ANALISIS PERBANDINGAN KLASIFIKASI METODE REGRESI LOGISTIK BINER DAN *RANDOM FOREST PADA BIG DATA*

TUGAS AKHIR

Disusun oleh:

Andi Nurhanna Manthovani
14 611 182

PROGRAM STUDI STATISTIKA

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM

UNIVERSITAS ISLAM INDONESIA

YOGYAKARTA

2018

ANALISIS PERBANDINGAN KLASIFIKASI METODE REGRESI LOGISTIK BINER DAN *RANDOM FOREST PADA BIG DATA*

TUGAS AKHIR

Diajukan Sebagai Salah Satu Ssyarat Untuk Memperoleh Gelar Sarjana
Jurusan Statistika

Disusun oleh:

Andi Nurhanna Manthovani
14 611 182

PROGRAM STUDI STATISTIKA

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM

UNIVERSITAS ISLAM INDONESIA

YOGYAKARTA

2018

HALAMAN PERSETUJUAN PEMBIMBING

TUGAS AKHIR

Judul : Analisis Perbandingan Klasifikasi Metode Regresi Logistik Biner dan *Random Forest* pada *Big Data*

Nama Mahasiswa : Andi Nurhanna Manthovani

Nomor Mahasiswa : 14 611 182

TUGAS AKHIR INI TELAH DIPERIKSA DAN DISETUJUI UNTUK DIUJIKAN

Yogyakarta, 14 Maret 2018

Pembimbing

Dr. Edy Widodo, S.Si., M.Si.

HALAMAN PENGESAHAN

TUGAS AKHIR

ANALISIS PERBANDINGAN KLASIFIKASI METODE REGRESI LOGISTIK BINER DAN *RANDOM FOREST* PADA *BIG DATA*

Nama Mahasiswa : Andi Nurhanna Manthovani

Nomor Mahasiswa : 14 611 182

TUGAS AKHIR INI TELAH DIUJIKAN

PADA TANGGAL 9 APRIL 2018

Nama Penguji	Tanda Tangan
1. Dr. Kartiko, M.Si.	.....
2. Ayundyah Kesumawati, M.Si.	.....
3. Dr. Edy Widodo, S.Si., M.Si.	.....

Mengetahui,

Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam

Drs. Allwar, M.Sc. Ph.D.

KATA PENGANTAR

Assalamu'alaikum Warahmatullahi Wabarakatuh

Alhamdulillah penulis ucapkan kepada Allah SWT, karena berkat limpahan Rahmat, Taufik, Hidayah, serta Inayah-Nya penulis dapat menyelesaikan Tugas Akhir yang berjudul **"Analisis Perbandingan Klasifikasi Metode Regresi Logistik Biner dan *Random Forest pada Big Data*"**. Puji syukur ke hadirat Allah SWT atas rahmat, kesehatan, karunia dan petunjuk yang telah diberikan. Shalawat serta salam penulis haturkan kepada Nabi Muhammad SAW beserta keluarga, sahabat, dan umatnya yang telah membawa kita dari kegelapan menuju cahaya Islam.

Penulis menyadari bahwa penulisan tugas akhir ini banyak memperoleh bantuan dari berbagai pihak, baik yang berupa saran, kritik, bimbingan maupun bantuan lainnya. Oleh karena itu, pada kesempatan ini penulis menyampaikan ucapan terima kasih kepada:

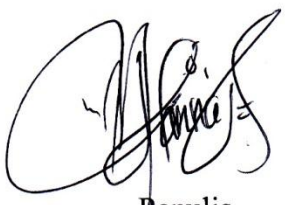
1. Bapak Drs. Allwar, M.Sc., Ph.D., selaku Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam yang telah memberikan izin penelitian.
2. Bapak Dr. RB Fajriya Hakim, S.Si., M.Si., selaku Ketua Program Studi Statistika Fakultas Matematika dan Ilmu Pengetahuan Alam yang telah banyak membantu dan memberi ilmu dan wawasan baru kepada penulis.
3. Bapak Dr. Edy Widodo, S.Si., M.Si., selaku Dosen Pembimbing Tugas Akhir serta Dosen Pembimbing Akademik, atas arahan dan bimbingan beliau selama ini.
4. Seluruh Dosen dan Staff Program Studi Statistika yang telah banyak memberikan bimbingan kepada penulis.
5. Mama, Papa, Adik-adik tersayang dan Keluarga Besar atas dukungan yang tidak ada hentinya, serta semangat dan do'a untuk penulis.

6. Sahabat seperjuangan: Septi, Yusi, Juju, Nanda, Ella, Tiwi, Tista, Dhila, Feby, Mei, Fizhan, Husni, Sem, Febrian, Sendhy, Aufa dan Alan. Terimakasih atas kebersamaan dan kekeluargaan yang selalu dijaga sejak awal datang ke Yogyakarta hingga saat ini.
7. Saktiwan Dwiatmono selaku *partner* yang selalu siap memberikan dukungan kepada penulis.
8. Teman-teman KKN unit 94; Putri, Arga, Yuris, Andi, Aulia, Yudha, Oriza, dan Audi atas motivasi dan senyum yang menenangkan.
9. Sahabat Statistika UII khususnya angkatan 2014 (XIX), yang banyak membantu penulis dalam usaha penyelesaian tugas akhir ini.
10. Semua pihak yang telah membantu dan tidak dapat penulis sebutkan satu persatu. Semoga Allah SWT selalu memberi rahmat dan anugerah-Nya kepada mereka semua tanpa henti. Aamiin.

Penulis menyadari bahwa dalam tugas akhir ini masih jauh dari kata sempurna, oleh karena itu segala kritik dan saran yang bersifat membangun selalu penulis harapkan. Semoga tugas akhir ini dapat bermanfaat bagi penulis khususnya dan bagi semua yang membutuhkan pada umumnya. Akhir kata, semoga Allah SWT selalu melimpahkan rahmat serta hidayah-Nya kepada kita semua, Aamiin aamiin ya robbal'alamin.

Wassalamu'alaikum Wr. Wb.

Yogyakarta, Maret 2018



Penulis

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

DAFTAR ISTILAH

APER	:*Apparent Error Rate* / Tingkat kesalahan klasifikasi dalam fungsi klasifikasi
Bebas	:Variabel yang memengaruhi / faktor yang diukur untuk menentukan hubungan antara fenomena yang diobservasi
Biner	:Sistem penulisan angka dengan menggunakan dua simbol, biasanya 0 dan 1.
CHAID	: *Chi-squared Automatic Interaction Detector*
Data Empiris	:Data yang dihasilkan dari percobaan atau pengamatan.
Dummy	:Variabel boneka / Variabel nominal yang digunakan untuk menunjukkan kelompok yang mendapat maupun yang tidak mendapatkan perlakuan
IPK	:Indeks Prestasi Kumulatif
Kualitatif	:Data informasi yang berbentuk kalimat verbal
Kuantitatif	:Data informasi yang berupa angka atau bilangan.
Linier	:Terletak pada suatu garis lurus
MDA	:*Mean Decrease Accuracy*
MDG	:*Mean Decrease Gini*
Multinomial	:Sistem penulisan angka dengan menggunakan lebih dari dua simbol
Parsial	:Sebagian dari suatu keseluruhan
Polikotomus	:Memungkinkan keadaan data kualitatif dengan kategori dua atau lebih, dan juga kuantitatif
RMSE	:*Root Mean Square Error*
Simultan	:Sesuatu yang terjadi bersamaan
Terikat	:Variabel yang dipengaruhi / faktor yang diobservasi dan diukur untuk menentukan adanya pengaruh variabel bebas.

