ANALISIS *MIXED EFFECT REGRESSION TREES* PADA DATA KONTINU BERKLUSTER

Skripsi
disusun sebagai salah satu syarat
untuk memperoleh gelar Sarjana Sains
Program Studi Matematika

oleh

Indana Lutfiani

4111415004

JURUSAN MATEMATIKA

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM

UNIVERSITAS NEGERI SEMARANG

2019

HALAMAN PENGESAHAN

Skripsi berjudul

Analisis *Mixed Effect Regression Trees* Pada Data Kontinu Berkluster

disusun oleh

Indana Lutfiani

4111415005

telah dipertahankan dihadapan sidang Panitia Ujian Skripsi FMIPA Universitas Negeri Semarang pada tanggal 30 Juli 2019 dan disahkan oleh Panitia Ujian.

Panitia:

Ketua

Dr. Sugianto, M.Si.
NIP. 196102191993031001

Sekretaris

Drs. Arief Agoestanto, M.Si.
NIP. 196807221993031005

Ketua Penguji

Dr. Scolastika Mariani, M.Si.
NIP. 196502101991022001

Anggota Penguji/
Penguji II

Drs. Sugiman, M.Si.
NIP.196401111989011001

Anggota Penguji/
Pembimbing I

Dr. Nur Karomah Dwidayati, M.Si.
NIP. 196605041990022001

PERNYATAAN

Dengan ini, saya

Nama : Indana Lutfiani
NIM : 4111415004
Program studi : Matematika

Menyatakan bahwa skripsi berjudul "Analisis *Mixed Effect Regression Trees* Pada Data Kontinu Berkluster" ini benar-benar karya saya sendiri bukan jiplakan dari karya orang lain atau pengutipan dengan cara-cara yang tidak sesuai dengan etika keilmuan yang berlaku baik sebagian atau seluruhnya. Pendapat atau temuan atau pihak lain yang terdapat dalam skripsi ini telah dikutip atau dirujuk berdasakan kode etik ilmiah. Atas pernyataan ini, saya secara pribadi siap menggung resiko/sanksi hukum yang dijatuhkan apabila ditemukan adanya pelanggaran terhadap etika keilmuan dalam karya ini.

Semarang, Agustus 2019

Indana Lutfiani

NIM. 4111415004

MOTO DAN PERSEMBAHAN

MOTO

Hidup hanyalah jalan, indah tidaknya tempat yang kita singgahi sepenuhnya berada pada pilihan kita, berusaha dan berdoa semoga senantiasa menjadi pengiring setiap jalan yang akan dilalui.

PERSEMBAHAN

Untuk Ibu Mudrikah, Bapak Abdul Rosyid, kakak Fina, keluarga di Kudus, sahabat, dan rekan yang selalu mendoakan dan ikut serta dalam perjalanan hidup yang bahagia ini.

PRAKATA

Segala puji dan syukur penulis ucapkan kehadirat Allah SWT atas segala limpahan rahmat-Nya sehingga penulis dapat menyelesaikan skripsi yang berjudul "Analisis *Mixed Effect Regression Trees* Pada Data Kontinu Berkluster". Skripsi ini disusun sebagai salah satu syarat meraih gelar Sarjana Sains pada Program Studi Matematika, Universitas Negeri Semarang. Shalawat serta salam disampaikan kepada junjungan kita Nabi Muhammad SAW, semoga mendapatkan syafaatnya di hari akhir nanti.

Penulis menyadari bahwa dalam penyusunan skripsi ini tidak terlepas dari bantuan dan bimbingan dari berbagai pihak. Untuk itu, penulis ingin menyampaikan terima kasih kepada:

1. Prof. Dr. Fathur Rokhman, M.Hum., Rektor Universitas Negeri Semarang.
2. Dr. Sugianto, M.Si., Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Negeri Semarang.
3. Drs, Nur Karomah Dwidayati, M.Si., Dosen Pembimbing yang telah memberikan bimbingan, arahan, dan saran kepada penulis dalam menyususn skripsi ini.
4. Dosen Penguji yang telah memberikan bimbingan, arahan, dan saran kepada penulis dalam menyusun skripsi ini.
5. Bapak dan Ibu dosen Jurusan Matematika, yang telah memberikan bimbingan dab ilmu kepada penulis selama menempuh pendidikan.
6. Teman-teman mahasiswa Program Studi Matematika, Universitas Negeri Semarang angkatan 2015, yang selalu berbagi rasa dalam suka dan duka, dan atas segala bantuan dan kerjasamnya dalam menempuh studi.
7. Teman-teman di Kos MHC gang Goda No.15, RT 7/ RW5, Banaran Gunungpati yang selalu menemani kesendirian dan kegabutan ini.
8. Teman-teman yang selalu bertanya kapan wisuda, sehingga memberi saya semangat untuk menyelesaikan skripsi ini

9. Semua pihak yang turut membantu dalam menyususn skripsi ini yang tidak dapat disebutkan namanya satu persatu.

Semoga skripsi ini dapat memberikan manfaat bagi penulis dan para pembaca. Terimakasih.

Semarang, Agustus 2019

Penulis

ABSTRAK

Lutfiani, Indana. 2019 Analisis *Mixed Effect Regression Trees* Pada Data Kontinu Berkluster. Skripsi, Jurusan Matematika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Negeri Semarang. Pembimbing Dr. Nur Karomah Dwidayati, M.Si.

Kata Kunci : *Mixed Effect Regression Trees (MERT), Standart Trees (STD), Predictive Mean Square Error (PMSE), Angka Kematian Ibu (AKI)*

Dilapangan terdapat berbagai macam tipe data salah satunya yaitu data berkluster. Data berkluster akan dianalisis menggunakan regresi, analisis regresi yang digunakan adalah dengan regresi pohon yang secara umum disebut dengan metode *Standart Tree* (STD), metode STD kurang optimal dalam menangani data berkluster karena dimungkinkan terdapat *missing data*, oleh sebab itu digunakanlah metode *Mixed Effect Regression Trees* (MERT) sebagai perpanjangan dari metode STD dengan penambahan algoritma *Expectation Maximation* (EM) untuk mengatasi *missing data.*

Penelitian ini bertujuan untuk mengetahui keoptimalan metode MERT dibandingkan dengan STD dilihat dari nilai *Predictive Mean Square Error* (PMSE). Data yang digunakan dalam studi kasus yaitu:

Y = Angka Kematian Ibu (AKI)
X_1 = Persentase Kunjungan Ibu Hamil dengan K1
X_2 = Persentase Kunjungan Ibu Hamil dengan K4
X_3 = Persentase Ibu Hamil Mendapat Fe_3
X_4 = Persentase Komplikasi Kebidanan yang Ditangani
X_5 = Persentase Persalinan Ditolong oleh Tenaga Kesehatan
Z_1 = 5 Provinsi di Pulau Jawa (Jateng, Jabar, Jabar, DIY, Banten)

dengan metode MERT dan STD akan diketahui faktor-faktor apa saja yang akan mempengaruhi tingkat Angka Kematian Ibu.

Metode yang digunakan adalah MERT dan STD dengan bantuan *software* R dan Ms.Excel. Data AKI sebanyak 113 dibagi menjadi 2 yaitu data *learning* sebanyak 56 dan data *testing* sebanyak 57. Analisis metode STD menggunakan *package* rpart dalam program R, sedangkan analisis metode MERT menggunakan *package* dari penelitian Ahlem Hajjem, yaitu modifikasi *package* rpart dengan menerapkan algoritma EM. Output dari metode MERT dan STD adalah suatu estimator untuk selanjutnya didapatkan nilai PMSE.

Berdasarkan hasil penelitian diperoleh Metode MERT lebih optimal dibanding metode STD dilihat dari hasil nilai PMSE.

DAFTAR ISI

HALAMAN SAMPUL i
HALAMAN PENGESAHAN ii
PERNYATAAN iii
MOTTO DAN PERSEMBAHAN iv
PRAKATA v
ABSTRAK vii
DAFTAR ISI viii
DAFTAR TABEL xii
DAFTAR GAMBAR xiv
DAFTAR LAMPIRAN xv
BAB I 1
1.1 Latar Belakang Masalah 1
1.2 Fokus Penelitian 4
1.3 Rumusan Masalah 4
1.4 Tujuan Penelitian 4
1.5 Manfaat Penelitian 4
1.5.1 Bagi Mahasiswa: 4
1.5.2 Bagi Pembaca: 5
1.6 Sistematika Penulisan 5
BAB II 7
2.1 Variabel Random 7
2.2 Matriks 8

2.3 Algoritma EM 8
2.4 Pohon Regresi (*Regression Tree*) 10
2.4.1 Pertumbuhan Pohon Regresi 11
2.4.2 Penghentian Pembentukan Pohon Regresi 13
2.4.3 Pemangkasan Pohon Regresi 13
2.4.4 Penentuan Ukuran Pohon Regresi Yang Optimal 15
2.5 *Linear Mixed Effect* (LME) 15
2.6 *Mixed Effects Regression Trees* (MERT) 16
2.7 *Software* R 17
2.7.1 Fungsi dalam *Software* R 18
2.7.2 *Package* dalam *Software* R 18
2.7.3 Jenis-jenis data objek pada *Software* R 19
2.8 Kerangka Berpikir 22
BAB III 23
3.1 Studi Literatur 23
3.2 Variabel Penelitian 23
3.2.1 Deskripsi Angka Kematian Ibu (AKI) 24
3.2.2 Deskripsi Kunjungan Ibu hamil dengan K1 24
3.2.3 Deskripsi Kunjungan Ibu hamil dengan K4 24
3.2.4 Deskripsi ibu hamil denga Fe3 25
3.2.5 Deskripsi komplikasi kebidanan yang ditangani 25
3.2.6 Deskripsi persalinan ditolong oleh tenaga kesehatan 25
3.2.7 Deskripsi tentang Provinsi 26

