria Indeks *Gini*

```
================================
TEST SAMPLE CLASSIFICATION TABLE
================================

Actual       Predicted Class                Actual
Class                 0           1         Total
-----------------------------------------------------
0                501.00      146.00       647.00
1                 30.00       95.00       125.00
-----------------------------------------------------
PRED. TOT.       531.00      241.00       772.00
CORRECT           0.774       0.760
SUCCESS IND.     -0.064       0.598
TOT. CORRECT      0.772

    SENSITIVITY:        0.774     SPECIFICITY:        0.760
FALSE REFERENCE:        0.056  FALSE RESPONSE:        0.606
REFERENCE = "0", RESPONSE = "1"
-------------------------------------------------------------


==========================================
TEST SAMPLE CLASSIFICATION PROBABILITY TABLE
==========================================

Actual       Predicted Class                Actual
Class                 0           1         Total
-----------------------------------------------------
0                 0.774       0.226        1.000
1                 0.240       0.760        1.000
-----------------------------------------------------


===================================
LEARNING SAMPLE CLASSIFICATION TABLE
===================================

Actual       Predicted Class                Actual
Class                 0           1         Total
-----------------------------------------------------
0              2066.00      513.00      2579.00
1               103.00      326.00       429.00
-----------------------------------------------------
PRED. TOT.     2169.00      839.00      3008.00
CORRECT           0.801       0.760
SUCCESS IND.     -0.056       0.617
TOT. CORRECT      0.795

    SENSITIVITY:        0.801     SPECIFICITY:        0.760
FALSE REFERENCE:        0.047  FALSE RESPONSE:        0.611
REFERENCE = "0", RESPONSE = "1"
-------------------------------------------------------------
```

Lampiran 4 (Lanjutan)

```
================================================
LEARNING SAMPLE CLASSIFICATION PROBABILITY TABLE
================================================

Actual       Predicted Class                Actual
Class                 0          1         Total
-------------------------------------------------
0                 0.801      0.199         1.000
1                 0.240      0.760         1.000
-------------------------------------------------


====================
VARIABLE IMPORTANCE
====================

                Relative     Number Of   Minimum
              Importance    Categories  Category
-------------------------------------------------
X3               100.000         2         0
X1                82.334
X2                 7.558         2         1
X4                 7.405
X9                 5.631
X8                 5.244
X5                 1.591         4         0
X7                 0.954         5         1
X6                 0.388         2         1

================
OPTION SETTINGS
================

Construction Rule          Gini  (priors altered by costs)
Estimation Method          Test Sample
Misclassification Costs    Unit
Tree Selection             0.000 se rule
Linear Combinations        No
```

Lampiran 4 (Lanjutan)

File: data variabel diolah.XLS
Target Variable: Y
Predictor Variables: X1, X4, X8, X9, X2, X3, X5, X6, X7

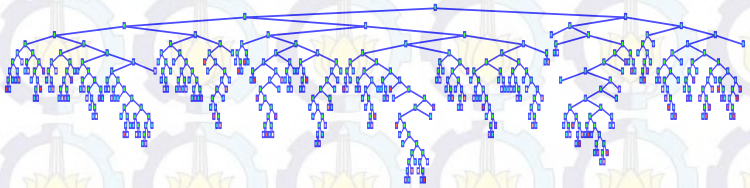
Tree Sequence

Tree Number	Terminal Nodes	Test Set Relative Cost	Resubstitution Relative Cost	Complexity
1	239	0.537 ± 0.045	0.130	-1.000
2	237	0.537 ± 0.045	0.130	1.00E-005
3	225	0.537 ± 0.045	0.131	5.85E-005
4	222	0.532 ± 0.045	0.131	7.46E-005
5	220	0.531 ± 0.045	0.132	0.000108
6	217	0.523 ± 0.045	0.133	0.000139
7	210	0.523 ± 0.045	0.135	0.000176
8	197	0.527 ± 0.045	0.140	0.000202
9	193	0.529 ± 0.045	0.142	0.000252
10	187	0.534 ± 0.045	0.146	0.000335
11	181	0.527 ± 0.045	0.150	0.000359
12	168	0.516 ± 0.044	0.160	0.000395
13	159	0.515 ± 0.044	0.168	0.000437
14	156	0.519 ± 0.044	0.171	0.000462
15	152	0.515 ± 0.044	0.175	0.000495
16	143	0.509 ± 0.044	0.184	0.000545
17	125	0.508 ± 0.044	0.205	0.000592
18	122	0.506 ± 0.044	0.209	0.000656
19	116	0.506 ± 0.044	0.217	0.000689
20	102	0.513 ± 0.043	0.239	0.000785
21	93	0.514 ± 0.044	0.254	0.000831
22	87	0.522 ± 0.044	0.264	0.000882
23	81	0.538 ± 0.045	0.275	0.000915
24	60	0.520 ± 0.044	0.314	0.000943
25	54	0.535 ± 0.045	0.326	0.000979
26	50	0.545 ± 0.045	0.335	0.001
27	42	0.525 ± 0.043	0.353	0.001
28	26	0.506 ± 0.044	0.392	0.001
29	24	0.483 ± 0.043	0.397	0.001
30	20	0.497 ± 0.043	0.408	0.001
31	18	0.496 ± 0.042	0.415	0.002
32	13	0.469 ± 0.042	0.431	0.002
33**	11	0.466 ± 0.042	0.439	0.002
34	9	0.486 ± 0.042	0.448	0.002
35	7	0.472 ± 0.042	0.456	0.002
36	5	0.493 ± 0.044	0.471	0.004
37	4	0.509 ± 0.042	0.482	0.005
38	2	0.507 ± 0.045	0.530	0.012
39	1	1.000 ± 3.05E-005	1.000	0.235

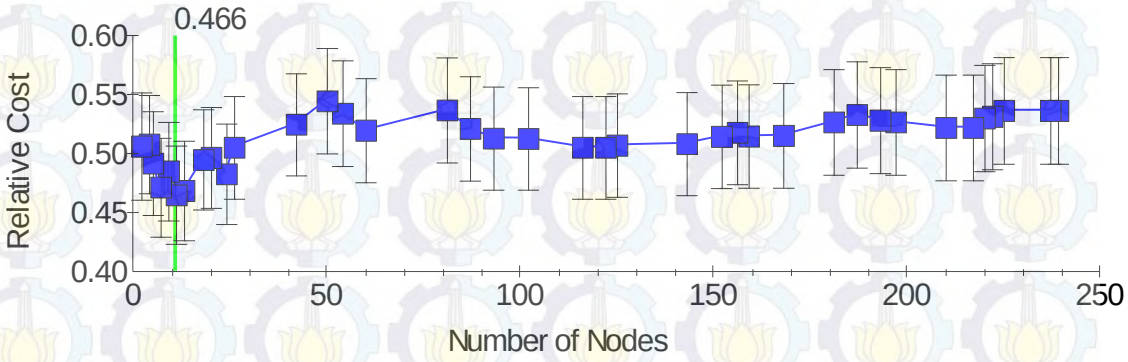
* Minimum Cost
** Optimal

Lampiran 4 (Lanjutan)