DAFTAR LAMPIRAN

PERNYATAAN

Dengan ini penulis menyatakan bahwa dalam Tugas Akhir ini tidak terdapat karya yang sebelumnya pernah diajukan untuk memperoleh gelar kesarjanaan di suatu perguruan tinggi dan sepanjang pengetahuan penulis tidak terdapat karya atau pendapat yang pernah ditulis atau diterbitkan oleh orang lain, kecuali yang di acu dalam naskah ini dan disebutkan dalam daftar pustaka.

Yogyakarta, Maret 2018

Penulis

ANALISIS PERBANDINGAN KLASIFIKASI METODE REGRESI LOGISTIK BINER DAN *RANDOM FOREST PADA BIG DATA*

Oleh : Andi Nurhanna Manthovani

Program Studi Statistika Fakultas Matematika dan Ilmu Pengetahuan Alam

Universitas Islam Indonesia

INTISARI

Seiring dengan perkembangan pesat di bidang teknologi dan informasi, berbagai macam data dapat dihasilkan dengan mudah dan memiliki jumlah yang tak terbatas di abad dua puluh, yang dikenal dengan era *Big Data*. Cara pengumpulan data pada era *Big Data* melengkapi teknik sampling dengan menghasilkan informasi yang lebih cepat dan relatif lebih murah daripada survei. *Machine Learning* sebagai salah satu ilmu data sains saat ini mulai dikenal oleh para peneliti yang terbiasa berkutat dengan statistika. Meskipun aplikasi dari disiplin ilmu Statistika dan *Machine Learning* kelihatan sangat berbeda, dua ilmu tersebut sangat berkaitan. Menyadari bahwa *Big Data* banyak berkorelasi dengan data biner maka Regresi Logistik Biner untuk Statistika dan *Random Forest* untuk *Machine Learning* dapat digunakan sebagai sarana pengolahan datanya. Pada *Big Data* yang dibangkitkan melalui *software* R, dilakukan perbandingan kemampuan menggunakan tingkat akurasi. Pada uji coba pertama Regresi Logistik Biner memiliki tingkat akurasi sebesar 61.18% dan 96.94% untuk *Random Forest*. Untuk uji coba kedua dengan pembagian data *training* dan data *test* didapatkan tingkat akurasi Regresi Logistik Biner sebesar 63.11% dan *Random Forest* sebesar 78.24%. Kedua hasil menunjukkan bahwa *Random Forest* lebih unggul dalam memprediksi dengan selisih 35.76% dan 15.13%.

Kata Kunci : *Machine Learning*, *Big Data*, Regresi Logistik Biner, *Random Forest*, dan *software* R.

CONSIDERATION ANALYSIS OF BINARY LOGISTIC REGRESSION AND RANDOM FOREST AT BIG DATA

By : Andi Nurhanna Manthovani

Department of Statistics Faculty of Mathematics and Science

Islamic University of Indonesia

ABSTRACT

Along with rapid development of technology and information, there are many easy ways to get data in 20th century and commonly known as the era of Big Data. At the era of Big Data, collecting data have been completed sampling techniques because it is faster and cheaper than survey. Machine Learning is one of data science which commonly known by many statistical researcher. There are the difference application of Statistics and Machine Learning but both of them are related. Big Data have many correlation with binary data so Binary Logistic Regression for Statistics and Random Forest for Machine Learning can be used to process that type of data. Comparison between Binary Logistic Regression and Random Forest in this research created by R with accuracy value. At the first test, Binary Logistic Regression has an accuration about 61.18% and 96.94% for Random Forest. With define data into training and test, the result of second test is Binary Logistic Regression has an accuration about 63.11% and Random Forest has 78.24%. Both of them showing that Random Forest is greater to predict with difference accuration about 35.76% and 15.13% than Binary Logistic Regression.

Keywords : *Machine Learning, Big Data, Binary Logistic Regression, Random Forest, and R software.*

BAB I

PENDAHULUAN

1.1 Latar Belakang

Menurut Badan Pusat Statistik pada tahun 2010, Statistika merupakan ilmu yang sudah berkembang sejak awal abad masehi, dimana saat itu sejarah mencatat bahwa bangsa romawi pernah melakukan kegiatan semacam sensus untuk mendata seluruh warga negaranya.

Kemudian memasuki era teknologi dan informasi, Statistika berkembang menjadi lebih aplikatif. pada era ini analisis statistik rumit pun bisa dipakai lebih mudah, cepat, dan tepat dengan bantuan komputerisasi. Perkembangan pesat di abad dua puluh seakan membuat banyak ahli merasa statistika sudah menjadi ilmu mapan yang sulit digoyahkan lagi.

Seiring dengan perkembangan pesat di bidang teknologi dan informasi, berbagai macam data dapat dihasilkan dengan mudah dan memiliki jumlah yang tak terbatas di abad dua puluh, yang dikenal dengan era *Big Data*. (Anova, 2013)

Permana (2016) menyatakan bahwa *Big Data* telah digunakan dalam banyak bisnis. Selain itu, *Big Data* sudah banyak dimanfaatkan untuk mempelajari profil konsumen, pola konsumsi, manajemen resiko, dan sebagainya. Cara pengumpulan data pada era *Big Data* melengkapi teknik sampling dengan menghasilkan informasi yang lebih cepat dan relatif lebih murah. Selain itu tidak menghasilkan beban responden seperti survei yaitu cara pengumpulan data yang menghasilkan data-set yang amat lekat dengan analisis statistik.

Machine Learning sebagai salah satu ilmu data sains saat ini mulai dikenal dan menjadi topik pembicaraan banyak kalangan, terutama para peneliti yang terbiasa berkutat dengan statistika.

Meskipun aplikasi dari disiplin ilmu Statistika dan *Machine Learning* kelihatan sangat berbeda, dua ilmu tersebut sangat berkaitan. Baik Statistika maupun *Machine Learning* merupakan ilmu tentang data. Teori-teori di disiplin

ilmu Statistika dan *Machine Learning* sebagian besar juga saling tumpang tindih. Kedua disiplin ilmu sama-sama berdasarkan teori peluang dan membahas dasar-dasar teori dan model yang sama. Perbedaan kedua ilmu tersebut terletak pada fokus yang berbeda. Statistika lebih fokus ke arah pengambilan kesimpulan, sedangkan *Machine Learning* fokus ke prediksi data baru. Dari persamaan dan perbedaan tersebut, tidak salah kalau Statistika dan *Machine Learning* disebut sebagai dua wajah berbeda dari satu kesatuan disiplin ilmu. (Fathony, 2015).

Metode yang umum digunakan pada Statistika yaitu Analisis Regresi. Analisis Regresi mempelajari bentuk hubungan antara variabel bebas dengan variabel terikat. Ketika data yang dianalisis memiliki variabel terikat berupa data kategorik maka digunakanlah Regresi Logistik.