3.3 Metode Pengumpulan Data 26
3.4 Metode 26
3.5 Diagram Alir Penelitian 28
3.6 Kesimpulan..... 29
BAB IV 30
4.1 HASIL 30
4.1.1 Deskripsi data pada studi kasus angka kematian ibu (AKI) 30
4.1.2 Analisis *Standart Tree* 30
4.1.2.1 Pertumbuhan Pohon Regresi 30
4.1.2.2 Pemangkasan Pohon Regresi..... 31
4.1.2.3 Penentuan Ukuran Pohon Regresi Yang Optimal 31
4.1.2.4 PMSE (Predictive Mean Square Error)..... 31
4.1.2.5 PMSE (Predictive Mean Square Error) untuk data learning 32
4.1.2.6 PMSE (*Predictive Mean Square Error*) untuk data testing 32
4.1.3 Analisis *Mixed Effect Regression Tree* 32
4.1.3.1 Pertumbuhan Pohon Regresi..... 32
4.1.3.2 Pemangkasan Pohon Regresi 33
4.1.3.3 Penentuan Ukuran Pohon Regresi yang Optimal..... 34
4.1.3.4 PMSE (Predictive Mean Square Error) 34
4.1.3.5 PMSE (Predictive Mean Square Error) untuk data learning 34
4.1.3.6 PMSE (Predictive Mean Square Error) untuk data testing..... 35
4.1.4 Perbandingan Metode *Standart Tree* Dengan Metode *Mixed Effect* Regression *Trees* 35

4.2 PEMBAHASAN 36
4.2.1 Hasil Penerapan Metode *Standart Tree* Pada Data Kecil 36
4.2.1.1 Analisis Numerik Metode *Standar Tree* Pada Data Kecil..... 36
4.2.1.2 Hasil penerapan metode *Standart Tree* pada data kecil dengan bantuan program R 45
4.2.1.3 Analisis numerik metode *Mixed Effect Regression Trees* pada data kecil 45
BAB V 46
5.1 SIMPULAN 46
5.2 SARAN 46
DAFTAR PUSTAKA 48

DAFTAR TABEL

Tabel 1.1 Tiga Peringkat Negara dengan Angka Kematian Ibu di ASEAN 3
Tabel 4.1 CP dari pohon awal 31
Tabel 4.2 Perbandingan MERT dan STD 35
Tabel 4.3 Data Kecil dengan Variabel *x* dan *y* 36
Tabel 4.4 Varian untuk y $(s2)$ 36
Tabel 4.5 Node yang dihasilkan dari x 36
Tabel 4.6 Titik-titik batas 37
Tabel 4.7 Varian dan Varian Total untuk $x < 6{,}5$ dan $x \geq 6{,}5$ 37
Tabel 4.8 Varian dan Varian Total untuk $x < 5{,}5$ dan $x \geq 5{,}5$ 37
Tabel 4.9 Varian dan Varian Total untuk $x < 4{,}5$ dan $x \geq 4{,}5$ 38
Tabel 4.10 Varian dan Varian Total untuk $x < 3$ dan $x \geq 3$ 38
Tabel 4.11 Varian total untuk setiap titik 38
Tabel 4.12 Selisih antara varian awal dengan varian total 39
Tabel 4.13 Data baru dengan variabel y dan $x \geq 3$ 39
Tabel 4.14 Varian untuk y $(s2)$ 39
Tabel 4.15 Node yang dihasilkan dari x 40
Tabel 4.16 Titik-titik batas 40
Tabel 4.17 Varian dan Varian Total untuk $x < 6{,}5$ dan $x \geq 6{,}5$ 40
Tabel 4.18 Varian dan Varian Total untuk $x < 5{,}5$ dan $x \geq 5{,}5$ 41
Tabel 4.19 Varian dan Varian Total untuk $x < 4{,}5$ dan $x \geq 4{,}5$ 41
Tabel 4.20 Varian total untuk setiap titik 41
Tabel 4.21 selisih antara varian awal dengan varian total 41
Tabel 4.22 Data baru untuk $x < 6{,}5$ 42
Tabel 4.23 varian untuk y $(s2)$ 42
Tabel 4.24 Node yang dihasilkan dari x 42
Tabel 4.25 Titik-titik batas 43
Tabel 4.26 Varian dan Varian Total untuk $x < 5{,}5$ dan $x \geq 5{,}5$ 43

Tabel 4.27 Varian dan Varian Total untuk $x < 4{,}5$ dan $x \geq 4{,}5$ 43
Tabel 4.28 Varian dan Varian Total untuk $x < 3$ dan $x \geq 3$ 44
Tabel 4.29 Varian total untuk setiap titik 44
Tabel 4.30 Selisih antara varian awal dan varian total 44

DAFTAR GAMBAR

Gambar 2.1 Kerangka Berpikir 22
Gambar 3.1 Flowchart Penelitian 28
Gambar 4.1 Pohon Awal *Standart Tree* 30
Gambar 4.2 Pohon Awal *Mixed Effect Regression Tree* 33
Gambar 4.3 Output "*pruned.tree*" pada iterasi terakhir (iterasi ke 144) 34

DAFTAR LAMPIRAN

Lampiran 1 Tabel data Angka Kematian Ibu pada Tahun 2017 53
Lampiran 2 Pembagian data *learning* dan data *testing* pada Tahun 2017 57
Lampiran 3 Syntaks dan output *Standart Tree* (STD) dengan data *learning*..... 61
Lampiran 4 Tabel prediksi, residual, dan error pada data *learning* (STD)..... 66
Lampiran 5 Syntaks dan output analisis *Mixed Effect Regression Tree* (MERT)..... 70
Lampiran 6 Tabel prediksi, residual, dan error pada data *learning* (MERT) 97
Lampiran 7 Syntaks dan hasil output *regression tree* pada data kecil..... 101

BAB I
PENDAHULUAN

1.1 Latar Belakang Masalah

Dewasa ini terdapat beragam metode yang dapat digunakan untuk menganalisis data, salah satunya yaitu dengan regresi. Analisis regresi dapat diterapkan jika dipunyai data dengan dua atau lebih variabel yang bertujuan untuk mempelajari cara bagaimana variabel-variabel tersebut berhubungan, hubungan tersebut umumnya dinyatakan dalam bentuk persamaan matematik yang menyatakan hubungan fungsional antara variabel-variabel (Sudjana, 2002), dengan kata lain analisis Regresi digunakan untuk mengetahui pengaruh antara variabel bebas (prediktor) terhadap variabel terikat (respon).

Regresi dapat diterapkan pada berbagai macam tipe data. Tipe data yang dulunya sulit untuk dianalisa secara regresi dikarenakan adanya faktor-faktor yang mempengaruhinya, kini oleh peneliti hampir semua data bisa dianalisis secara regresi. Regresi umumnya diterapkan untuk data tidak berkluster, sedangkan untuk data berkluster tidak banyak digunakan dalam statistika terlebih pada regresi. Data berkluster merupakan data yang dikelompokkan berdasarkan kesamaan karakteristik antar objek-objek yang diamati seperti contohnya data mahasiswa di Indonesia dengan klusternya adalah Universitas.

Pada umumnya data tidak berkluster dilakukan analisis klustering untuk mengklasifikasi objek sehingga setiap objek yang memiliki sifat yang mirip akan mengelompok ke dalam satu kluster yang sama, namun untuk data berkluster belum begitu banyak penelitian lanjutan terlebih untuk penelitian dengan menerapkan analisis regresi, hal itu dikarenakan adanya variabel Z sebagai identitas kluster untuk setiap data.

Pada kesempatan kali ini penulis ingin membahas regresi dengan pendekatan statistika nonparametrik. Statistika nonparametrik itu sendiri merupakan uji statistik yang tidak memerlukan adanya asumsi-asumsi mengenai sebaran data populasi (Djarwanto, 1991). Analisis yang dipilih yaitu metode Pohon regresi yang secara umum disebut *standart tree*.

Metode *standart tree* mempunyai kelebihan yaitu dapat mendeteksi secara otomatis kemungkinan interaksi yang signifikan antara kovariat, dan model yang diusulkan mudah diinterpretasi sekaligus ditampilkan secara grafis, namun saat data yang dianalisis merupakan data berkluster, metode *standart tree* tidak lagi bisa digunakan karena dimungkinkan setiap simpul pohon dapat mencakup pengamatan yang termasuk dalam kelompok yang berbeda dengan kata lain dimungkinkan adanya *missing data* oleh sebab itu (Hajjem, 2010) mengusulkan metode *Mixed Effect Regression Tree* untuk menyelesaikan permasalahan *missing data* yang tidak bisa diselesaikan dengan *standart tree.*

Metode *Mixed Effect Regression Tree* merupakan perluasan dari *Standart Tree* dengan mengaplikasikan Algoritma *Expectation Maximation* (EM) dalam menyelesaikan kasus *missing data*. *Missing data* merupakan vaktor penentu kualitas suatu data, karena jika jumlah data yang missing cukup besar dapat mempengaruhi keakuratan dan kualitas kesimpulan yang dihasilkan (Sirait, 2013).