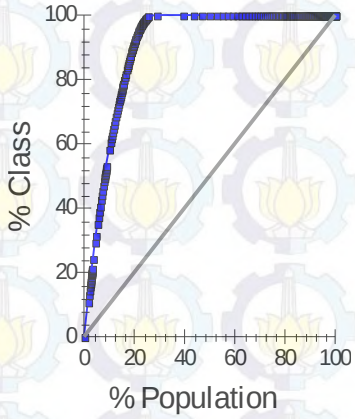
Classification tree topology for: Y



Relative Cost vs Number of Nodes



Gains for Class 1



Variable Importance

Variable	Score	
X1	100.00	
X3	87.86	
X4	64.68	
X8	57.74	
X9	39.34	
X5	31.20	
X7	28.37	
X2	17.86	
X6	6.13	

Lampiran 4 (Lanjutan)

Misclassification for Learn Data

Class	N Cases	N Mis-classed	Pct Error	Cost
1	429	4	0.93	0.01
0	2579	311	12.06	0.12

Misclassification for Test Data

Class	N Cases	N Mis-classed	Pct Error	Cost
0	647	125	19.32	0.19
1	125	42	33.60	0.34

Lampiran 4 (Lanjutan)

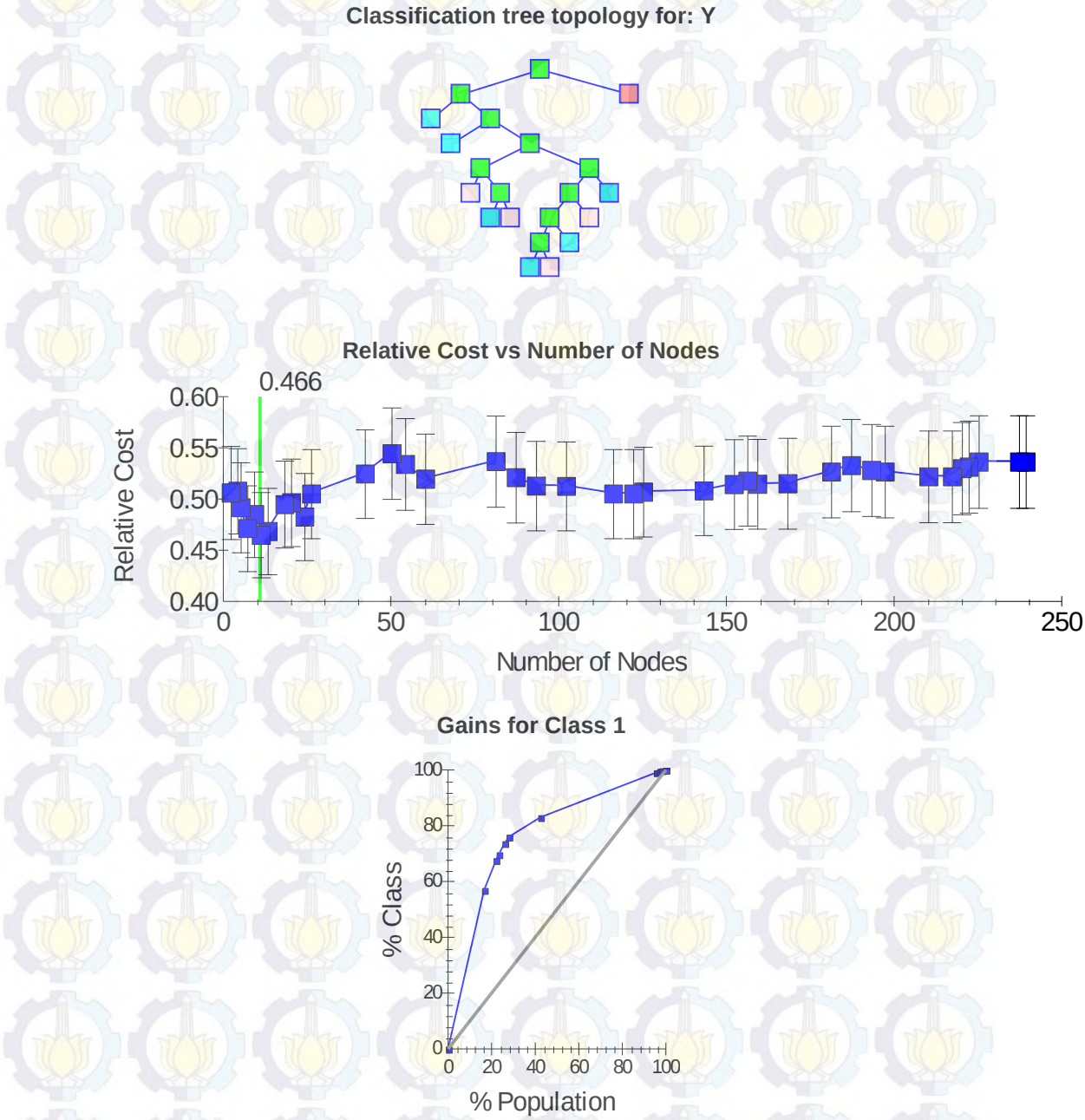
File: data variabel diolah.XLS
Target Variable: Y
Predictor Variables: X1, X4, X8, X9, X2, X3, X5, X6, X7