Regresi Logistik tidak mengasumsikan hubungan linier antar variabel bebas dan terikat dikarenakan bentuk variabel terikat yang kategorik. Dua nilai yang biasa digunakan sebagai variabel terikat yang diprediksi adalah 0 dan 1 yang menyatakan dua kondisi (biner) dengan kondisi yang bertolak belakang. Kondisi tersebut dapat ditemukan pada pengujian Statistika dengan menggunakan metode Regresi Logistik Biner.

Big Data banyak berkorelasi dengan data biner. Sebagai contoh menyatakan suatu kondisi; Ya atau Tidak, Berhasil atau Gagal, Ringan atau Berat, dan sebagainya. Terdapat salah satu algoritma *Machine Learning* yang mampu menangani kondisi serupa Regresi Logistik, yaitu *Random Forest.* Algoritma tersebut didasarkan pada teknik pohon keputusan sehingga mampu mengatasi masalah nonlinier dengan kondisi yang sama yaitu bekerja pada data dengan variabel terikat yang kategorik.

Berdasarkan latar belakang tersebut, peneliti menyadari bahwa Statistika dan *Machine Learning* memiliki beberapa kesamaan dan perbedaan dalam pengaplikasian disiplin ilmu maupun alat analisis, salah satunya yaitu pada metode Regresi Logistik biner untuk statistika dan *Random Forest* utuk *Machine Learning*. Kedua metode tersebut sangat menarik untuk digunakan dikarenakan kondisi variabel terikat yang kategorik dengan dua kategori atau biasa disebut

biner. Sehingga topik tersebut dirasa perlu untuk dikaji untuk menentukan metode mana yang lebih efisien dan sesuai dengan kebutuhan penelitian.

1.2 Rumusan Masalah

Berdasarkan uraian latar belakang, maka permasalahan yang dapat diidentifikasi dalam penelitian ini adalah sebagai berikut:

1. Bagaimana penerapan Regresi Logistik biner serta *Random Forest* pada Big Data yang dibangkitkan menggunakan R?
2. Bagaimanakah perbandingan kemampuan Regresi Logistik Biner dengan *Random Forest* setelah dianalisis?

1.3 Batasan Masalah

Pembatasan suatu masalah digunakan untuk menghindari adanya penyimpangan maupun pelebaran pokok masalah agar penelitian lebih terarah dan memudahkan dalam pembahasan sehingga tujuan penelitian akan tercapai. Beberapa batasan masalah dalam penelitian ini adalah sebagai berikut:

1. Luas lingkup data meliputi kategori-kategori kasus kecelakaan sepeda motor oleh Ditlantas Polda DIY dengan data simulasi yang dibandingkan menggunakan *software* R.
2. Hasil perbandingan berlaku untuk metode Regresi Logistik Biner dan *Random Forest* pada kasus yang disajikan pada penelitian ini.

1.4 Tujuan Penelitian

Tujuan yang hendak dicapai dari penelitian ini adalah sebagai berikut:

1. Menerapkan Regresi Logistik Biner serta *Random Forest* pada *Big Data* yang dibangkitkan menggunakan R.
2. Membandingkan kemampuan Regresi Logistik Biner dengan *Random Forest* setelah dianalisis.

1.5 Manfaat Penelitian

Adapun manfaat dari penelitian ini adalah sebagai berikut:

1. Bagi penulis akan bermanfaat untuk lebih memperdalam tentang Regresi Logistik Biner dan *Random Forest*.
2. Dapat mengetahui perbedaan dan persamaan kemampuan antara Regresi Logistik Biner dan *Random Forest*.
3. Mengetahui efisiensi dari Regresi Logistik Biner dan *Random Forest* dan menentukan penggunaannya sesuai kondisi yang diinginkan.
4. Dengan adanya penelitian ini akan membuka peluang diadakannya penelitian perbandingan dari metode-metode yang ada pada Statistika dan *Machine Learning* lainnya.

BAB II
KAJIAN PUSTAKA

Penelitian yang dilakukan kali ini melihat dari referensi jurnal serta beberapa skripsi yang berhubungan dengan penelitian yang akan dilakukan. Adapun tinjauan pustaka yang digunakan adalah sebagai berikut:

Pada tahun 2011, Dewi melakukan penelitian dengan mengangkat studi kasus *Random Forest* pada *Driver Analysis*. *Driver Analysis* merupakan istilah yang digunakan secara luas meliputi berbagai metode analisis dan dilakukan untuk memahami pengaruh variabel bebas terhadap variabel terikat sehingga dapat diketahui prioritas setiap variabel bebas dalam menggerakkan variabel terikat (Wiener dan J., 2005). Pada kasus tersebut, peneliti menggunakan *Random Forest* sebagai alat analisis yang mampu mengatasi masalah nonlinier. Selain itu *Random Forest* juga menghasilkan ukuran tingkat kepentingan variabel bebas sehingga peneliti bertujuan untuk mengetahui ukuran *Random Forest* yang berakurasi prediksi tinggi dan stabil serta menghasilkan *driver analysis* yang stabil pula. Dari hasil penelitian tersebut diperoleh kesimpulan, *Random Forest* dengan ukuran lebih dari 500 memberikan akurasi prediksi yang tinggi dan stabil, yaiu dengan tingkat misklasifikasi berkisar antara 34.5% dan 35.5% dengan nilai rataannya sebesar 34.5%.

Kemudian persamaan dan perbedaan disiplin ilmu Statistika dan *Machine Learning* dikaji oleh Fathony (2015). Penelitian tersebut berisikan teori dengan membandingkan masing-masing pondasi dasar kedua disiplin ilmu, dilanjuttkan dengan cara pengambilan kesimpulan, interpretasi, juga prediksi. Kemudian dilihat pula area-area yang sama-sama didalami oleh kedua ilmu, maupun area-area yang kurang didalami peneliti dari satu ilmu ke ilmu lainnya. Selain itu peneliti juga membuka cara berfikir matematis dan algoritmis, kultur jurnal dan konferensi, serta bahasa pemrograman masing-masing dari statistika maupun *Machine Learning*. Dari hasil penelitian tersebut diperoleh kesimpulan bahwa disiplin ilmu Statistika dan *Machine Learning* mempunyai banyak persamaan dan juga perbedaan. Kedua disiplin ilmu sama-sama berdasarkan teori peluang dan membahas dasar-dasar teori dan model yang sama. perbedaan keduanya terletak pada fokus yang

berbeda. Statistika lebih fokus ke arah pengambilan kesimpulan, sedangkan *Machine Learning* fokus ke prediksi data baru. Dari persamaan dan perbedaan tersebut, peneliti menilai tidak salah jika Statistika dan *Machine Learning* disebut sebagai dua wajah berbeda dari satu kesatuan disiplin ilmu.

Statistika dan *Machine Learning* kembali dibandingkan pada tahun 2016 dalam penelitian yang dilakukan oleh Rumaendra. Penelitian ini lebih spesifik dikarenakan langsung membandingkan Regresi Logistik Biner untuk statistika dengan algoritma C4.5 untuk *Machine Learning* pada penyakit hipertensi UPT Puskesmas Ponjong I Gunungkidul. Kedua metode tersebut dipilih dikarenakan tepat dengan tujuan penelitian yaitu mengklasifikasian penyakit hipertensi. Dikarenakan data yang digunakan memiliki variabel respon bertipe kategorik maka peneliti menggunakan Regresi Logistik Biner dan membandingkan ketepatan klasifikasinya dengan algoritma C4.5 yang merupakan salah satu metode klasifikasi dari data mining yang digunakan untuk mengkonstruksikan pohon keputusan. Menggunakan kedua metode tersebut peneliti dapat mengetahui nilai ketepatan klasifikasi. Dari hasil penelitian tersebut diperoleh kesimpulan, klasifikasi penyakit hipertensi dengan metode Regresi Logistik Biner diperoleh nilai APER=27,4648% dan ketepatan klasifikasi sebesar 72,5352%, sedangkan menggunakan algoritma C4.5 diperoleh nilai APER=35,9155% dan ketepatan klasifikasi sebesar 64,0845%.