Studi Kasus yang digunakan adalah pengaruh faktor-faktor Angka Kematian Ibu (AKI) di 5 Provinsi di Pulau Jawa. Menurut WHO (Jurnal Matematika dan Pendidikan Matematika, 2016:2) kematian ibu adalah kematian dari setiap wanita selama kehamilan, bersalin atau dalam 42 hari sesudah berakhirnya kehamilan oleh sebab apapun, tanpa melihat usia dan lokasi kehamilan oleh setiap penyebab yang berhubungan dengan atau diperberat oleh kehamilan atau penanganannya tetapi bukan oleh kecelakaan.

Data Angka Kematian Ibu digunakan sebagai penelitian dikarenakan Derajat kesehatan suatu bangsa dapat dinilai dengan menetapkan beberapa alat ukur yaitu usia harapan hidup, Angka Kematian Ibu (AKI), Angka Kematian Bayi (AKB), angka-angka tersebut dapat menggambarkan tingkat kemajuan suatu bangsa (Helmizar, 2014), oleh sebab itu Angka Kematian Ibu (AKI) dan Angka Kematian Bayi (AKB) menjadi tujuan dalam pembangunan Millenium Development Goals (MDGs) yaitu menurunkan Angka Kematian Ibu (AKI) pada tahun 2015 menjadi 102 per 100.000 kelahiran hidup (Arkandi dan Winahju, 2015). Faktanya di tahun 2015 Angka Kematian Ibu mencapai 305 per 100.000 kelahiran hidup dan menjadikan Indonesia menempati peringkat kedua sebagai

Negara dengan Angka Kematian Ibu tertinggi di ASEAN, data dapat dilihat pada tabel di bawah.

Tabel 1.1 Tiga Peringkat Negara dengan Angka Kematian Ibu di ASEAN

No.	Negara	Tahun						Target
		1950	1995	2000	2005	2010	2015	
1.	Laos	680	605	530	405	280	357	170
2.	Indonesia	390	353	320	268	228	305	98
3.	Philipina	164	170	176	144	129	221	41

Catatan: nilai diatas per 100.000 kelahiran hidup

Sumber: ASEAN Statistical Report On Millenium Development Goals 2017

Tabel 1.1 menjelaskan bahwa Angka Kematian Ibu di Indonesia sangat jauh dari target yang diharapkan yaitu ditemukannya kematian ibu sejumlah 305 per 100.000 kelahiran hidup padahal ditargetkan pada tahun 2015 hanya mencapai 98 per 100.000 kelahiran hidup. Angka Kematian Ibu yang berangsur turun hingga tahun 2000 dan kembali mengalami kenaikan pada tahun 2005 hingga 2015.

Menurut Saddiyah Rangkuti (Jurnal Ilmiah Research, 2015:3) faktor penyebab kematian ibu dapat disebabkan oleh 2 faktor yaitu penyebab langsung dan penyebab tidak langsung. Penyebab langsung berupa pendarahan, eklampsia, partus lama, komplikasi aborsi, dan infeksi. Sedangkan penyebab tidak langsung berupa status perempuan dalam keluarga, keberadaan anak, sosial budaya, pendidikan, sosial ekonomi, dan geografis daerah.

Data yang digunakan merupakan data berkluster yang diambil dari data Angka Kematian Ibu (AKI) di Pulau Jawa pada tahun 2017. Penggunaan data berkluster biasanya mencakup dua komponen yaitu komponen efek tetap dan kompoen efek campuran, untuk menangani komponen dari efek campuran penggunaan metode *standart tree* tidak dapat menyelesaikan permasalahan yang ada, sebagai solusinya digunakanlah *metode Mixed Effect Regression Trees* sebagai penyelesaian terhadap masalah terkait komponen efek campuran, sehingga dari data berkluster tersebut dapat diambil kesimpulan tentang faktor-

faktor yang mempengaruhi Angka Kematian Ibu (AKI) di 5 Provinsi di Pulau Jawa.

1.2 Fokus Penelitian

Fokus penelitian ini adalah menganalisis metode *Mixed Effect Regression Trees* dan membandingkannya dengan *Standart Trees* untuk menentukan metode yang lebih optimal terhadap data berkluster, dengan studi kasus faktor-faktor yang mempengaruhi Angka Kematian Ibu (AKI) di 5 Provinsi di Pulau Jawa.

1.3 Rumusan Masalah

Berdasarkan latar belakang yang telah dijelaskan, maka rumusan masalah yang termuat dalam penelitian ini adalah:

1. Manakah yang lebih optimal antara metode *Mixed Effect Regression Trees* dengan metode *Standart Tree*; dan
2. Faktor-faktor apa saja yang mempengaruhi Angka Kematian Ibu (AKI) berdasarkan analisis dengan metode *Mixed Effect Regression Trees* dan metode *standart tree.*

1.4 Tujuan Penelitian

Skripsi ini disusun sebagai salah satu syarat untuk memperoleh gelar sarjana S1 Program Studi Matematika, Jurusan Matematika, Fakultas Matematika dan Ilmu Pengetahuan Alam. Sedangkan tujuan untuk penulisan skripsi ini adalah sebagai berikut:

1. Mengetahui metode yang lebih optimal antara metode *Mixed Effect Regression Trees* dengan metode *Standart Tree;* dan
2. Mengetahui faktor-faktor yang mempengaruhi Angka Kematian Ibu (AKI) di 5 Provinsi di Pulau Jawa dengan analisis *Mixed Effect Regression Trees* dan metode *Standart Tree*.

1.5 Manfaat Penelitian

Manfaat yang diperoleh dalam penelitian ini, diantaranya:

1.5.1 Bagi Mahasiswa:

Penelitian ini dapat memberikan Manfaat Kepada Mahasiswa yaitu:

1. Mahasiswa dapat memahami analisis regresi pada data berkluster yang diaplikasikan pada metode *Standart Tree*;

2. Mahasiswa dapat memahami analisis regresi pada data berkluster yang diaplikasikan pada metode *Mixed Effect Regression Tree*; dan
3. Mahasiswa dapat membandingkan metode yang lebih efektif dalam menerapkan data berkuster dengan studi kasus pengaruh angka kematian ibu antara metode *Standart Tree* dengan *Mixed Effect Regression Tree*.

1.5.2 Bagi Pembaca:

Penelitian ini dapat memberikan Manfaat Kepada Pembaca yaitu:

1. Menambah dan memperkaya kajian pustaka pada Prodi Matematika khususnya pada ilmu statistika;
2. Menambah Pengetahuan bahwa analisis regresi bisa dilakukan pada berbagai macam tipe data salah satunya pada data berkluster; dan
3. Menambah pengetahuan tentang pengaruh angka kematian ibu berdasarkan hasil analisis dengan model terbaik.

1.6 Sistematika Penulisan

Sistematika penulisan skripsi ini adalah sebagai berikut:

1) Bab 1 Pendahuluan

   Bab ini berisi latar belakang masalah, fokus penelitian, rumsan masalah, tujuan penelitian, manfaat penelitian, serta sistematika penulisan.

2) Bab 2 Tinjauan Pustaka

   Tinjauan pustaka berisi kajian teori dan hasil-hasil penelitian terdahulu yang menjadi kerangka pikir penyelesaian masalah penelitian yang disajikan ke dalam beberapa sub-bab.

3) Bab 3 Metode Penelitian

   Bab ini berisi studi literatur, variabel penelitian, metode pengumpulan data, metode penelitian, diagram alir penelitian, dan kesimpulan.

4) Bab 4 Hasil dan Pembahasan

   Bagian ini berisi analisis dari hasil pengolahan contoh data dan pembahasan mengenai prosedur bagaimana cara Analisis menggunakan metode *Standart Tree* dan *Mixed Effect Regression Trees* pada data berkluster dengan studi kasus Pengaruh Angka Kematian Ibu (AKI).

5) Bab 5 Penutup

Pada bab ini berisi simpulan dari hasil penelitian serta saran-saran sebagai masukan untuk pengembangan penelitian selanjutnya.

BAB II
TINJAUAN PUSTAKA

2.1 Variabel Random

X dikatakan variabel random, jika X adalah fungsi yang didefinisikan pada ruang sampel S, dengan X merupakan fungsi bernilai real. X dapat ditulis sebagai $X(c) = x$, c adalah setiap hasil yang mungkin di S (Bain and Engelhardt, 1991,53). Spiegel dkk memisalkan untuk setiap titik ruang sampel yang berupa angka, kemudian memiliki fungsi yang didefinisikan pada ruang sampel, fungsi ini disebut variabel random atau lebih tepatnya fungsi acak yang biasanya dilambangkan dengan huruf kapital X atau Y (Spiegel dkk, 2004).

Huruf kapital seperti X, Y, dan Z akan digunakan untuk menunjukkan variabel random, sedangkan huruf a,b,c,... akan digunakan untuk menunjukkan kemungkinan nilai yang dapat diperoleh oleh variabel random yang sesuai, dalam matematika perlu untuk membatasi jenis fungsi yang dianggap variabel random.