Tree Sequence

Tree Number	Terminal Nodes	Test Set Relative Cost	Resubstitution Relative Cost	Complexity
1	239	0.537 ± 0.045	0.130	-1.000
2	237	0.537 ± 0.045	0.130	1.00E-005
3	225	0.537 ± 0.045	0.131	5.85E-005
4	222	0.532 ± 0.045	0.131	7.46E-005
5	220	0.531 ± 0.045	0.132	0.000108
6	217	0.523 ± 0.045	0.133	0.000139
7	210	0.523 ± 0.045	0.135	0.000176
8	197	0.527 ± 0.045	0.140	0.000202
9	193	0.529 ± 0.045	0.142	0.000252
10	187	0.534 ± 0.045	0.146	0.000335
11	181	0.527 ± 0.045	0.150	0.000359
12	168	0.516 ± 0.044	0.160	0.000395
13	159	0.515 ± 0.044	0.168	0.000437
14	156	0.519 ± 0.044	0.171	0.000462
15	152	0.515 ± 0.044	0.175	0.000495
16	143	0.509 ± 0.044	0.184	0.000545
17	125	0.508 ± 0.044	0.205	0.000592
18	122	0.506 ± 0.044	0.209	0.000656
19	116	0.506 ± 0.044	0.217	0.000689
20	102	0.513 ± 0.043	0.239	0.000785
21	93	0.514 ± 0.044	0.254	0.000831
22	87	0.522 ± 0.044	0.264	0.000882
23	81	0.538 ± 0.045	0.275	0.000915
24	60	0.520 ± 0.044	0.314	0.000943
25	54	0.535 ± 0.045	0.326	0.000979
26	50	0.545 ± 0.045	0.335	0.001
27	42	0.525 ± 0.043	0.353	0.001
28	26	0.506 ± 0.044	0.392	0.001
29	24	0.483 ± 0.043	0.397	0.001
30	20	0.497 ± 0.043	0.408	0.001
31	18	0.496 ± 0.042	0.415	0.002
32	13	0.469 ± 0.042	0.431	0.002
33**	11	0.466 ± 0.042	0.439	0.002
34	9	0.486 ± 0.042	0.448	0.002
35	7	0.472 ± 0.042	0.456	0.002
36	5	0.493 ± 0.044	0.471	0.004
37	4	0.509 ± 0.042	0.482	0.005
38	2	0.507 ± 0.045	0.530	0.012
39	1	1.000 ± 3.05E-005	1.000	0.235

* Minimum Cost
** Optimal

Lampiran 4 (Lanjutan)



Lampiran 4 (Lanjutan)

Variable Importance

Variable	Score	
X3	100.00	
X1	82.33	
X2	7.56	
X4	7.00	
X8	5.14	
X9	5.02	
X5	1.59	
X7	0.75	
X6	0.00	

Misclassification for Learn Data

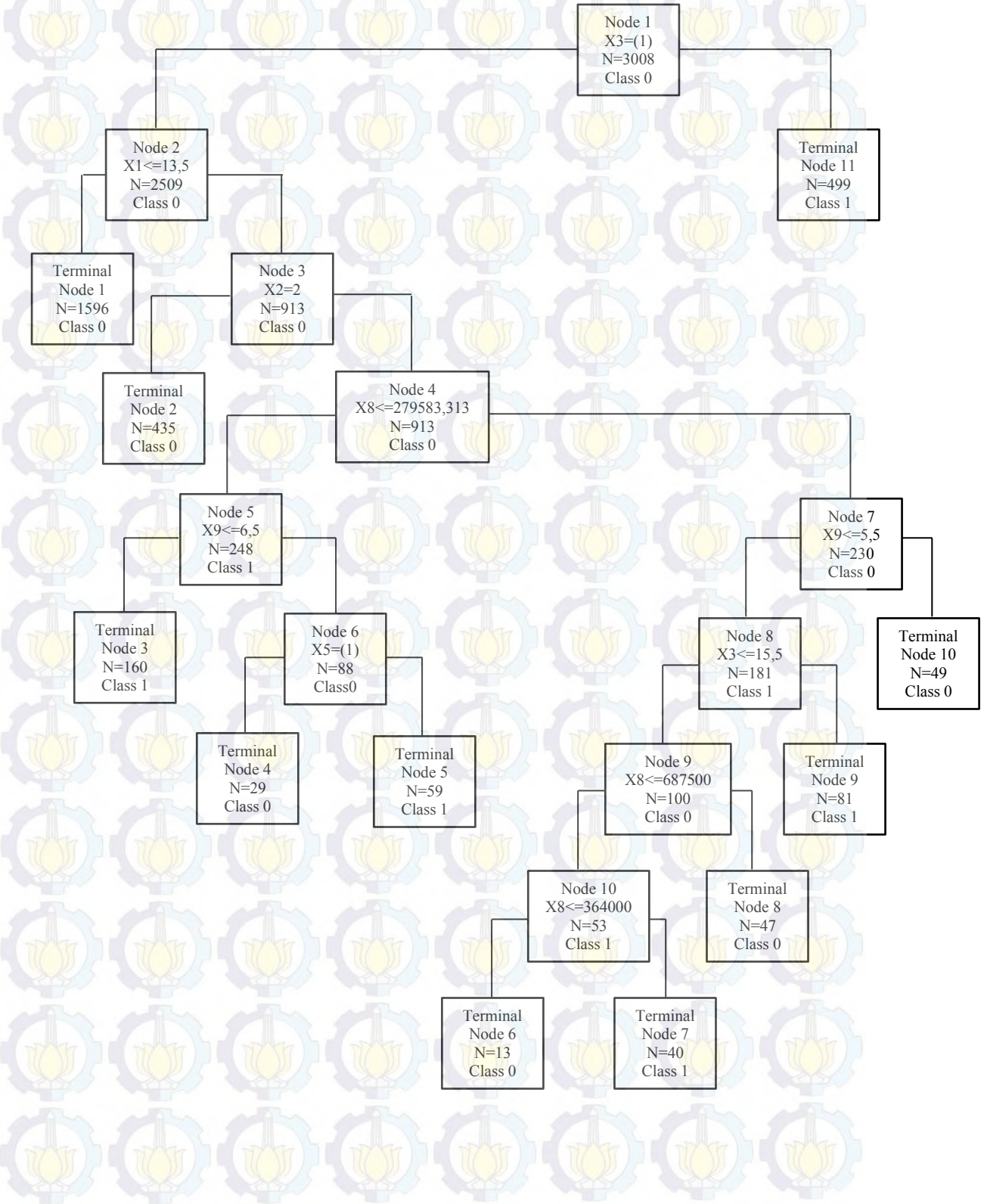
Class	N Cases	N Mis-classed	Pct Error	Cost
0	2579	513	19.89	0.20
1	429	103	24.01	0.24

Misclassification for Test Data

Class	N Cases	N Mis-classed	Pct Error	Cost
0	647	146	22.57	0.23
1	125	30	24.00	0.24

Lampiran 4 (Lanjutan)