Mambang dan Byna (2017) melakukan penelitian perbandingan analisis algoritma C4.5, *Random Forest* dan CHAID *Decision Tree* untuk megklasifikasikan tingkat kecemasan ibu hamil. Peneliti Menggunakan ketiga metode tersebut dengan tujuan membandingkan algoritma-algoritma yang ada berdasarkan klasifikasi galat dan tingkat akurasinya. Berdasarkan penelitian yang dilakukan didapatkan hasil akurasi dengan menggunakan algoritma pohon keputusan C4.5, *Random Forest* dengan CHAID *Decision Tree* menghasilkan akurasi yang lebih baik yaitu berada pada angka 64% dan 62.67%. Pada pengujian *training* dan *testing* yang dilakukan dapat pula disimpulkan bahwa metode *Random Forest*, C4.5 dan CHAID *Decision Tree* dapat diterapkan. *Random Forest* menghasilkan hasil akurasi yang paling unggul dengan nilai 64% dan RMSE sebesar 0.584.

Penerapan *Big Data* pada salah satu metode yang digunakan pada penelitian ini yakni Regresi Logistik Biner pernah dikaji oleh Ilham (2017) pada studi kasus *Airline On-time Performance 2005*. Pada penelitian tersebut data *Airline On-time Performance 2005* diyakini penelitinya sebagai *Big Data* karena ukuran data melebihi kemampuan *software* yang umum digunakan. Berdasarkan penelitian yang dilakukan didapatkan model Regresi Logistik Biner yaitu Arr Delay = 0.0043 + 1.0018*ActualElapsedTime* – 1.0014 *CRSElapsedTime* + 0.0003*AirTime* + 0.9999*DepDelay* – 0.0001*Distance* + 0.0003*TaxiIn* – 0.0016*TaxiOut* dan diketahui semua variabel bebas yang digunakan berpengaruh signifikan terhadap model Regresi Logistik Biner yang dihasilkan.

Berdasarkan kelima kajian pustaka yang dicantumkan, didapatkan pengetahuan bahwa penggunaan analisis yang ada pada statistika turut pula diolah menggunakan *Machine Learning*. Memasuki era *Big Data*, penggunaan *Random Forest* yaitu salah satu alat pada *Machine Learning* sudah digunakan pada data serupa Regresi Logistik Biner atau alat pada Statistika. Sebelumnya Regresi Logistik Biner pernah dibandingkan dengan algoritma C4.5. Begitu pula algoritma C4.5 yang dibandingkan dengan *Random Forest*. Namun belum ada yang benar-benar membandingkan metode Regresi Logistik Biner dengan *Random Forest* secara langsung. Maka dari itu, penelitian ini hadir sebagai sarana perbandingan Regresi Logistik Biner dan *Random Forest* yang diharapkan dapat memberikan hasil baik, manfaat serta membuka jalan untuk perbandingan metode yang ada pada Statistika dan *Machine Learning* lain kedepannya.

Tabel 2.1. Kajian Pustaka

Tahun	Nama	Judul	Data/Variabel	Metode	Hasil Penelitian
2011	Nariswari Karina Dewi, Utami Dyah Syafitri, dan Soni Yadi Mulyadi. Institut Peranian Bogor	Penerapan Metode *Random Forest* dalam *Driver Analysis*	*Driver Analysis* pada data perusahaan riset pemasaran di Indonesia.	*Random Forest*	*Random Forest* memberikan akurasi prediksi yang tinggi dan stabil dengan rata-rata misklasifikasi sebesar 34.5%.
2015	Rizal Fathony. *University of Illinois Chicago*.	Statistika dan Machine Learning: Satu Ilmu Dua Wajah	Teori disiplin ilmu *Statistika* dan *Machine Learning*	Perbandingan pondasi dasar, cara pengambilan kesimpulan,	Disiplin ilmu Statistika dan *Machine Learning* mempunyai banyak persamaan dan

Tahun	Nama	Judul	Data/Variabel	Metode	Hasil Penelitian
				interpretasi dan prediksi	perbedaan. Sehingga disebut sebagai dua wajah berbeda dari satu kesatuan disiplin ilmu
2016	Wella rumaendra. Universitas Dipenogoro Semarang	Perbandingan Klasifikasi Penyakit Hipertensi Menggunakan Regresi Logistik Biner dan Algoritma C4.5.	Penyakit hipertensi UPT Puskesmas Ponjong I Gunungkidul	Regresi Logistik Biner dan algoritma C4.5	klasifikasi penyakit hipertensi dengan metode Regresi Logistik Biner diperoleh ketepatan klasifikasi sebesar 72,5352%, dan 64,0845% untuk algoritma C4.5.
2017	Mambang dan Agus Byna. Universitas Amikom Yogyakarta	Analisis Perbandingan Algoritma C4.5, *Random Forest* Dengan *CHAID Decision Tree* Untuk Klasifikasi Tingkat Kecemasan Ibu Hamil	Tingkat Kecemasan Ibu Hamil, Stikes Sari Mulia Banjarmasin	Algoritma C4.5, *Random Forest*, dan CHAID *Decision Tree*	*Random Forest* menghasilkan hasil akurasi yang paling unggul dengan nilai 64% dan RMSE sebesar 0.584.
2017	Fajar P Ilham, Mardiana Nur Wahidah, Qudhrotul Zahro' Khoiriya, dan Anindya Fauzianizahra. Universitas Gadjah Mada.	Aplikasi *Big Data* pada *Airline On-time Performance* 2005 dengan Regresi Logistik Biner	*Airline On-time Performance 2005 (Big Data)*	Regresi Logistik Biner	Didapatkan model Regresi Logistik Biner dan diketahui semua variabel bebas yang digunakan berpengaruh signifikan terhadap model Regresi Logistik Biner yang dihasilkan.

BAB III
LANDASAN TEORI

3.1 Regresi Logistik Biner

Analisis Regresi pada dasarnya adalah studi mengenai ketergantungan variabel terikat dengan satu atau lebih variabel bebas, dengan tujuan untuk mengestimasi dan memprediksi populasi atau nilai-nilai variabel terikat berdasarkan nilai variabel bebas yang diketahui (Ghozali, 2005).

Banyak kasus dalam analisis regresi memiliki kondisi variabel terikat yang bersifat kualitatif. Variabel terikat ini bisa mempunyai dua kelas atau kategori (biner) dan lebih dari dua kelas (multinomial). Salah satu pendekatan yang digunakan untuk mengestimasi model Regresi dengan variabel terikat bersifat kualitatif adalah dengan model probabilitas logistik atau disingkat logit (Widarjono, 2010).