Variabel random X disebut variabel random diskrit jika himpunan semua nilai yang mungkin dari variable random X merupkan himpunan yang dapat dihitung (countable), dengan fungsinya yaitu $f(x) = P[X = x]$, $x = x_1, x_2, x_3, \ldots$, yang menetapkan peluang untuk setiap nilai X yang mungkin, X akan disebut sebagai fungsi kepadatan peluang diskrit, sehingga fungsi distribusi kumulatifnya dapat dituliskan sebagai berikut:

$$F(x) = P[X \leq x] = \sum_{x_i \leq x} f(x_i) \tag{2.1}$$

Variabel random X disebut variabel random kontinu jika fungsi $f(x)$ disebut fungsi kepadatan peluang dari X , sedemikian hingga fungsi distribusi kumulatifnya dapat direpresentasikan sebagai $F(x) = \int_{-\infty}^{x} f(t)\, dt$, Sehingga fungsi kepadatan peluang untuk variabel random kontinu dapat dituliskan sebagai berikut:

$$f(x) = \frac{d}{dx} F(x) = F'(x). \tag{2.2}$$

2.2 Matriks

Matriks merupakan susunan segi empat siku-siku dari bilangan-bilangan, dengan bilangan-bilangan dalam susunan tersebut dinamakan entri dalam matriks (Howard Anton, 1997) Suatu matriks berukuran m x n atau matriks m x n adalah suatu jajaran bilangan berbentuk persegi panjang yang terdiri dari m baris dan n kolom.

Dalam menyatakan suatu matriks biasanya digunakan huruf kapital atau huruf besar dalam susunan Alphabet misal: A, B, dan C. Sedangkan dalam menyatakan unsur atau elemen atau anggota digunakan huruf kecil dalam susunan Alphabet, misal: a, b, dan c.

Dalam menunjukkan sebuah matriks kadang kala digunakan sepasang tanda kurung ( ), dan garis tegak ganda ‖ ‖. Selanjutnya akan dipakai penulisan sepasang kurung siku. Pada saatnya matriks (1) akan disebut " matriks $[a_{ij}]$, $m \times n$" atau " matriks $A = [a_{ij}]$, "$m \times n$" . Bilamana ukuran (ordo) sudah dikembangkan, cukup dituliskan " matriks A" saja.

Trace Matriks merupakan jumlahan elemen-elemen pada diagonal utama matriks (A.K. Gupta, D.K. Nagar, 1999). Misalkan $A = [a_{ij}]$ matriks bujur sangkar untuk $n \times n$, Trace dari matriks A, dinotasikan $Tr(A)$, didefinisikan sebagai jumlahan dari elemen-elemen diagonal utama A, yaitu

$$Tr(A) = a_{1,1} + a_{2,2} + \cdots a_{n,n} = \sum_{i=1}^{n} a_{i,i} \qquad (2.3)$$

2.3 Algoritma EM

Algoritma EM adalah algoritma untuk menduga suatu parameter dalam suatu fungsi dengan menggunakan MLE, dimana fungsi tersebut mengandung data yang tidak lengkap (Hoog, 2005). Definisi yang sama tentang algoritma EM oleh (N Dwidayati, 2013) bahwa pada hakekatnya Algoritma EM merupakan komputasi iterasi dari *Maximum Likelihood Estimators* (MLE) jika observasi dikategorikan sebagai *incomplete*. Algoritma EM merupakan teknik komputasi yang sering dignakan dalam komputasi statistika (Gentle, 2002).

N Dwidayati & Zaenuri (2017) menyatakan bahwa Algoritma EM diformulasikan untuk menangani data yang hilang yaitu dengan:

1. Mengganti nilai yang hilang dengan nilai yang diestimasi;
2. Mengestimasi parameter;
3. Reestimate nilai yang hilang dengan asumsi estimasi parameter baru sudah benar; dan
4. Reestimate parameter dan seterusnya melalui iterasi hingga konvergen.

Tiap iterasi dari algoritma EM terdiri dari dua proses yaitu tahap-E dan tahap-M. Tahap pendugaan, tahap-E menghitung dugaan kemungkinan dengan mempertimbangkan peubah laten jika peubah tersebut diamati. Tahap maksimisasi, tahap-M menghitung dugaan kemungkinan maksimum parameter dengan maksimisasi dugaan kemungkinan pada tahap-E.

Parameter didapatkan pada tahap-M selanjutnya digunakan untuk tahap-E berikutnya, dan proses tersebut dilakukan secara berulang, Selanjutnya sebaran bersyarat data hilang jika diketahui data amatan dapat dinyatakan sebagai :

$$p(z,y|\theta) = \frac{p(y,z|\theta)}{p(y|\theta)} = \frac{p(y|z,\theta)p(z|\theta)}{\int p(y|\hat{z},\theta)p(\hat{z},\theta)d\hat{z}} \tag{2.4}$$

Dengan

y =nilai peubah teramati (menunjukkan data tidak lengkap)

z = nilai peubah tidak teramati (menunjukkan data hilang)

$y + z$ = nilai peubah dengan data lengkap (y dan z secara bersama-sama membentuk data lengkap)

p = fungsi kepadatan peluang bersama dari data lengkap dengan parameter diberikan oleh vektor θ: $p(y, z|\theta)$

Algoritma EM secara iteratif meningkatkan dugaan awal θ_0 dengan mencari dugaan baru θ_1, θ_2 dan seterusnya. Tiap tahapan yang menurunkan θ_{n+1} dari θ_n dengan bentuk sebagai berikut:

$$\theta_{n+1} = ar\,g\,max_{\theta}\,Q(\theta) \tag{2.5}$$

Dimana $Q(\theta)$ adalah nilai harapan *log-likelihood*. Q diberikan oleh

$$Q(\theta) = \sum_z p(z|y, \theta_n) \log p(y, z|\theta) \quad (2.6)$$

Atau secara umum

$$Q(\theta) = E_z[\log p(y, z|\theta)|y] \quad (2.7)$$

Dengan kata lain, θ_{n+1}adalah nilai yang memaksimumkan (M) dugaan bersyarat (E) dari *log-likelihood* data lengkap jika diketahui peubah teramati pada nilai parameter sebelumnya. Pendugaan $Q(\theta)$ pada kasus kontinu diberikan oleh:

$$\begin{aligned} Q(\theta) &= E_z[\log p(y, z|\theta)|y] \\ &= \int_{-\infty}^{\infty} p(z|y, \theta_n) \log p(y, z|\theta)\, dz \end{aligned} \quad (2.8)$$

2.4 Pohon Regresi (*Regression Tree*)

CART adalah metode yang terdiri dari dua analisis, yaitu *classification trees* dan *regression trees*. Jika variabel dependen yang dimiliki bertipe kategorik maka CART menghasilkan pohon klasifikasi (*classification trees*), sedangkan jika variabel dependen yang dimiliki bertipe kontinu atau numeric (interval atau rasio) maka CART menghasilkan pohon regresi (*regression trees*) (Mardika, dkk, 2016).

Sama halnya dengan metode regresi biasa, pohon regresi juga menjelaskan bagaimana hubungan antara peubah respon dan peubah-peubah penjelasnya. Perbedaannya adalah bahwa pada metode pohon regresi, pengaruh peubah penjelas serta pendugaan responnya dilakukan pada kelompok-kelompok pengamatan yang ditentukan berdasarkan peubah-peubah penjelas, sehingga interpretasi hasil dari metode ini lebih mudah dilakukan. Hal ini karena identifikasi pengaruh dari peubah penjelas dari pohon regresi dilakukan dalam masing-masing subgrup data bukan dalam keseluruhan data seperti halnya regresi biasa (Wieta B, 2007). Pohon regresi secara umum bisa disebut dengan *standart tree.*

Metode berbasis pohon (*standart tree*) digunakan untuk menemukan sub kelompok dalam data yang berbeda sehubungan dengan memodelkan parameter. Dalam beberapa aplikasi sangat wajar untuk menjaga beberapa parameter tetap umum untuk semua pengamatan. Penerapan dari pohon model berbasis pohon

(*standart tree*) hanya bisa berhubungan dengan perencanaan bilamana semua parameter tergantung pada sub kelompok (Heidi, 2016:1).

Metode berbasis pohon sangat berguna untuk tujuan eksplorasi karena mereka dapat menangani banyak variabel prediktor potensial dan secara otomatis dapat mendeteksi (tingkat tinggi) interaksi antara variabel prediktor (Strobl, Malley, & Tutz, 2009).

Metode *standart tree* adalah teknik data mining klasik yang mempunyai banyak kelebihan dibandingkan dengan metode parametrik yaitu:

1. Dapat mendeteksi secara otomatis kemungkinan interaksi yang signifikan antara kovariat,
2. Dapat mengusulkan model yang mudah diinterpretasi yang dapat ditampilkan secara grafis,

Namun dipihak lain metode *standart tree* memiliki kekurangan, yaitu:

1. Tidak dapat mengesistensikan pengamatan terhadap kovariat (yaitu variasi waktu) untuk menjadi kandidat dalam proses pemisahan akibatnya ketidak acakan atau efek spesifik subjek dari kovariat ini diperbolehkan
2. Semua pengulangan pengamatan yang diulang dari subjek yang diberikan tidak dapat dipisah dinode yang berbeda.

2.4.1 Pertumbuhan Pohon Regresi

Pohon regresi terbentuk dari hasil pemilahan data pada setiap simpul induk menjadi dua simpul anak, dengan aturan pemilahan yaitu:

1. Tiap Pemilahan bergantung pada satu nilai pemilah yang hanya berasal dari satu variabel prediktor.
2. Untuk variabel prediktor kontinu X_j, pemilahan berasal dari $x_j \leq c_i$ untuk $x_j \leq \Re^j$, dengan $\Re^j$ merpakan ruang sampel dari variabel X_j, jika ruang sampelnya berukuran N dan terdapat sebanyak-banyaknya n nilai amatan berbeda pada variabel X_j, maka akan terdapat sebanyak $n-1$ pemilahan yang berbeda yang dibentuk oleh gugus $x_j \leq c_i$ dengan $i = 1,2,\dots,n-1$ dan c_i adalah nilai tengah antara dua nilai amatan variabel X_j berukuran berbeda.