Optimal Tree



Lampiran 5. Output Pengolahan Data Pohon Maksimal Dengan Kriteria Indeks *Twoing*

```
CART VERSION 4.0.0.20

RECORDS READ: 3780
RECORDS WRITTEN IN LEARNING SAMPLE: 3008
RECORDS WRITTEN IN TEST SAMPLE: 772
=============
TREE SEQUENCE
=============

Dependent variable: Y

  Terminal          Test Set          Resubstitution   Complexity
Tree Nodes       Relative Cost       Relative Cost    Parameter
-------------------------------------------------------------------
   1    236       0.540 +/- 0.045          0.128          0.000
  29     20       0.497 +/- 0.043          0.408          0.001
  30     18       0.496 +/- 0.042          0.415          0.002
  31     13       0.469 +/- 0.042          0.431          0.002
  32**   11       0.466 +/- 0.042          0.439          0.002
  33      9       0.486 +/- 0.042          0.448          0.002
  34      7       0.472 +/- 0.042          0.456          0.002
  35      5       0.493 +/- 0.044          0.471          0.004
  36      4       0.509 +/- 0.042          0.482          0.005
  37      2       0.507 +/- 0.045          0.530          0.012
  38      1       1.000 +/- .305176E-04    1.000          0.235

Initial misclassification cost = 0.500
Initial class assignment = 0

================================
TEST SAMPLE CLASSIFICATION TABLE
================================

Actual       Predicted Class                 Actual
Class                 0           1          Total
-----------------------------------------------------
0                520.00      127.00       647.00
1                 43.00       82.00       125.00
-----------------------------------------------------
PRED. TOT.       563.00      209.00       772.00
CORRECT           0.804       0.656
SUCCESS IND.     -0.034       0.494
TOT. CORRECT      0.780

   SENSITIVITY:       0.804     SPECIFICITY:       0.656
FALSE REFERENCE:      0.076  FALSE RESPONSE:       0.608
REFERENCE = "0", RESPONSE = "1"
-------------------------------------------------------------
```

Lampiran 5 (Lanjutan)

```
==========================================
TEST SAMPLE CLASSIFICATION PROBABILITY TABLE
==========================================

Actual        Predicted Class              Actual
Class                  0          1        Total
-----------------------------------------------------
0                  0.804      0.196        1.000
1                  0.344      0.656        1.000
-----------------------------------------------------

====================================
LEARNING SAMPLE CLASSIFICATION TABLE
====================================

Actual        Predicted Class              Actual
Class                  0          1        Total
-----------------------------------------------------
0               2278.00     301.00     2579.00
1                  5.00     424.00      429.00
-----------------------------------------------------
PRED. TOT.      2283.00     725.00     3008.00
CORRECT            0.883      0.988
SUCCESS IND.       0.026      0.846
TOT. CORRECT       0.898

    SENSITIVITY:       0.883     SPECIFICITY:       0.988
FALSE REFERENCE:       0.002  FALSE RESPONSE:       0.415
REFERENCE = "0", RESPONSE = "1"
-------------------------------------------------------------

================================================
LEARNING SAMPLE CLASSIFICATION PROBABILITY TABLE
================================================

Actual        Predicted Class              Actual
Class                  0          1        Total
-----------------------------------------------------
0                  0.883      0.117        1.000
1                  0.012      0.988        1.000
-----------------------------------------------------
====================
VARIABLE IMPORTANCE
====================

               Relative     Number Of   Minimum
             Importance    Categories  Category
-------------------------------------------------
X1              100.000
X3               86.952        2          0
X4               63.442
X8               57.521
X9               40.426
X5               30.937        4          0
X7               27.843        5          1
X2               17.561        2          1
X6                6.570        2          1
```

Lampiran 6. Output Pengolahan Data Pohon Optimal Dengan Kriteria Indeks *Twoing*

```
================================
TEST SAMPLE CLASSIFICATION TABLE
================================

Actual       Predicted Class                Actual
Class                 0          1         Total
-----------------------------------------------------
0                501.00     146.00        647.00
1                 30.00      95.00        125.00
-----------------------------------------------------
PRED. TOT.       531.00     241.00        772.00
CORRECT           0.774      0.760
SUCCESS IND.     -0.064      0.598
TOT. CORRECT      0.772

   SENSITIVITY:       0.774     SPECIFICITY:       0.760
FALSE REFERENCE:      0.056  FALSE RESPONSE:       0.606
REFERENCE = "0", RESPONSE = "1"
-------------------------------------------------------------


============================================
TEST SAMPLE CLASSIFICATION PROBABILITY TABLE
============================================

Actual       Predicted Class                Actual
Class                 0          1         Total
-----------------------------------------------------
0                 0.774      0.226         1.000
1                 0.240      0.760         1.000
-----------------------------------------------------


===================================
LEARNING SAMPLE CLASSIFICATION TABLE
===================================

Actual       Predicted Class                Actual
Class                 0          1         Total
-----------------------------------------------------
0               2066.00     513.00      2579.00
1                103.00     326.00       429.00
-----------------------------------------------------
PRED. TOT.      2169.00     839.00      3008.00
CORRECT           0.801      0.760
SUCCESS IND.     -0.056      0.617
TOT. CORRECT      0.795

   SENSITIVITY:       0.801     SPECIFICITY:       0.760
FALSE REFERENCE:      0.047  FALSE RESPONSE:       0.611
REFERENCE = "0", RESPONSE = "1"
-------------------------------------------------------------
```

Lampiran 6 (Lanjutan)

```
================================================
LEARNING SAMPLE CLASSIFICATION PROBABILITY TABLE
================================================

Actual        Predicted Class                Actual
Class                  0            1        Total
-------------------------------------------------------
0                  0.801        0.199        1.000
1                  0.240        0.760        1.000
-------------------------------------------------------


====================
VARIABLE IMPORTANCE
====================

                  Relative     Number Of   Minimum
                Importance    Categories  Category
-------------------------------------------------
X3                100.000         2         0
X1                 82.334
X2                  7.558         2         1
X4                  7.405
X9                  5.631
X8                  5.244
X5                  1.591         4         0
X7                  0.954         5         1
X6                  0.388         2         1

================
OPTION SETTINGS
================

Construction Rule           Twoing
Estimation Method           Test Sample
Misclassification Costs     Unit
Tree Selection              0.000 se rule
Linear Combinations         No
```

Lampiran 6 (Lanjutan)

File: data variabel diolah.XLS
Target Variable: Y
Predictor Variables: X1, X4, X8, X9, X2, X3, X5, X6, X7