Menurut Hosmer dan Lemeshow (2000), Regresi Logistik Biner merupakan suatu metode analisis data yang digunakan untuk mencari hubungan antara variabel terikat (y) yang memiliki kategori biner dengan variabel bebas (*x*) yang bersifat polikotomus. Keluaran dari variabel terikat terdiri dari 2 kategori yang biasanya dinotasikan dengan *y* =1 yang berarti sukses dan *y* = 0 yang artinya gagal.

Hosmer dan Lemeshow (2000) merumuskan suatu fungsi probabilitas, fungsi Regresi logistik, model Regresi Logistik, dan transformasi logit. Fungsi probabilitas untuk setiap observasi yaitu;

$$f(y) = \begin{cases} \pi^{y}(1-\pi)^{1-y} & untuk\ y = 0,1 \\ 0 & untuk\ y \neq 0,1 \end{cases} \qquad (1)$$

dengan;

π = probabilitas sukses.

Jika $y = 0$ maka $f(y) = (1 - \pi)$, dan jika $y = 1$ maka $f(y) = \pi$.

Sedangkan untuk fungsi Regresi Logistik dapat dituliskan sebagai berikut;

$$f(z) = \begin{cases} \frac{1}{1+e^{-z}} & untuk -\infty < z < +\infty \\ 0 & untuk\ lainnya \end{cases} \quad (2)$$

dengan $z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \cdots + \beta_k x_k$.

β_i (i = 0, 1, 2, .., k) merupakan koefisien dalam model regresi; dan x_j (j = 1, 2, .., k) adalah variabel bebas.

Nilai z antara -∞ dan +∞ sehingga nilai $f(z)$ terletak antara 0 dan 1. Hal ini menunjukan bahwa model logistik menggambarkan probabilitas atau resiko dari suatu objek. Secara umum, model Regresi Logistik ditulis dalam bentuk;

$$\pi(x) = \begin{cases} \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k)}} & untuk\ 0 \le \pi(x) \le 1 \\ 0 & untuk\ \pi(x) \le 0 \text{ dan } \pi(x) > 1 \end{cases} \quad (3)$$

dengan $\pi(x)$ adalah peluang kejadian sukses dengan nilai probabilitas $0 \le \pi(x) \le 1$ jika 1 dinyatakan sebagai kejadian sukses dan 0 berarti gagal.

$\pi(x)$ adalah fungsi yang nonlinier, sehingga perlu dilakukan transformasi ke dalam bentuk logit untuk memperoleh fungsi yang linier agar dapat dilihat hubungan antara variabel bebas dan variabel terikat. Pendugaan parameter model Regresi Logistik dapat diuraikan dengan menggunakan transformasi logit dari $\pi(x)$ yaitu;

$$\ln\left(\frac{\pi(x)}{1-\pi(x)}\right)$$

Karena

$$\frac{\pi(x)}{1-\pi(x)} = e^{(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k)} \quad (4)$$

maka

$$g(x) = \text{logit}\,[\pi(x)] = \ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k \quad (5)$$

g(x) merupakan fungsi hubungan dari model Regresi Logistik yang disebut sebagai fungi hubungan logit.

Sebagai contoh, pada penelitian Tampil (2016) dilakukan analisis Regresi Logistik dengan sasaran mahasiswa FMIPA Universitas Sam Ratulangi dengan variabel terikat (y) yaitu IPK yang dinotasikan dengan 0 untuk $y \leq$ rata-rata IPK dan 1 untuk $y >$ rata-rata. Serta variabel bebas yaitu Jenis kelamin (x_1) dinotasikan dengan 0 untuk perempuan dan 1 untuk laki-laki, Jurusan (x_2) dinotasikan dengan 0 untuk kimia dan 1 untuk matematika, Tempat tinggal (x_3) dinotasikan dengan 0 untuk bukan kost dan 1 untuk kost, Menerima Beasiswa (x_4) dinotasikan dengan 0 untuk tidak dan1 untuk ya, Daerah asal (x_5) dinotasikan dengan 0 untuk luar Sulawesi Utara dan 1 untuk Sulawesi Utara, Asal sekolah (x_6) dinotasikan dengan 0 untuk SMK dan 1 untuk SMA, Pekerjaan orang tua (x_7) dinotasikan dengan 0 untuk bukan pegawai negeri dan 1 untuk pegawai negeri, Biaya hidup tiap bulan (x_8) dinotasikan dengan 1 untuk $\leq$ 1.000.000 dan dinotasikan dengan 0 untuk > 1.000.000.

Pada penelitian tersebut dapat dilihat terdapat satu variabel terikat dengan dua kategori dan delapan variabel bebas yang semuanya merupakan data kategori biner. Model umum regresi dari kasus tersebut yaitu;

$$\pi(x) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_8 x_8)}}$$

Dari pengujian yang dilakukan, diperoleh model Regresi Logistik Biner dari IPK mahasiswa sebagai berikut;

$$\pi(x) = \frac{1}{1+e^{\begin{matrix}-(1.268+0.027x_1+1.294x_2-1.151x_3+0.318x_4 \\ -0.738x_5-1.001x_6+0.805x_7+0.03x_8)\end{matrix}}}$$

Lalu dilakukan transformasi logit dari $\pi(x)$ sehingga didapatkan fungsi logit;

$$g(x) = 1.268 + 0.027x_1 + 1.294x_2 - 1.151x_3 + 0.318x_4$$

$$- 0.738x_5 - 1.001x_6 + 0.805x_7 + 0.03x_8$$

3.1.1 Pengujian Parameter

Pengujian terhadap parameter-parameter estimasi model dilakukan untuk mengetahui peran seluruh variabel prediktor baik secara simultan maupun secara parsial.

Menurut Hosmer dan Lemeshow (2000), uji simultan disebut juga uji model chi-square, dilakukan sebagai upaya memeriksa peranan variabel terikat dalam model secara bersama-sama.

Hipotesis :

$H_0 : \delta_1 = \delta_2 = \ldots\ldots = \delta_k = 0$

H_1 : paling sedikit ada satu $\delta_i \neq 0$ ($i = 1,2,\ldots,k$) (6)

Statistik uji yang digunakan adalah statistik uji G^2 atau uji rasio likelihood.

$$G^2 = -2ln \frac{L_1}{L_0}$$

$$G^2 = -2ln \left(\frac{\left(\frac{n_1}{n}\right)^{n_1} \left(\frac{n_0}{n}\right)^{n_0}}{\prod_{i=1}^{n} \pi_1^{y_i} (1-\pi_1)^{(1-y_i)}} \right)$$

dengan:

n_1 = banyaknya observasi berkategori 1

n_0 = banyaknya observasi berkategori 0

n = banyaknya observasi (n_1+n_0)

L_1 = Likelihood tanpa variabel terikat tertentu

L_0 = Likelihood dengan variabel terikat tertentu

Statistik uji G^2 mengikuti distribusi chi-square, sehingga untuk memperoleh keputusan dilakukan perbandingan dengan χ^2 tabel. Dimana derajat bebas = k (banyaknya variabel terikat). Kriteria penolakan (tolak H_0) jika nilai $G^2 > \chi^2_{(db,\gamma)}$ atau *p-value* < γ.

Sedangkan pengujian parameter β secara parsial dilakukan dengan membandingkan model terbaik yang dihasilkan oleh uji simultan terhadap model tanpa variabel bebas di dalam model terbaik. Pengujian hipotesis yang dilakukan yaitu: Hasil pengujian secara parsial akan menunjukan apakah

suatu variabel terikat layak untuk masuk dalam model atau tidak (Alan Agresti, 2007).