Pohon regresi dibentuk melalui suatu pemilahan rekursif berdasarkan aturan pemilahan tersebut (Luqman, 2015). Proses pemilahan dilakukan pada tiap simpul untuk membentuk pohon regresi dengan aturan yaitu:

1. Mencari semua kemungkinan pemilahan pada tiap variabel prediktor berdasarkan aturan pemilahan tersebut.
2. Memilih pemilahan yang terbaik pada masing-masing variabel prediktor, kemudian melakukan pemilahan-pemilahan terbaik dari kumpulan pemilahan terbaik tersebut.

Pemilahan terbaik adalah pemilahan yang memaksimumkan ukuran kehomogenan di dalam masing-masing simpul anak relatif terhadap simpul induknya dan yang memaksimumkan ukuran pemisahan antara kedua simpul anak tersebut. Pemilahan terbaik dihitung berdasarkan selisih selisih jumlah kuadrat deviasi antara simpul induk dengan kedua simpul anak pemilahnya. Selisih terbesar akan dijadikan pemilah terbaik. Misalkan diketahui simpul t berisi sampel $\{(\ x_{n(t)}, y_{n(t)})\}$, sedangkan $n(t)$ adalah jumlah amatan dalam simpul t maka rataan respon dalam simpul t adalah:

$$\bar{y}(t) = \frac{1}{n(t)} \sum_{x_n \in t} y_{n(t)} \tag{2.9}$$

Dengan $y_{n(t)}$= nilai individu atau amatan variabel respon dalam simpul t

$n(t)$= jumlah amatan dalam simpul t

Sehingga jumlah kuadrat deviasi yang digunakan sebagai kriteria kehomogenan pada suatu simpul t adalah:

$$R(T) = \sum_{x_{n(t)} \in t} \left(y_{n(t)} - \bar{y}(t) \right)^2 \tag{2.10}$$

Jika terdapat pemilah s yang memilih t menjadi simpul anak kiri t_L dan simpul anak kanan t_R, maka kriteria selisih jumlah kuadrat deviasi adalah:

$$\Delta R(s,t) = R(t) - R(t_L) - R(t_R) \tag{2.11}$$

Dengan $R(t)$ = jumlah kuadrat deviasi suat simpul t

$R(t_L)$ = jumlah kuadrat deviasi suatu simpul anak kiri t_L

$R(t_R)$ = jumlah kuadrat deviasi suatu simpul anak kanan t_R

Pemilah terbaik s* dari t adalah pemilahan pada S yang sedemikian hingga

$$\Delta R(s^*, t) = \max_{s \in S} \Delta R(s, t) \quad ` \quad (2.12)$$

Dengan S merupakan gugus yang berisi semua kemungkinan pemilahan. Pohon regresi dibentuk melalui pemilahan simpul secara rekursif, yaitu dengan memaksimumkan fungsi $\Delta R(s, t)$.

2.4.2 Penghentian Pembentukan Pohon Regresi

Proses pohon regresi akan berhenti apabila sudah tidak dimungkinkan lagi dilakukannya proses pemilhan. Proses pemilahan akan berhenti apabila hanya terdapat satu amatan yang ada didalam sebuah simpul yang merupakan anggota nilai respon yang relatif homogen. Pohon regresi yang terbentuk sebagai hasil dari proses ini dinamakan regresi maksimal (T_{max}).

2.4.3 Pemangkasan Pohon Regresi

Pohon regresi yang dibentuk melalui proses pemilahan secara rekursif akan berukuran sangat besar. Hal ini disebabkan karena aturan penghentian yang digunakan hanya berdasarkan banyaknya amatan pada simpul terminal atau besarnya penurunan tingkat keragaman dalam tiap simpul anak hasil pemilahan. Semakin banyak pemilahan yang dilakukan maka tingkat kesalahan prediksi juga akan semakin kecil. Namun pohon regresi yang terbesar atau maksimal terlalu sulit untuk dipahami dan menyebabkan *overfitting* untuk data baru. Masalah tersebut diatasi dengan pemangkasan pada pohon regresi maksimal untuk mendapatkan pohon regresi dengan ukuran yang optimal.

Langkah awal pemangkasan dilakukan terhadap T_1, yaitu suatu sub pohon daro pohon terbesar T_{max}, untuk mendapakan T_1 dari T_{max} diambil t_L dan t_R yang merupakan simpul anak kiri dan simpul anak kanan dari T_{max} yang dihasilkan dari pemilahan pada setiap simpul induk t. Karena $R\,(t) \geq R\,(t_L) + R\,(t_R)$, maka ketika terdapat dua simpul anak dan simpl induk yang memenuhi persamaan $R\,(t) = R\,(t_L) + R\,(t_R)$, simpl anak t_L dan t_R tersebut dipangkas. Proses ini diulangi sampai tidak memungkinkan lagi dilakukam pemangkasan. Proses pemangkasan kompleksitas kesalahan terkecil adalah pemotongan hubungan terlemah pada pohon regresi. Untuk sembarang T_t yang merupakan cabang dari T_1, besar rataan kuadrat deviasi didefinisikan sebagai

$$R(T_t) = \sum_{t \in \tilde{T}_t} R(t') \tag{2.13}$$

Dengan $R(t')$= rataan kuadrat deviasi pada simpul t

$$R(t') = \frac{1}{n(t)} \sum_{x_{n(t)} \in t} \left(y_{n(t)} - \bar{y}(t) \right)^2 \tag{2.14}$$

Untuk setiap simpul $t \in T_1$, didefinisikan $\{t\}$ sebagai suatu sub cabang dari T_t yang hanya terdiri dari satu simpul. Ukran kompleksitas kesalahan dari sub cabang $\{t\}$ adalah:

$$R_\alpha(\{t\}) = R(t') + \alpha \tag{2.15}$$

Dan ukuran kompleksitas kesalahan dari cabang atau pohon T_t adalah

$$R_\alpha(\{T_t\}) = R(T_t) + \alpha|\tilde{T}_t| \tag{2.16}$$

Dengan α= parameter kompleksitas mengenai keslahan bagi penambahan satu simpul akhir pada pohon T_t

$|\tilde{T}_t|$= banyaknya simpul akhir yang dimiliki pohon T_t

Nilai α merupakan suatu parameter kompleksitas mengenai kesalahan bagi penambahan satu simpul akhir pada pohon T_t. Semakin besar nilai α maka ukuran pohon yang dihasilkan akan kecil, sebaliknya jika nilai α kecil maka kompleksitas simpulnya juga kecil sehingga ukuran pohon yang dihasilkan akan besar. Contoh sebuah T_{max} memiliki simpul akhir yang berisi hanya satu objek atau homogen sehingga nilai α untuk T_{max} adalah 0. Nilai α akan terus meningkat selama proses pemangkasan berlangsung dan akan mencapai nilai terbesar pada saat simpul akhir = simpul akar.

Nilai kompleksitas pemangkasan menentukan pohon bagian $(T_t)(\alpha)$ yang meminimumkan $R_\alpha(T_t)$ pada seluruh pohon bagian untuk setiap nilai α. Nilai parameter kompleksitas α akan secara perlahan meningkat selama proses pemangkasan. Selanjutnya pencarian pohon bagian $T(\alpha) < T_{max}$ yang dapat meminimumkan $R_\alpha(T_t)$ yaitu:

$$R_\alpha((T_t(\alpha))) = \min_{T < T_{max}} R_\alpha(T_t) \tag{2.17}$$

Hasil dari proses pemangkasan adalah berupa deretan sub pohon dengan ukuran yang semakin mengecil, yaitu $T_1 > T_2 > \cdots > \{t_1\}$, artinya pohon T_1 adalah induk bagi pohon T_2, pohon T_2 adalah induk dari pohon T_3, demikian

setersnya dengan deretan α dalam urutan meningkat, yaitu $\{\alpha_1, \alpha_{2,\dots}\alpha_K\}$ dengan $\alpha_k < \alpha_{k+1}$ untuk $k \geq 1$, dan $\alpha_k = 0$ untuk $k = 1$.

2.4.4 Penentuan Ukuran Pohon Regresi Yang Optimal

Ukuran pohon regresi yang besar akan menyebabkan nilai kompleksitas kesalahan yang tinggi, tetapi semakin besar pohon regresi maka tingkat kesalahan prediksinya juga akan semakin kecil, sehingga perlu dipilij pohon regresi optimal yang berukuran sederhana tetapi juga memberikan nilai kesalahan prediksi yang cukup kecil. Pohon regresi optimal dinotasikan dengan T_k yaitu sebuah sub pohon terkecil dari T_{max} yang meminimumkan tingkat kesalahan prediksi. Ada beberapa cara yang digunakan untuk menduga tingkat kesalahan prediksi dari suatu model pohon regresi. Salah satunya adalah dengan penduga validasi silang lipat V.