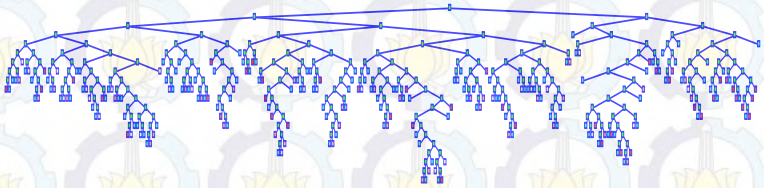
Tree Sequence

Tree Number	Terminal Nodes	Test Set Relative Cost	Resubstitution Relative Cost	Complexity
1	236	0.540 ± 0.045	0.128	-1.000
2	234	0.540 ± 0.045	0.128	1.00E-005
3	226	0.540 ± 0.045	0.129	5.85E-005
4	223	0.535 ± 0.045	0.130	7.46E-005
5	221	0.534 ± 0.045	0.130	0.000108
6	218	0.526 ± 0.045	0.131	0.000139
7	211	0.526 ± 0.045	0.133	0.000176
8	199	0.527 ± 0.045	0.138	0.000204
9	195	0.529 ± 0.045	0.140	0.000252
10	189	0.534 ± 0.045	0.144	0.000335
11	183	0.527 ± 0.045	0.148	0.000359
12	170	0.516 ± 0.044	0.158	0.000395
13	161	0.515 ± 0.044	0.166	0.000437
14	158	0.519 ± 0.044	0.168	0.000462
15	154	0.515 ± 0.044	0.172	0.000495
16	145	0.509 ± 0.044	0.182	0.000545
17	127	0.509 ± 0.044	0.203	0.000592
18	124	0.507 ± 0.044	0.207	0.000656
19	118	0.507 ± 0.044	0.215	0.000689
20	93	0.514 ± 0.044	0.254	0.000785
21	87	0.522 ± 0.044	0.264	0.000882
22	81	0.538 ± 0.045	0.275	0.000915
23	60	0.520 ± 0.044	0.314	0.000943
24	54	0.535 ± 0.045	0.326	0.000979
25	50	0.545 ± 0.045	0.335	0.001
26	42	0.525 ± 0.043	0.353	0.001
27	26	0.506 ± 0.044	0.392	0.001
28	24	0.483 ± 0.043	0.397	0.001
29	20	0.497 ± 0.043	0.408	0.001
30	18	0.496 ± 0.042	0.415	0.002
31	13	0.469 ± 0.042	0.431	0.002
32**	11	0.466 ± 0.042	0.439	0.002
33	9	0.486 ± 0.042	0.448	0.002
34	7	0.472 ± 0.042	0.456	0.002
35	5	0.493 ± 0.044	0.471	0.004
36	4	0.509 ± 0.042	0.482	0.005
37	2	0.507 ± 0.045	0.530	0.012
38	1	1.000 ± 3.05E-005	1.000	0.235

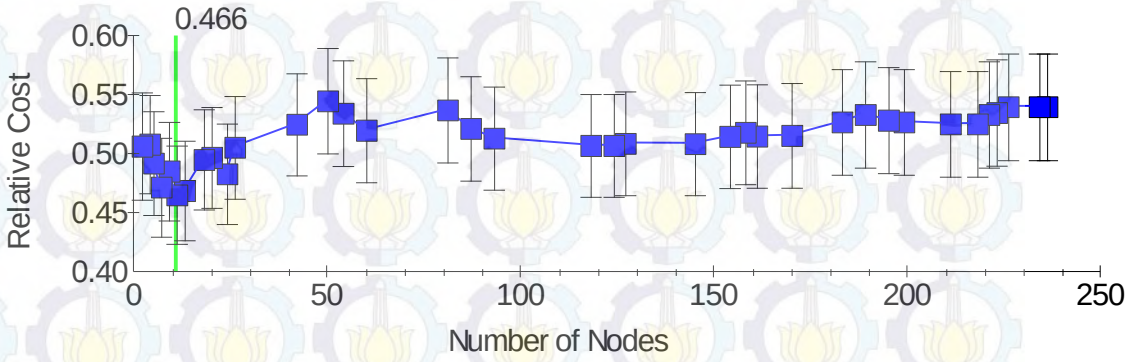
* Minimum Cost
** Optimal

Lampiran 6 (Lanjutan)

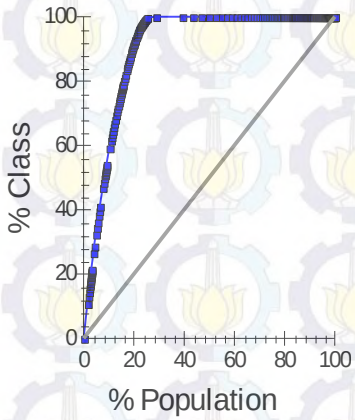
Classification tree topology for: Y



Relative Cost vs Number of Nodes



Gains for Class 1



Variable Importance

Variable	Score	
X1	100.00	
X3	88.47	
X4	64.79	
X8	58.83	
X9	39.59	
X5	31.01	
X7	28.18	
X2	17.99	
X6	6.17	

Lampiran 6 (Lanjutan)

Misclassification for Learn Data

Class	N Cases	N Mis-classed	Pct Error	Cost
1	429	3	0.70	0.01
0	2579	313	12.14	0.12

Misclassification for Test Data

Class	N Cases	N Mis-classed	Pct Error	Cost
0	647	127	19.63	0.20
1	125	42	33.60	0.34

Lampiran 6 (Lanjutan)

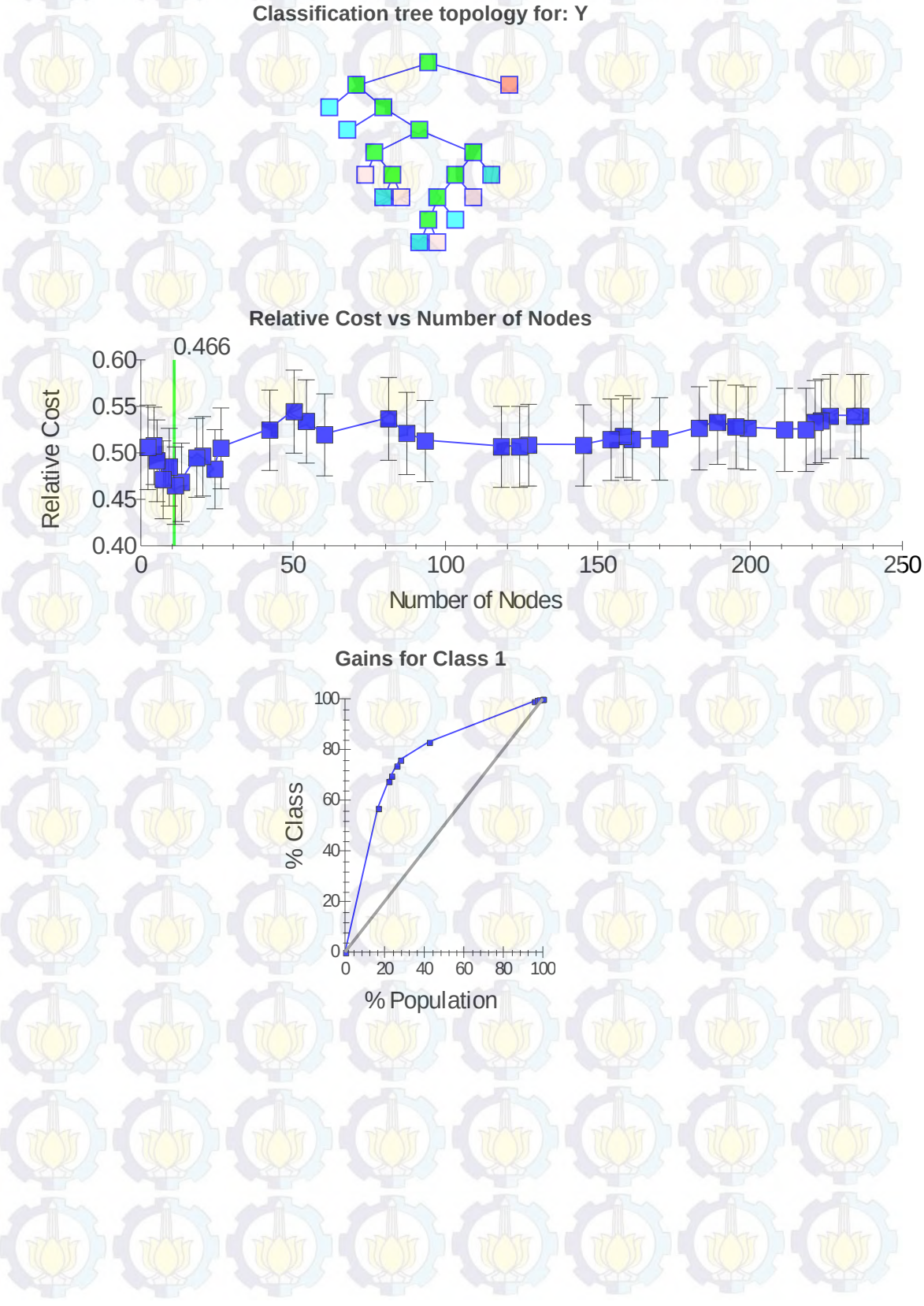
File: data variabel diolah.XLS
Target Variable: Y
Predictor Variables: X1, X4, X8, X9, X2, X3, X5, X6, X7