Hipotesis :

$H_0 : \delta_i = 0 \quad (i = 1,2,\ldots,k)$

$H_1 : \delta_i \neq 0 \quad (i = 1,2,\ldots,k)$ (7)

Statistik Uji : Wald (W) = $\frac{\delta_i}{SE(\delta_i)}$

Rasio yang dihasilkan dari statistik uji, dibawah hipotesis H_0 akan mengikuti sebaran normal baku (Hosmer dan Lemeshow, 2000). Sehingga untuk memperoleh keputusan dilakukan perbandingan dengan distribusi normal baku (Z). kriteria penolakan (tolak H_0) jika nilai $W > Z_{\gamma/2}$ atau *p-value* $< \gamma$.

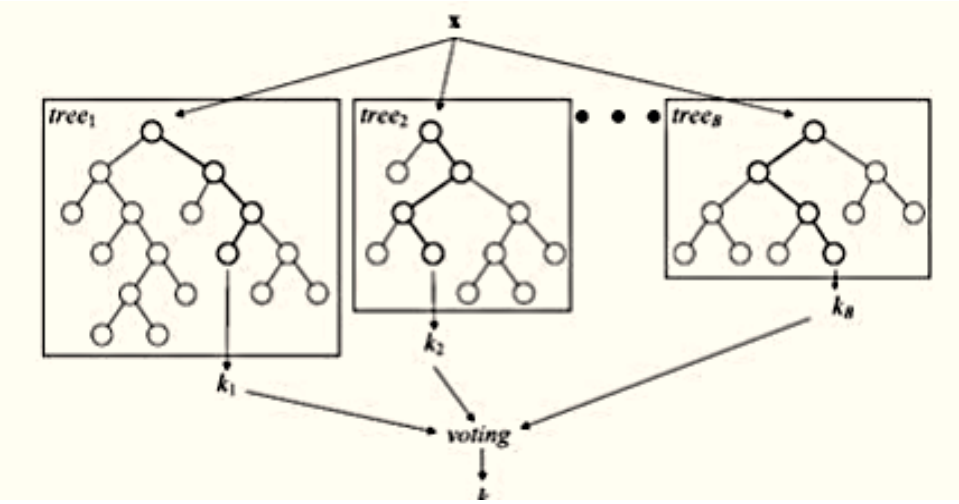
3.2 *Random Forest*

Skema *Random Forest* pertama kali dicetuskan oleh Breiman (2000) untuk membangun prediktor dengan sekumpulan pohon keputusan yang berkembang secara acak pada subruang data. Metode *random forest* merupakan model klasifikasi yang dilakukan dengan mengembangkan beberapa pohon keputusan berdasarkan seleksi data dan variabel yang dilakukan secara acak. Operator tersebut menghasilkan satu set sejumlah tertentu pohon acak yaitu menghasilkan *forest* (hutan;kumpulan pohon) acak. Model yang dihasilkan adalah model suara pilihan dari semua pohon.

Operator *Random Forest* menghasilkan satu set pohon acak. Kelas yang dihasilkan dari proses klasifikasi dipilih dari kelas yang paling banyak (modus) yang dihasilkan oleh pohon acak yang ada. (Biau, 2012)

Dengan membuat banyak pohon keputusan secara acak, maka sebenarnya banyak dari pohon-pohon yang dibuat oleh metode *Random Forest* menjadi kurang berguna. Namun *Random Forest* mampu menjadi sebuah metode klasifikasi yang cukup baik, karena beberapa pohon keputusan yang ikut dibuat saat konstruksi, ternyata memiliki kemampuan prediksi yang baik. Saat dilakukan pemilihan untuk menentukan klasifikasi secara keseluruhan, pohon-pohon yang buruk akan membuat prediksi yang acak dan saling bertentangan, sehingga

jawaban dari beberapa pohon keputusan yang merupakan prediktor yang baik akan muncul sebagai jawaban. (Nugroho dan Emiliyawati. 2017).



Sumber: Nugroho dan Emiliyawati (2017)

Gambar 3.1. Ilustrasi *Random Forest*

Dalam *Random Forest* terdapat ukuran kepentingan, yaitu MDA dan MDG. Dewi (2011) menyarankan untuk menggunakan banyak pohon ketika penelitian mempertimbangkan ukuran kepentingan dan saat dihadapkan pada variabel bebas yang banyak agar ukuran kepentingan yang dihasilkan semakin stabil.

3.2.1 Ukuran Tingkat Kepentingan

MDA (*Mean Decrease Accuracy*) merupakan salah satu ukuran tingkat kepentingan (*variable importance*) variabel bebas yang dihasilkan oleh metode *Random Forest*. MDA menampilkan seberapa besar tambahan observasi yang mengalami misklasifikasi jika satu persatu variabel bebas tidak diikutsertakan kedalam pengujian.

Ukuran kepentingan lainnya yaitu MDG (*Mean Decrease Gini*). Ukuran tersebut digunakan untuk melihat kestabilan tiap variabel bebas dalam *Random Forest*. Menurut Breiman (2000), semakin tinggi nilainya maka semakin baik.

3.3 Prosedur Klasifikasi

Menurut Johnson dan Wichern (2007) prosedur klasifikasi adalah suatu evaluasi untuk melihat peluang kesalahan klasifikasi (misklasifikasi) yang dilakukan oleh suatu fungsi klasifikasi. Prosedur klasifikasi yang baik ditentukan dengan nilai misklasifikasi yang kecil. Satu hal penting untuk menghasilkan prosedur klasifikasi ialah dengan menghitung tingkat *error* atau probabilitas misklasifikasi. Terdapat alat ukur yang dapat digunakan untuk menentukan kesalahan klasifikasi yang tidak bergantung pada distribusi populasi dan dapat mempermudah perhitungan berbagai prosedur klasifikasi.

Salah satu ukuran yang digunakan adalah *Apparent Error Rate* (APER) yang merupakan fraksi observasi dalam sampel yang salah diklasifikasikan pada fungsi klasifikasi. Penentuan kesalahan pengklasifikasian dapat diketahui melalui tabel klasifikasi. Kebalikannya yaitu *Total Accuracy Rate* atau Akurasi merupakan ukuran yang digunakan untuk mengetahui probabilitas atau persentase ketepatan klasifikasi. Untuk mendapatkan nilai ketepatan klasifikasi digunakan rumus: 1 – APER.

Tabel klasifikasi merupakan tabel kontingensi ($k \times k$) berdasarkan data empiris dari variabel terikat. Pembuata tabel klasifikasi dirujuk pada **Tabel 3.1**.