Keseluruhan data akan digunakan untuk membentuk deretan pohon $\{T_k\}$ dan deretan parameter kompleksitas $\{\alpha_k\}$. Untuk memperoleh nilai penduga validasi silang lipat V, amatan induk £ yang berukuran N dibagi secara acak menjadi V kelompok, yaitu $£_1, £_2, \dots, £_V$, yang berukuran sama yaitu n. Sampel *learning* ke-v adalah $£^{(v)} = £ - £_v$, v=1,2,...,V yang digunakan untuk memperoleh nilai penduga $d_k^{(v)}(x)$. Penduga validasi siliang lipat V dirumuskan dengan

$$R^{CV}(T_k) = \frac{1}{N}\sum_{v=1}^{V}\sum_{(x_n,x_y)\in £_v}\left(y_n - d_k^{(v)}(X_n)\right)^2 \qquad (2.18)$$

Dengan y_n = nilai amatan variabel respon dari data *learning* ke-V

$d_k^{(v)}(X_n)$= nilai penduga variabel respon dari data *learning* ke-V sebagai ukuran tingkat kesalahan prediksi.

Pohon regresi optimal adalah T_{k0} yang memenuhi kriteria

$$R^{CV}(T_{k0}) = \min_k R^{CV}(T_k) \qquad (2.19)$$

2.5 *Linear Mixed Effect* (LME)

Model umum *Linear Mixed Effect* adalah sebagai berikut:

$$y_i = X_i\beta + Z_i b_i + \epsilon_i$$
$$b_i \sim N(0.D), \epsilon_i \sim N(0, R_i)$$
$$`i = 1,2,\dots,n \qquad (2.20)$$

Dengan :

y_i= variabel respon berukuran $n_i \times 1$ untuk pengamatan ke i

X_i= matriks variabel efek tetap berukuran $n_i \times p$ untuk pengamatan ke i

Z_i= matriks variabel efek random berukuran $n_i \times q$ untuk pengamatan ke i

β= koefisien tetep berukuran $p \times 1$

b_i= koefisien efek random berukuran $q \times 1$ untuk grup ke i

ϵ_i= tingkat kesalahan/ error berukuran $n_i \times 1$ untuk pengamatan grup ke i

D= matriks kovariansi berukuran $q \times q$ untuk pengamatan grup ke i

R_i= matriks kovariansi berukuran $n_i \times n_i$ untuk error grup ke i

Siklus utama untuk *Maksimum Likelihood* berbasis algoritma EM pada model *Linear Mixed Effect*, adalah sebagai berikut:

Step 0. Himpunan $r = 0$, misalkan $\hat{\sigma}^2_{(0)} = 1$, dan $\widehat{D}_{(0)} = I_q$.

Step 1. Himpunan $r = r + 1$, memperbarui $\hat{\beta}_{(r)}$ dan $\hat{b}_{i(r)}$

$$\hat{\beta}_{(r)} = \left(\sum_{i=1}^{n} X_i^T \hat{V}^{-1}_{i(r-1)} X_i\right)^{-1} \left(\sum_{i=1}^{n} X_i^T \hat{V}^{-1}_{i(r-1)} y_i\right) \quad (2.21)$$

$$\hat{b}_{i(r)} = \widehat{D}_{(r-1)} Z_i^T \hat{V}^{-1}_{i(r-1)} (y_i - X_i)\hat{\beta}_{(r)}, i = 1,2,3, \dots, n, \quad (2.22)$$

Dimana

$$\hat{V}^{-1}_{i(r-1)} = Z_i \widehat{D}_{(r-1)} Z_i^T + \hat{\sigma}^2_{(r-1)} I_{n_i}, \quad i = 1,2,3, \dots, n. \quad (2.23)$$

Step 2. Memperbarui $\hat{\sigma}^2_{(r)}$, dan $\widehat{D}_{(r)}$ menggunakan

$$\hat{\sigma}^2_{(r)} = N^{-1} \sum_{i-1}^{n} \left\{\hat{\epsilon}^T_{(r)} \hat{\epsilon}_{i(r)} + \hat{\sigma}^2_{(r-1)} [n_i - \hat{\sigma}^2_{(r-1)} trace\left(\hat{V}_{i(r-1)}\right)]\right\} \quad (2.24)$$

$$\widehat{D}_{(r)} = n^{-1} \sum_{i-1}^{n} \left\{\hat{b}^T_{i(r)} \hat{b}_{i(r)} + [\widehat{D}_{(r-1)} - \widehat{D}_{(r-1)} Z_i^T \hat{V}^{-1}_{i(r-1)} Z_i^T \widehat{D}_{(r-1)}]\right\} \quad (2.25)$$

Step 3. Ulangi Step 1 dan Step 2 hingga konvergen.

2.6 *Mixed Effects Regression Trees* (MERT)

Model umum *Mixed Effect Regression Trees* adalah sebagai berikut:

$$y_i = f(X_i) + Z_i b_i + \epsilon_i$$
$$b_i \sim N(0.D), \epsilon_i \sim N(0, R_i)$$
$$i = 1,2, \dots, n \quad (2.26)$$

Model diatas diperoleh dengan mengganti bagian efek tetap pada model linear yaitu $X_i\beta$ dengan fungsi $f(X_i)$. Fungsi $f(X_i)$ dapat diestimasi dengan *standart regression tree model*. Bagian random $Z_i b_i$ masih diasumsikan linear.

Algoritma dari *Mixed Effect Regression Trees* adalah *Maksimum Likelihood* berbasis algoritma EM dengan mengganti struktur linear dengan menggunakan estimasi dari bagian tetap dari model struktur pohon, algoritmanya adalah sebagai berikut:

Step 0. Himpunan $r = 0$, misalkan $\hat{b}_{i(0)} = 0$ $\hat{\sigma}^2_{(0)} = 1$, dan $\widehat{D}_{(0)} = I_q$.

Step 1. Himpunan $r = r + 1$, memperbarui $y^*_{i(r)}, \hat{f}(X_i)_{(r)}$dan $\hat{b}_{i(r)}$

1. $y^*_{i(r)} = y_i - Z_i \hat{b}_{i(r-1),} i - 1,2,3, \dots, n,$ (2.27)

2. Misalkan $\hat{f}(X_i)_{(r)}$diestimasikan dari $f(X_i)$ diperoleh dari algoritma *standart tree* dengan $y^*_{i(r)}$ sebagai repson dan $X_i, i = 1,2,3, \dots, n$ sebagai kovariat. Perhatikan bahwa pohon dibangun menggunakan semua pengamatan N individu sebagai input bersama dengan vektor kovariat mereka.

3. $\hat{b}_{i(r)} = \widehat{D}_{(r-1)} Z_i^T \hat{V}^{-1}_{i(r-1)} \left(y_i - \hat{f}(X_i)_{(r)}\right), i = 1,2,3, \dots, n,$ (2.28)

Dimana $\hat{V}^{-1}_{i(r-1)} = Z_i \widehat{D}_{(r-1)} Z_i^T + \hat{\sigma}^2_{(r-1)} I_{n_i}, \quad i = 1,2,3, \dots, n.$ (2.29)

Step 2. Memperbarui $\hat{\sigma}^2_{(r)}$, dan $\widehat{D}_{(r)}$ menggunakan

$$\hat{\sigma}^2_{(r)} = N^{-1} \sum_{i-1}^{n} \left\{ \hat{\epsilon}^T_{(r)} \hat{\epsilon}_{i(r)} + \hat{\sigma}^2_{(r-1)} [n_i - \hat{\sigma}^2_{(r-1)} trace\left(\hat{V}_{i(r-1)}\right)] \right\} \quad ` \qquad (2.30)$$

$$\widehat{D}_{(r)} = n^{-1} \sum_{i-1}^{n} \left\{ \hat{b}^T_{i(r)} \hat{b}_{i(r)} + [\widehat{D}_{(r-1)} - \widehat{D}_{(r-1)} Z_i^T \hat{V}^{-1}_{i(r-1)} Z_i^T \widehat{D}_{(r-1)}] \right\} \qquad (2.31)$$

Step 3. Ulangi Step 1 dan Step 2 hingga konvergen.

2.7 Software R

R merupakan suatu sistem analisis statistika yang relatif lengkap, sebagai hasil dari kolaborasi riset berbagai statistikawan di seluruh belahan dunia. R saat ini dapat dikatakan merupakan *lingua franca* (bahasa standar) untuk keperluan komputasi statistika modern.

Versi paling awal R dibuat tahun 1992 di Universitas Auckland, New Zealand oleh Ross Ihaka dan Robert Gentleman (yang menjadi asal muasal akronim nama R untuk *software* ini). Saat ini *source code kernel* R dikembangkan oleh R Development Core Team, sedangkan pengembangan dan kontribusi berupa kode/*library*, melaporkan *error* dan *bugs* dan membuat dokumentasi untuk R, dilakukan oleh masyarakat statistikawan di seluruh penjuru dunia.

R bersifat *multiplatforms*, dengan file instalasi binary/file tar tersedia untuk sistem operasi Windows, Mac OS, Mac OS X, Free BSD, NetBSD, Linux, Irix, Solaris, AIX dan HPUX. Sintaks dari bahasa R secara umum ekuivalen dengan *software* statistika komersil S+. Untuk sebagian besar keperluan analisis statistika, pemrograman dengan R hampir identik dengan pemrograman dengan S+ (Rosad, 2017).

Tahapan dalam penggunaan *software* R adalah Sebagai Berikut:

1. Pastikan sudah install *software* R
2. Jalankan R dengan klik ikon R,
3. Prompt pada R adalah >
4. Untuk keluar dari R ketikkan

   >q ()

   Atau dengan memilih Exit pada menu File.