Tree Sequence

Tree Number	Terminal Nodes	Test Set Relative Cost	Resubstitution Relative Cost	Complexity
1	236	0.540 ± 0.045	0.128	-1.000
2	234	0.540 ± 0.045	0.128	1.00E-005
3	226	0.540 ± 0.045	0.129	5.85E-005
4	223	0.535 ± 0.045	0.130	7.46E-005
5	221	0.534 ± 0.045	0.130	0.000108
6	218	0.526 ± 0.045	0.131	0.000139
7	211	0.526 ± 0.045	0.133	0.000176
8	199	0.527 ± 0.045	0.138	0.000204
9	195	0.529 ± 0.045	0.140	0.000252
10	189	0.534 ± 0.045	0.144	0.000335
11	183	0.527 ± 0.045	0.148	0.000359
12	170	0.516 ± 0.044	0.158	0.000395
13	161	0.515 ± 0.044	0.166	0.000437
14	158	0.519 ± 0.044	0.168	0.000462
15	154	0.515 ± 0.044	0.172	0.000495
16	145	0.509 ± 0.044	0.182	0.000545
17	127	0.509 ± 0.044	0.203	0.000592
18	124	0.507 ± 0.044	0.207	0.000656
19	118	0.507 ± 0.044	0.215	0.000689
20	93	0.514 ± 0.044	0.254	0.000785
21	87	0.522 ± 0.044	0.264	0.000882
22	81	0.538 ± 0.045	0.275	0.000915
23	60	0.520 ± 0.044	0.314	0.000943
24	54	0.535 ± 0.045	0.326	0.000979
25	50	0.545 ± 0.045	0.335	0.001
26	42	0.525 ± 0.043	0.353	0.001
27	26	0.506 ± 0.044	0.392	0.001
28	24	0.483 ± 0.043	0.397	0.001
29	20	0.497 ± 0.043	0.408	0.001
30	18	0.496 ± 0.042	0.415	0.002
31	13	0.469 ± 0.042	0.431	0.002
32**	11	0.466 ± 0.042	0.439	0.002
33	9	0.486 ± 0.042	0.448	0.002
34	7	0.472 ± 0.042	0.456	0.002
35	5	0.493 ± 0.044	0.471	0.004
36	4	0.509 ± 0.042	0.482	0.005
37	2	0.507 ± 0.045	0.530	0.012
38	1	1.000 ± 3.05E-005	1.000	0.235

* Minimum Cost
** Optimal

Lampiran 6 (Lanjutan)



Lampiran 6 (Lanjutan)

Variable Importance

Variable	Score	
X3	100.00	
X1	82.33	
X2	7.56	
X4	7.00	
X8	5.14	
X9	5.02	
X5	1.59	
X7	0.75	
X6	0.00	

Misclassification for Learn Data

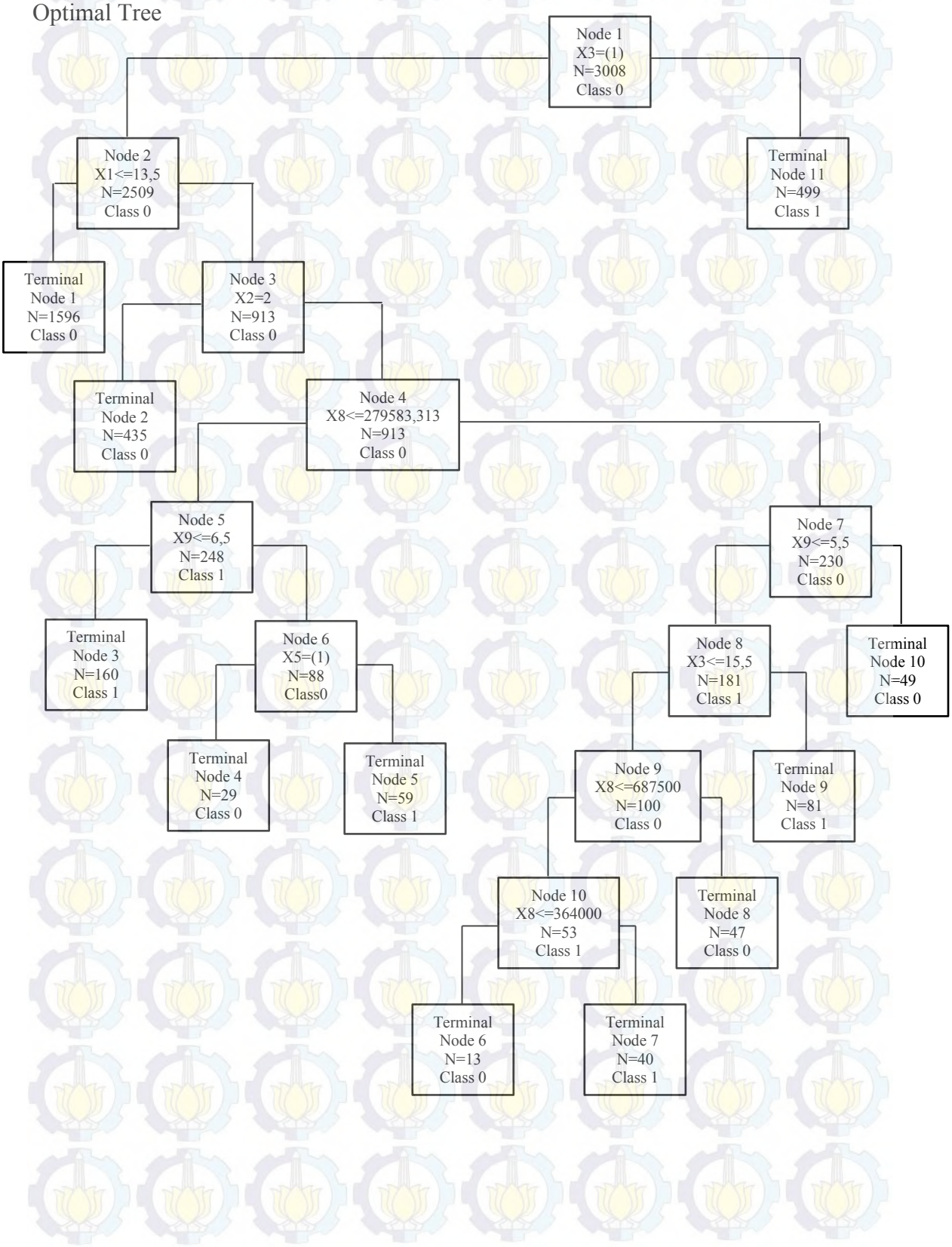
Class	N Cases	N Mis-classed	Pct Error	Cost
0	2579	513	19.89	0.20
1	429	103	24.01	0.24