Tabel 3.1. Tabel Klasifikasi

Keanggotaan sebenarnya	Keanggotaan prediksi		Total
	$\hat{\pi}_1$	$\hat{\pi}_2$	
π_1	n_{11}	n_{12}	A
π_2	n_{21}	n_{22}	B
Total	C	D	E

Keterangan:

n_{11} : jumlah y_i dari π_1 yang tepat diklasifikasikan sebagai $\hat{\pi}_1$

n_{12} : jumlah y_i dari π_1 yang tidak tepat diklasifikasikan sebagai $\hat{\pi}_2$

n_{21} : jumlah y_i dari π_2 yang tidak tepat diklasifikasikan sebagai $\hat{\pi}_1$

n_{22} : jumlah y_i dari π_2 yang tepat diklasifikasikan sebagai $\hat{\pi}_2$

A : Jumlah keseluruhan y_i yang ada pada π_1

B : Jumlah keseluruhan y_i pada ada pada π_2

C : Jumlah keseluruhan y_i yang diklasifikasikan sebagai $\hat{\pi}_1$

D : Jumlah keseluruhan y_i yang diklasifikasikan sebagai $\hat{\pi}_2$

E : Jumlah keseluruhan observasi

Sehingga diperoleh rumus ketepatan klasifikasi secara keseluruhan nilai Tingkat Akurasi adalah:

$$\text{Akurasi} = \left(\frac{n_{11}+n_{22}}{\text{E}}\right) \qquad (8)$$

Kemudian, untuk mendapatkan nilai kesalahan klasifikasi digunakan rumus;

$$\text{APER} = \left(\frac{n_{12}+n_{21}}{\text{E}}\right) \text{ atau APER} = 1 - \text{Akurasi}$$

3.4 Distribusi Beta

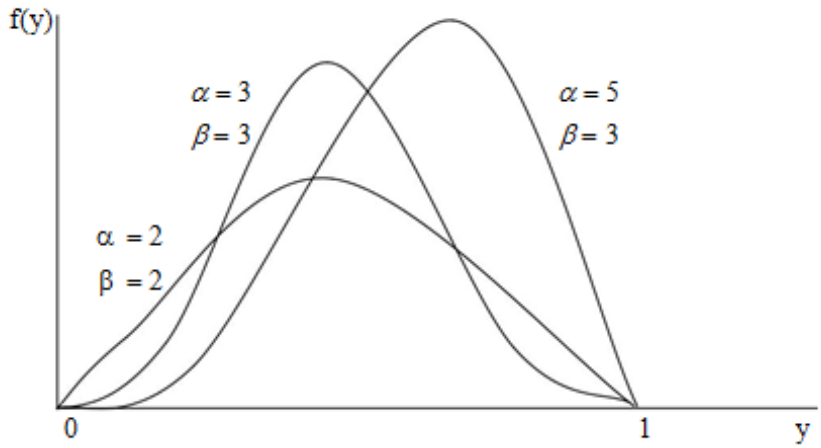
Fungsi Densitas Beta didefinisikan pada interval tutup $0 \leq y \leq 1$. Distribusi Beta sering digunakan sebagai model untuk proporsi, sebagai contoh yaitu proporsi ketakmurnian produk kimia atau proporsi waktu sebuah mesin diwaktu perbaikan. Walpole (1993) menyatakan variabel acak Y mempunyai distribusi peluang Beta dengan parameter $\alpha > 0$ *dan* $\beta > 0$, jika dan hanya jika fungsi densitas dari Y adalah;

$$f(y) = \begin{cases} \frac{y^{\alpha-1}(1-y)^{\beta-1}}{B(\alpha,\beta)} & untuk\ 0 \leq y \leq 1;\ \alpha > 0; \beta > 0 \\ 0 & untuk\ lainnya \end{cases} \qquad (9)$$

dengan fungsi Beta $B(\alpha,\beta)$ adalah;

$$B(\alpha,\beta) = \begin{cases} \int_0^1 y^{\alpha-1}(1-y)^{\beta-1}\,dy & untuk\ \alpha > 0; \beta > 0 \\ 0 & untuk\ lainnya \end{cases} \quad (9)$$

Grafik fungsi densitas Beta mengasumsikan perbedaan yang lebar dari bentuk untuk berbagai nilai dari dua parameter α dan β. Beberapa diantaranya digambarkan seperti pada **Gambar 3.2**.



Sumber: Yendra (2008)

Gambar 3.2. Grafik Fungsi Densitas Beta

3.5 Big Data

Menurut Eaton (2012), *Big Data* merupakan istilah yang berlaku untuk informasi yang tidak dapat diproses atau dianalisis menggunakan alat tradisional.

Menurut Dumbill (2012), *Big Data* adalah data yang melebihi proses kapasitas dari kovensi sistem database yang ada. Data yang terlalu besar dan terlalu cepat atau tidak sesuai dengan struktur arsitektur *database* yang ada. Untuk mendapatkan nilai dari data, maka harus memilih jalan altenatif untuk memprosesnya.

3.6 Variabel *Dummy*

Variabel dummy adalah variabel yang digunakan untuk mengkuantitatifkan variabel yang bersifat kualitatif (misal: jenis kelamin, ras, agama, perubahan kebijakan pemerintah, perbedaan situasi dan lain-lain). Variabel *dummy* merupakan variabel yang bersifat kategorikal yang diduga mempunyai pengaruh terhadap variabel yang bersifat kontinu.

Variabel *dummy* sering juga disebut variabel boneka, *binary*, kategorik atau dikotom. Penggunaan Variabel *dummy* dalam regresi dapat berupa dua kategori maupun lebih dari dua kategori.

3.7 Simulasi

Menurut Utami (2015), simulasi adalah proses implementasi model menjadi program komputer (*software*) atau rangkaian elektronik dan mengeksekusi *software* tersebut sedemikian rupa sehingga perilakunya menirukan atau menyerupai sistem nyata tertentu untuk tujuan mempelajari perilaku sistem, pelatihan atau permainan yang melibatkan sistem nyata (realitas).

Menurut Hasan (2002), simulasi merupakan suatu model pengambilan keputusan dengan mencontoh atau mempergunakan gambaran sebenarnya dari suatu sistem kehidupan dunia nyata tanpa harus mengalaminya pada keadaan yang sesungguhnya.

Simulasi memeiliki beberapa kelebihan, diantaranya;

1. Simulasi mampu menggambarkan suatu prosedur operasional untuk rentang waktu yang lebih singkat dari perencanaan.
2. Simulasi mampu menyajikan sistem nyata yang lebih besar dan rumit atau kompleks, dibandingkan dengan model matematika yang masih konvensional.
3. Dengan simulasi, penggunanya dapat menjadikan hasil simulasi sebagai pengambilan keputusan misalnya untuk penerapan sistem maupun memutuskan langkah-langkah prefentif aspek lainnya.

Selain kelebihan, tentunya simulasi juga memiliki beberapa kelemahan, seperti;

1. Simulasi bukan merupakan proses optimasi, tetapi menghasilkan cara untuk menilai suatu solusi, simulasi tidak menghasilkan solusi.
2. Pembuatan simulasi memerlukan waktu yang cukup lama mengingat harus merepresentasikan kondisi nyata dan juga biaya yang diperlukan cukup besar untuk simulasi kasus yang kompleks.
3. Tidak semua kasus dapat disimulasikan karena untuk kasus yang menuntut kepastian akan sangat sulit menggunakan simulasi.

3.8 *Software* R

R adalah suatu kesatuan *software* yang terintegrasi dengan beberapa fasilitas untuk manipulasi, perhitungan dan penampilan grafik yang handal. R berbasis pada bahasa pemrograman S, yang dikembangkan oleh *AT&T Bell Laboratories* (sekarang *Lucent Technologies*) pada akhir tahun ’70 an. R merupakan versi gratis dari bahasa S dari *software* (berbayar) yang sejenis yakni S-PLUS yang banyak digunakan para peneliti dan akademisi dalam melakukan kegiatan ilmiahnya.