2.7.1 Fungsi dalam *Software* R

Fungsi merupakan perintah dalam R, fungsi ditulis dengan diakhiri oleh tanda ( ). Didalam tanda kurung tersebut kadangkala diisi dengan satu atau lebih argumen. Beberapa fungsi ada yang tidak membutuhkan argumen, hal ini dikarenakan semua argumennya mempunyai nilai default (dapat diubah) atau karena tidak ada argumen yang didefinisakan.

Ada banyak fungsi yang tersedia pada R, pengguna dapat membuat fungsi baru sesuai dengan kebutuhan. Tirta, I Made (2015, 74) menjelaskan beberapa fungsi yang akan digunakan dalam program R yaitu

1. source(“nama file”) digunakan untuk membaca file tanpa membuka file
2. sink(“nama file”) digunakan untuk mengarahkan penulisan hasil ke file sink
3. #komentar digunakan untuk menulis komentar yang tidak dieksekusi R

2.7.2 *Package* dalam *Software* R

Package merupakan kumpulan perintah-perintah yang digunakan untuk analisis tertentu. Default lingkungan R telah memiliki banyak fungsi-fungsi yang dapat digunakan untuk berbagai keperluan. Lingkungan R dapat ditambahkan fungsi-fungsi baru. Fungsi-fungsi terbaru tersebut biasanya dikemas dalam bentuk *package* (Faisal, 2017).

Package berisi fungsi dasar pada R yang berfungsi sebagai bahasa: aritmatika, input/output, dukungan pemrograman dasar, dll. Untuk fungsi yang lengkap, gunakan library(help="base") (The R development Core Team, 2008).

Terdapat ribuan *package* yang dapat diunduh secara gratis dan dapat disesuaikan dengan analisis yang akan digunakan. Misalkan untuk analisis Pohon Regresi tersedia *package* "rpart", dan *package* "rpart.plot" untuk penampilan grafik pohon regresi. Cara termudah dalam analisa dengan metode pohon adalah dengan menggunakan "rpart.plot" (Milborrow, Stephen, 2018).

2.7.3 Jenis-jenis data objek pada *software* R

Jenis-jenis data objek pada R yaitu data vektor, data matriks, data frame, dan data list (Suhartono, 2008).

2.7.3.1 Data Vektor

Vektor merupakan suatu array atau himpunan bilangan, karakter atau string, *logical value*, dan merupakan objek paling dasar yang dikenal dalam R. Pada data vektor harus digunakan mode tunggal pada data, sehingga gabungan dua data atau lebih yang berbeda mode tidak dapat dilakukan kedalam satu objek vektor. Contoh penggunaan data vektor pada R adalah:

c(1:10)

Outputnya adalah

[1] 1 2 3 4 5 6 7 8 9 10

2.7.3.2 Data Matriks

Matriks atau data array dua dimensi banyak digunakan pada program statistik, seperti untuk penyelesaian suatu persamaan linear. Proses entry data matriks dapat dilakukan dengan menggunakan fungsi matrix. Argumen yang diperlukan adalah elemen-elemen dari matriks, dan argumen optional yaitu banyaknya baris nrow dan banyaknya kolom ncolom, contoh penggunaan data matriks pada R yaitu:

Matriks.a = matrix(c(1,2,3,4,5,6),nrow=2,ncol=3)

Outpunya adalah:

Matriks.a

	[,1]	[,2]	[,3]
[1,]	1	3	5
[2,]	2	4	6

2.7.3.3 Data Frame

Data frame merupakan objek yang mempunyai bentuk sama dengan matriks, yaitu terdiri dari atas baris dan kolom. Perbedaannya adalah data frame dapat terdiri atas mode data yang berbeda-beda untuk setiap kolomnya, misalnya kolom pertama adalah numeric, kolom kedua adalah string, dan kolom ketiga adalah logical. Objek data frame dapat dibuat dengan menggunakan perintah data.frame. contoh penggunaan data frame pada R yaitu:

data.frame(c(1:4),c(T,T,F,F))

Outputnya adalah:

	$c.1.4$	$c.T..T..F..F.$
1	1	$True$
2	2	$True$
3	3	$False$
4	4	$False$

2.7.3.4 Data List

Data list merupakan objek yang paling umum dan paling fleksibel pada R. List adalah suatu vektor terurut dari sekumpulan komponen dapat sembarang data objek, yaitu vektor, matriks, data frame, atau data list. Tiap komponen pada data list dapat mempunyai mode yang berbeda. Contoh penggunaan data list pada R yaitu:

List(c(1:3),c(T,F,T,T),data.frame(nama=c("Ana","Ani","Ano"),nilai=c(8 :10)))

Outputnya adalah:

[[1]]

[1]1 2 3

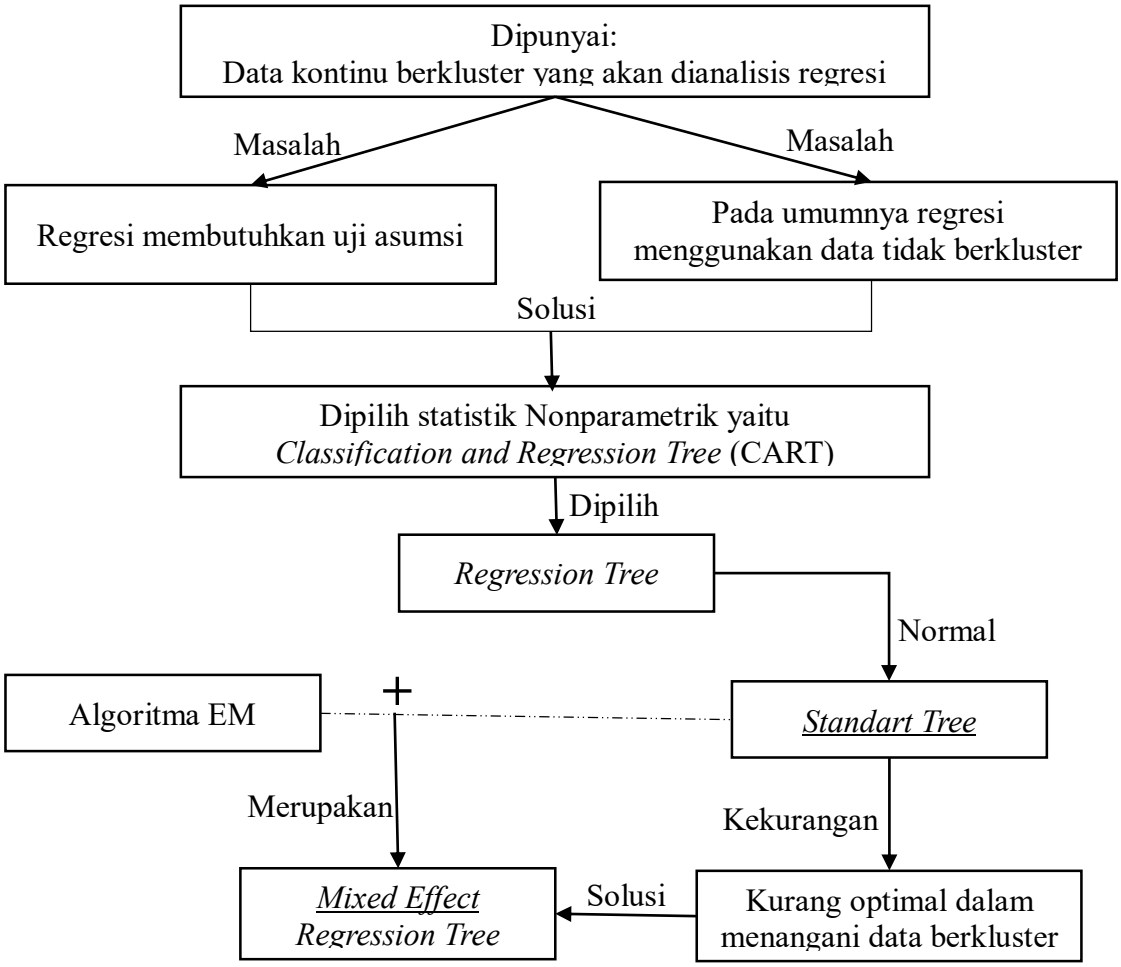
[[2]]

[1] TRUE FALSE TRUE TRUE

[[3]]

	nama	*nilai*
1	*Ana*	8
2	*Ani*	9
3	*Ano*	10

2.8 Kerangka Berpikir



Gambar 2.1 Kerangka Berpikir

BAB V
PENUTUP

5.1 SIMPULAN

Berdasarkan penelitian yang dilakukan dengan menggunakan metode *Standart Tree* dan metode *Mixed Effect Regression Tree* dapat disimpulkan bahwa:

1. Metode *Mixed Effect Regression Tree* lebih optimal dibanding metode *Standart Tree* dalam hal menangani data berkluster (studi kasus: Angka Kematian Ibu di 5 provinsi pada Pulau Jawa), ditunjukkan pada nilai PMSE metode *Mixed Effect Regression Tree* kurang dari nilai PMSE metode *Standart Tree*, yaitu 60,65 untuk metode *Mixed Effect Regression Tree* dan 150,96 untuk metode *Standart Tree*.
2. Analisis dengan metode *Mixed Effect Regression Tree* memberikan hasil bahwa Persentase kunjungan ibu hamil dengan K1, Persentase kunjungan ibu hamil dengan K4, Persentase ibu hamil mendapat tablet Fe3, Persentase komplikasi kebidanan yang ditangani, Persentase persalinan ditolong oleh tenaga kesehatan memberikan pengaruh terhadap Angka Kematian Ibu di 5 Provinsi di Pulau Jawa. Sedangkan analisis dengan metode *Standart Tree* memberikan hasil bahwa Persentase kunjungan ibu hamil dengan K1, Persentase kunjungan ibu hamil dengan K4, Persentase komplikasi kebidanan yang ditangani memberikan pengaruh terhadap Angka Kematian Ibu di 5 Provinsi di Pulau Jawa.