Misclassification for Test Data

Class	N Cases	N Mis-classed	Pct Error	Cost
0	647	146	22.57	0.23
1	125	30	24.00	0.24

Lampiran 6 (Lanjutan)

Optimal Tree



Lampiran 7. Output Pengolahan *Bagging* 25 Kali

```
==================
COMMITTEE SUMMARY
==================

                  % CORRECT
FIRST TREE ONLY        0.784
BOOTSTRAP SET          0.864


==================================
BOOTSTRAP SET CLASSIFICATION TABLE
==================================

I=INITIAL TREE, B=BOOTSTRAP SET OF TREES

Actual       Predicted Class                  Actual
Class                     0           1       Total
------------------------------------------------------------
I 0                  282.00       60.00      342.00
B                    321.00       21.00
I 1                   29.00       41.00       70.00
B                     35.00       35.00
------------------------------------------------------------
INITIAL TREE
PRED. TOT.           311.00      101.00      412.00
CORRECT               0.825       0.586
SUCCESS IND.         -0.006       0.416
Tot. Correct          0.784
------------------------------------------------------------
BOOTSTRAP SET OF TREES
PRED. TOT.           356.00       56.00
CORRECT               0.939       0.500
SUCCESS IND.          0.108       0.330
Tot. Correct          0.864
------------------------------------------------------------
INITIAL TREE
    SENSITIVITY:      0.825     SPECIFICITY:        0.586
FALSE REFERENCE:      0.093  FALSE RESPONSE:        0.594
REFERENCE = "0", RESPONSE = "1"
------------------------------------------------------------
BOOTSTRAP SET OF TREES
    SENSITIVITY:      0.939     SPECIFICITY:        0.500
FALSE REFERENCE:      0.098  FALSE RESPONSE:        0.375
------------------------------------------------------------

C:\Users\Administrator\Documents\data variabel diolah.XLS[xls5]: 3780
RECORDS.
```

Lampiran 8. Output Pengolahan *Bagging* 50 kali

```
 ==================
 COMMITTEE SUMMARY
 ==================

                  % CORRECT
 FIRST TREE ONLY      0.784
 BOOTSTRAP SET        0.867


 ===================================
 BOOTSTRAP SET CLASSIFICATION TABLE
 ===================================

 I=INITIAL TREE, B=BOOTSTRAP SET OF TREES

 Actual       Predicted Class                  Actual
 Class                    0          1         Total
 ---------------------------------------------------
 I 0                 282.00      60.00       342.00
 B                   323.00      19.00
 I 1                  29.00      41.00        70.00
 B                    36.00      34.00
 ---------------------------------------------------
 INITIAL TREE
 PRED. TOT.          311.00     101.00       412.00
 CORRECT               0.825      0.586
 SUCCESS IND.         -0.006      0.416
 Tot. Correct          0.784
 ---------------------------------------------------
 BOOTSTRAP SET OF TREES
 PRED. TOT.          359.00      53.00
 CORRECT               0.944      0.486
 SUCCESS IND.          0.114      0.316
 Tot. Correct          0.867
 ---------------------------------------------------------------
 INITIAL TREE
    SENSITIVITY:       0.825    SPECIFICITY:        0.586
 FALSE REFERENCE:      0.093  FALSE RESPONSE:       0.594
 REFERENCE = "0", RESPONSE = "1"
 ---------------------------------------------------------------
 BOOTSTRAP SET OF TREES
    SENSITIVITY:       0.944    SPECIFICITY:        0.486
 FALSE REFERENCE:      0.100  FALSE RESPONSE:       0.358
 ---------------------------------------------------------------

 C:\Users\Administrator\Documents\data variabel diolah.XLS[xls5]: 3780
RECORDS.
```

Lampiran 9. Output Pengolahan *Bagging* 75 Kali

```
==================
COMMITTEE SUMMARY
==================

                  % CORRECT
FIRST TREE ONLY        0.784
BOOTSTRAP SET          0.862


==================================
BOOTSTRAP SET CLASSIFICATION TABLE
==================================

I=INITIAL TREE, B=BOOTSTRAP SET OF TREES

Actual       Predicted Class                    Actual
Class                     0           1         Total
------------------------------------------------------------
I 0                  282.00       60.00       342.00
B                    322.00       20.00
I 1                   29.00       41.00        70.00
B                     37.00       33.00
------------------------------------------------------------
INITIAL TREE
PRED. TOT.           311.00      101.00       412.00
CORRECT               0.825       0.586
SUCCESS IND.         -0.006       0.416
Tot. Correct          0.784
------------------------------------------------------------
BOOTSTRAP SET OF TREES
PRED. TOT.           359.00       53.00
CORRECT               0.942       0.471
SUCCESS IND.          0.111       0.302
Tot. Correct          0.862
------------------------------------------------------------------
INITIAL TREE
    SENSITIVITY:      0.825     SPECIFICITY:        0.586
FALSE REFERENCE:      0.093  FALSE RESPONSE:        0.594
REFERENCE = "0", RESPONSE = "1"
------------------------------------------------------------------
BOOTSTRAP SET OF TREES
    SENSITIVITY:      0.942     SPECIFICITY:        0.471
FALSE REFERENCE:      0.103  FALSE RESPONSE:        0.377
------------------------------------------------------------------

C:\Users\Administrator\Documents\data variabel diolah.XLS[xls5]: 3780
RECORDS.
```