5.2 SARAN

1. Para Akademisi, metode *Mixed Effect Regresion Tree* dapat digunakan sebagai referensi oleh mahasiswa dari berbagai bidang baik matematika, statistik, maupun bidang lain dalam melakukan analisis data berkluster dikarenakan keoptimalan metode yang relatif tinggi, tanpa harus memenuhi syarat normalitas dan uji asumsi klasik.

2. Bagi Dinas Kesehatan, dengan adanya analisis Pengaruh Angka Kematian Ibu dengan metode *Mixed Effect Regresion Tree* diharapkan pemerintah dapat menerapkan kebijakan yang tepat dalam hal menurunkan Angka Kematian Ibu yaitu dengan melakukan sosialisasi terhadap pentingnya kunjungan K1 hingga K4, Pemberian Fe_3 kepada ibu hamil secara berkala sebagai program penanggulangan anemia, memberikan penanganan komplikasi kebidanan yang kompeten, dan meningkatkan tenaga kesehatan dalam memberikan pertolongan saat persalinan.

DAFTAR PUSTAKA

Afshartous, David, Leeuw, Jan De. 2005. *Decomposition of Prediction Error in Multilevel Models*. Florida: Taylor & Francis, Inc. ISSN: 0361-0918.

A.K. Gupta, D.K. Nagar. 1999. *Matrix Variate Distributions*. Ohio: Chapman & Hall/CRC.

Anton, Howard. 1997. *Aljabar Linear Elementer*, Edisi Kelima, terjemahan. Jakarta: Erlangga.

Afsah. 2016. *Analisis Jalur Faktor-Faktor Yang Mempengaruhi Angka Kematian Ibu Di Jawa Timur. Jurnal Matematika dan Pendidikan Matematika* Vol. 1 No.2.

Arkandi, Indi. dan Winahju, Wiwiek Setya. 2015. *Analisis Faktor Risiko Kematian Ibu dan Kematian Bayi dengan Pendekatan Regresi Poisson Bivariat di Provinsi Jawa Timur Tahun 2013*. JURNAL SAINS DAN SENI ITS, 4:2337-3520.

Black, R.E, Allen, L.H., Bhutta, Z.A., Caufield, L.E., De Onis, M., Ezzati, M., Rivera, J. 2008. *Maternal and child under nutrition: Global and regional exposures and health consequences. The Lancet Series on Maternal and Child Undernutrition 1*. Lancet 2008.

Breiman L, Friedman J, Stone CJ, Olshen RA (1984). *Classification and Regression Trees.* Wadsworth.

Depkes RI. *Profil Kesehatan Indonesia 2011*. Pusat data dan informasi. Jakarta: Departemen Kesehatan RI. 2012.

Dinkes Prov Banten. 2018. *Profil Kesehatan Provinsi Banten Tahun 2017*. Banten.

Dinkes Prov DIY. 2018. *Profil Kesehatan Provinsi DI Yogyakarta Tahun 2017*. Yogyakarta.

Dinkes Prov Jabar. 2018. *Profil Kesehatan Provinsi Jawa Barat Tahun 2017*. Bandung.

Dinkes Prov Jateng. 2018. *Profil Kesehatan Provinsi Jawa Tengah Tahun 2017*. Semarang.

Dinkes Prov Jatim. 2018. *Profil Kesehatan Provinsi Jawa Timur Tahun 2017*. Surabaya.

Djarwanto. 1991. *Statistik Non Parametrik*. Edisi 2.Yogyakarta: BPFE.

Edyanti, Deal Baby, dkk. 2014. *Faktor pada Ibu yang Berhubungan dengan Kejadian Komplikasi Kebidanan*. Jurnal Biometrika dana Kependudukan, Vol.3, No. 1 Juli 2014:1-7.

Faisal, Reza M. 2017. *Seri Belajar Data Science Pemrograman R untuk Data Scientist*.

Gentle, J.E. 2002. *Elements of Computational Statistics*. Springer.

H. Dewi, Y. Wilandari, and S. Sudarno,"*Analisis Faktor-Faktor yang Mempengaruhi Indek Mutu Benang Menggunakan Metode Pohon Regresi (Studi Kasus di PT. Industri Sandang Nusantara Unit Patal Grati)*," Media Statistika, Vol. 5, no.2, pp.75-86, Dec.2012.

Hajjem, Ahlem. 2010. *Mixed Effects Regression Trees For Clustered Data*, Mixed Effects Trees and Forests for Klustered Data. 93:14-43.

Helmizar. 2014. *Evaluasi Kebijakan Jaminan Persalinan (Jampersal) Dalam Penurunan Angka Kematian Ibu Dan Bayi Di Indonesia*. KEMAS, 9:197-205.

Hoog, R. V., McKean J. W., &Craig,A. T. 2005. *Introduction to Mathematical Statistic (6ed.)*. United States of America : Pearson Education.

Kementerian Kesehatan Republik Indonesia. 2018. *Profil Kesehatan Indonesia Tahun 2017*. Jakarta.

Kerlinger. 2006. *Asas-Asas Penelitian Behavioral*. Yogyakarta: Gadjah Mada University Press.

Luqman, Shaumal, dkk. 2015. *Identifikasi Variabel yang Mempengaruhi Besar Pinjaman dengan Metode Pohon Regresi*. JURNAL GAUSSIAN, Volume 4, Nomor 4, Tahun 2015, Halaman 1027-1035.

Mardika, Z. W., Mukid, M. A. & Yasin, H. 2016. *Pembentukan Pohon Klasifikasi Biner Dengan Algoritma Cart (Classification And Regression Trees)*. JURNAL GAUSSIAN, Volume 5, pp. 583-592.

Milborrow, Stephen. 2018. *Plotting rpart trees with the rpart.plot package.*

Mubarak. 2012. *Ilmu Kesehatan Masyarakat*. Jakarta: Salemba Medika.

N Dwidayati, SH Kartiko, Subanar. 2013. *Konvergensi Estimator Dalam Model Mixture Berbasis Missing Data. Semarang*. Jurnal MIPA 36 (2):185-192.

N Dwidayati, Zaenuri. 2017. *Convergence Properties of the EM Algorithm in the Mixture Model with Missing Data*. International Conference on Mathematics, Science and Education 20017 (ICMSE2017).

Norma, Eka, dkk. 2012. *Cakupan Kunjungan Pertama Ibu Hamil pada Pelayanan Antenatal Care*. Jurnal Ilmiah Mahasiswa, Vol.2 No.1.

Rangkuti, Saddiyah. 2015. *Upaya Menekan Angka Kematian Ibu Melahirkan*. Jurnal Ilmiah Research Vol. 1 No. 3.

Rosadi, Dedi. 2017. *Analisis Statistika dengan R*. Yogyakarta. Badan Penerbit Dan Publikasi Universitas Gadjah Mada.

Rustono, dkk. 2018. *Panduan Penulisan Karya Ilmiah. Semarang*. UNNES PRESS.

Septiari, Laras. 2017. *Penyelesaian Model Pohon Regresi Efek Campuran untuk Memodelkan Data Berkluster*. Skripsi:Universitas Gadjah Mada.

Sirait, Law Rencus, Shaufiah, Kurniati, Angelina Kurniati. 2013. *Analisis Algoritma Expectation Maximization (EM) Dalam Penanganan Missing Value*. Tugas Akhir. Universitas TELKOM.

Somantri, Ating, Sambas Ali Muhidin. 2006. *Aplikasi STATISTIKA Dalam Penelitian*. Bandung. CV Pustaka Setia.

Sudjana. 2002. *Metode Statistika*. Bandung. Penerbit Tarsito.

Sugianto, Aris. 2016. *Jenis-Jenis Data Variabel (Variabel Diskrit dan Variabel Kontinu)*. PalangKaraya.

Sugiyono. 2006. *Metode Penelitian Kuantitatif Kualitatif dan R&D*. Bandung:Alfabeta.

Sugiyono. 2013. *Metode Penelitian Pendidikan Pendekatan Kuantitatif, Kualitatif, dan R&D*. Bandung: Alfabeta.

Suhartono. 2008. *Analisis Data Statistik dengan R*. Surabaya. Lab. Statistik Komputasi.

Spiegel, R.M., Schiller, J., Srinivasan, R.A. 2004. *Schaum's outlines of Probabilitas dan statistik.* Edisi Kedua. Indriasari R, penerjemah: Simarmata L, editor. Penrbit Erlangga. Terjemahan dari: Schaum's outlines of theory and problems of Probability and Statistics, second edition.

The R Development Core Team. 2008. R: *A Language and Environment for Statistical Computing Reference Index*. R foundation for Statistical Computing. ISBN 3-900051-07-0.

Tirta, I Made. 2015. *Buku Panduan Program Statistika*. UNEJ.

Wieta B. 2007. *Metode Pohon Regresi untuk Eksplorasi Data dengan Peubah yang Banyak dan Kompleks*. Informatika Pertanian Volem 16 No.1.