Lampiran 10. Output Pengolahan *Bagging* 100 Kali

```
 ==================
 COMMITTEE SUMMARY
 ==================

                  % CORRECT
 FIRST TREE ONLY      0.784
 BOOTSTRAP SET        0.867


 ==================================
 BOOTSTRAP SET CLASSIFICATION TABLE
 ==================================

 I=INITIAL TREE, B=BOOTSTRAP SET OF TREES

 Actual       Predicted Class                  Actual
 Class                  0          1           Total
 ---------------------------------------------------------
 I 0                282.00      60.00         342.00
 B                  323.00      19.00
 I 1                 29.00      41.00          70.00
 B                   36.00      34.00
 ---------------------------------------------------------
 INITIAL TREE
 PRED. TOT.         311.00     101.00         412.00
 CORRECT              0.825      0.586
 SUCCESS IND.        -0.006      0.416
 Tot. Correct         0.784
 ---------------------------------------------------------
 BOOTSTRAP SET OF TREES
 PRED. TOT.         359.00      53.00
 CORRECT              0.944      0.486
 SUCCESS IND.         0.114      0.316
 Tot. Correct         0.867
 ---------------------------------------------------------------
 INITIAL TREE
    SENSITIVITY:      0.825    SPECIFICITY:        0.586
 FALSE REFERENCE:     0.093  FALSE RESPONSE:       0.594
 REFERENCE = "0", RESPONSE = "1"
 ---------------------------------------------------------------
 BOOTSTRAP SET OF TREES
    SENSITIVITY:      0.944    SPECIFICITY:        0.486
 FALSE REFERENCE:     0.100  FALSE RESPONSE:       0.358
 ---------------------------------------------------------------

 C:\Users\Administrator\Documents\data variabel diolah.XLS[xls5]: 3780
RECORDS.
```

Lampiran 11. Output Pengolahan *Bagging* 150 Kali

```
==================
COMMITTEE SUMMARY
==================

                  % CORRECT
FIRST TREE ONLY        0.784
BOOTSTRAP SET          0.871


==================================
BOOTSTRAP SET CLASSIFICATION TABLE
==================================

I=INITIAL TREE, B=BOOTSTRAP SET OF TREES

Actual       Predicted Class                    Actual
Class                     0           1         Total
------------------------------------------------------------
I 0                  282.00       60.00       342.00
B                    324.00       18.00
I 1                   29.00       41.00        70.00
B                     35.00       35.00
------------------------------------------------------------
INITIAL TREE
PRED. TOT.           311.00      101.00       412.00
CORRECT               0.825       0.586
SUCCESS IND.         -0.006       0.416
Tot. Correct          0.784
------------------------------------------------------------
BOOTSTRAP SET OF TREES
PRED. TOT.           359.00       53.00
CORRECT               0.947       0.500
SUCCESS IND.          0.117       0.330
Tot. Correct          0.871
------------------------------------------------------------------
INITIAL TREE
    SENSITIVITY:        0.825     SPECIFICITY:        0.586
FALSE REFERENCE:        0.093  FALSE RESPONSE:        0.594
REFERENCE = "0", RESPONSE = "1"
------------------------------------------------------------------
BOOTSTRAP SET OF TREES
    SENSITIVITY:        0.947     SPECIFICITY:        0.500
FALSE REFERENCE:        0.097  FALSE RESPONSE:        0.340
------------------------------------------------------------------

C:\Users\Administrator\Documents\data variabel diolah.XLS[xls5]: 3780
RECORDS.
```

Lampiran 12. Output Pengolahan *Bagging* 200 Kali

```
 ==================
 COMMITTEE SUMMARY
 ==================

                  % CORRECT
 FIRST TREE ONLY      0.784
 BOOTSTRAP SET        0.874


 ===================================
 BOOTSTRAP SET CLASSIFICATION TABLE
 ===================================

 I=INITIAL TREE, B=BOOTSTRAP SET OF TREES

 Actual       Predicted Class                  Actual
 Class                   0          1          Total
 ---------------------------------------------------------
 I 0                282.00      60.00        342.00
 B                  325.00      17.00
 I 1                 29.00      41.00         70.00
 B                   35.00      35.00
 ---------------------------------------------------------
 INITIAL TREE
 PRED. TOT.         311.00     101.00        412.00
 CORRECT              0.825      0.586
 SUCCESS IND.        -0.006      0.416
 Tot. Correct         0.784
 ---------------------------------------------------------
 BOOTSTRAP SET OF TREES
 PRED. TOT.         360.00      52.00
 CORRECT              0.950      0.500
 SUCCESS IND.         0.120      0.330
 Tot. Correct         0.874
 ---------------------------------------------------------------
 INITIAL TREE
    SENSITIVITY:      0.825    SPECIFICITY:        0.586
 FALSE REFERENCE:     0.093  FALSE RESPONSE:       0.594
 REFERENCE = "0", RESPONSE = "1"
 ---------------------------------------------------------------
 BOOTSTRAP SET OF TREES
    SENSITIVITY:      0.950    SPECIFICITY:        0.500
 FALSE REFERENCE:     0.097  FALSE RESPONSE:       0.327
 ---------------------------------------------------------------

 C:\Users\Administrator\Documents\data variabel diolah.XLS[xls5]: 3780
RECORDS.
```

BIOGRAFI PENULIS

Penulis bernama lengkap Mohammad Fajri, dengan nama panggilan Ipung. Lahir di Kabupaten Donggala, Sulawesi Tengah, pada tanggal 4 Maret 1990 dan merupakan anak pertama dari lima bersaudara. Penulis menempuh pendidikan formal di TK Al Khairaat Donggala (1993-1995), SD Negeri 3 Donggala (1995-2001), SMP Negeri 2 Donggala (2001-2004), dan SMA Negeri 1 Donggala (2004-2007).

Setelah lulus SMA, penulis melanjutkan studi pada Jurusan Matematika FMIPA Universitas Tadulako, Palu (2007-2012). Pada tahun 2013 penulis melanjutkan studi Magister di Program Pascasarjana Jurusan Statistika, FMIPA Institut Teknologi Sepuluh Nopember, Surabaya.

Saran dan kritik yang berhubungan dengan Tesis ini dapat ditujukan ke alamat email : mohammadfajri.nd@gmail.com