R dapat berinteraksi dengan program statisik, manipulasi, perhitungan dan penampilan grafik lainnnya, seperti SPSS, yang cukup popular dan juga Microsoft Excel dengan menyediakan fasilitas impor dan ekspor data. Selain *software* di atas, R dapat melakukan impor *file* dari *software* lainnya seperti, Minitab, SAS, Stat, Systat dan EpInfo.

R mempunyai beberapa kelebihan dan fitur-fitur yang canggih dan berguna, diantaranya:

a. Efektif dalam pengelolaan data dan fasilitas penyimpanan. Ukuran file yang disimpan jauh lebih kecil dibanding *software* lainnya.
b. Lengkap dalam operator perhitungan array
c. Lengkap dan terdiri dari koleksi *tools* statistik yang terintegrasi untuk analisis data, diantaranya, mulai statistik deskriptif, fungsi probabilitas, berbagai macam uji statistik, hingga *time series*.

d. Tampilan grafik yang menarik dan fleksibel ataupun costumized.
e. Dapat dikembangkan sesuai keperluan dan kebutuhan dan sifatnya yang terbuka, setiap orang dapat menambahkan fitur-fitur tambahan dalam bentuk paket ke dalam *software* R.

BAB IV
METODELOGI PENELITIAN

4.1 Data

Pada penelitian ini peneliti menggunakan data simulasi dengan melakukan pembangkitan data simulasi menggunakan *software* R. Data disimulasikan ke dalam kasus pelaku kecelakaan sepeda motor dengan mengasumsikan terdapat enam variabel bebas yaitu penyebab kecelakaan, pendidikan, jenis kelamin, waktu kejadian, kategori usia, dan penggunaan helm dengan variabel terikat kategori biner yaitu luka-luka dan meninggal dunia.

Pengkategorian pada variabel bebas yang disertakan pada kasus pelaku kecelakaan lalu lintas didapatkan dari Direktorat Lalu Lintas Kepolisian Daerah DI Yogyakarta (Ditlantas Polda DIY) sebagai acuan pemilihan jumlah variabel bebas, variabel terikat serta jumlah kategori dalam tiap variabel seperti yang terlihat pada **Tabel 4.1**.

Tabel 4.1. Pengkategorian Variabel Bebas

Penyebab Kecelakaan (x_1)	Pendidikan (x_2)	Jenis Kelamin (x_3)	Waktu Kejadian (x_4)	Kategori Usia (x_5)	Penggunaan Helm (x_6)
1. Lengah 2. Mengantuk 3. Sakit 4. Tidak tertib 5. Tekanan psikologi 6. Pengaruh alkohol 7. Batas kecepatan	1. Sekolah Dasar 2. Sekolah Menengah Pertama 3. Sekolah Menengah Atas 4. Perguruan Tinggi 5. Lain-lain	1. Laki-laki 2. Perempuan	1. 00.01–06.00 2. 06.01–12.00 3. 12.01–18.00 4. 18.01–00.00	1. Balita 2. Kanak-kanak 3. Remaja 4. Dewasa 5. Lansia 6. Manula	1. Standar 2. Tidak Standar 3. Tidak Menggunakan Helm

dengan;

variabel bebas pertama = x_1 = 7 kategori

variabel bebas kedua = x_2 = 5 kategori

variabel bebas ketiga = x_3 = 2 kategori
variabel bebas keempat = x_4 = 4 kategori
variabel bebas kelima = x_5 = 6 kategori
variabel bebas keenam = x_6 = 3 kategori

Lalu Pengkategorian variabel bebas disimulasikan pada 9 wilayah kabupaten/kota di provinsi bali sebagai yang tertera pada **Tabel 4.2**.

Tabel 4.2. Pengkategorian Variabel Bebas pada 9 wilayah

No	Kabupaten/kota	Nama Wilayah
1	Kabupaten	Badung
2	Kabupaten	Bangli
3	Kabupaten	Buleleng
4	Kabupaten	Gianyar
5	Kabupaten	Jembrana
6	Kabupaten	Karang Asem
7	Kabupaten	Klungkung
8	Kabupaten	Tabanan
9	Kota	Denpasar

maka tiap kategori variabel bebas diulang = r = 9 kali.

Ketika dikombinasikan jumlah data keseluruhan berdasarkan x_1, x_2, x_3, x_4, x_5, x_6, dan r yaitu 7 x 5 x 2 x 4 x 6 x 3 x 9 didapatkan data sebanyak 45,360 observasi. Data variabel bebas dan perulangan dibangkitkan menggunakan sistem perulangan pada *software* R.

Untuk menyerupai realitas dengan tujuan mempelajari perilaku sistem, pelatihan atau permainan yang melibatkan sistem nyata yang disandang Utami (2015) sebagai definisi simulasi, maka data simulasi yang digunakan pada penelitian ini didasarkan pada simulasi kasus pelaku kecelakaan sepeda motor di provinsi Bali.

Selain untuk menciptakan efek *Big Data* dengan jumlah pengkategorian wilayah yang banyak yaitu sebanya 9 buah, Kasus kecelakaan sepeda motor di

provinsi Bali dipilih karena pernyataan pada artikel yang ditulis oleh Parama pada Tribun Bali (2017) yang menyatakan bahwa angka korban meninggal dunia akibat kecelakaan lalu lintas di Bali cukup mengkhawatirkan. Rata-rata terdapat 600 korban jiwa setiap tahunnya, 65% di antaranya merupakan pengendara yang masih berusia produktif. Menurut Putra (2017) pada artikel Republika, Kepolisian Daerah Bali mencatat kecelakaan lalu lintas di Bali selama tahun 2017 mencapai 1,698 kasus atau meningkat 14% jika dibandingkan dengan tahun 2016. Polisi mencatat posisi pertama kecelakaan lalu lintas melibatkan sepeda motor.

Kebijakan catatan data kecelakaan di Bali didasarkan peraturan yang sama dengan pencatatan Ditlantas Polda DIY. Hal tersebut tercatat pada Peraturan Kapolri Nomor 15 tahun 2013 tentang tata cara penanganan kecelakaan lalu lintas. Pendataan Kecelakaan Lalu Lintas tertuang pada Bab XIV Bagian Kesatu yang berisi Pasal 95 ayat 1, 2, dan 3 yang dituliskan sebagai berikut:

1. Ayat (1): Petugas yang melakukan olah tempat kejadian perkara wajib memasukkan data ke lembar formulir data kecelakaan lalu lintas.
2. Ayat (2): Formulir data kecelakaan lalu lintas sebagaimana dimaksud pada ayat (1) paling rendah berisi identitas dan jumlah korban, kondisi korban, identitas pelaku, identitas kendaraan, lokasi dan waktu kejadian, penyebab terjadinya kecelakaan, kondisi jalan, situasi lingkungan, jenis kecelakaan serta kronologis terjadinya kecelakaan lalu lintas.
3. Ayat (3): Format formulir data kecelakaan lalu lintas sebagaimana dimaksud pada ayat (2) tercantum dalam lampiran “F” yang merupakan bagian tidak terpisahkan dari peraturan ini.

Pendataan kecelakaan sepeda motor diatur dengan peraturan dan formulir yang sama untuk setiap daerah di seluruh Indonesia, sehingga pencatatan data kecelakaan sepeda motor di Bali dapat didasarkan dengan pencatatan Ditlantas Polda DIY.

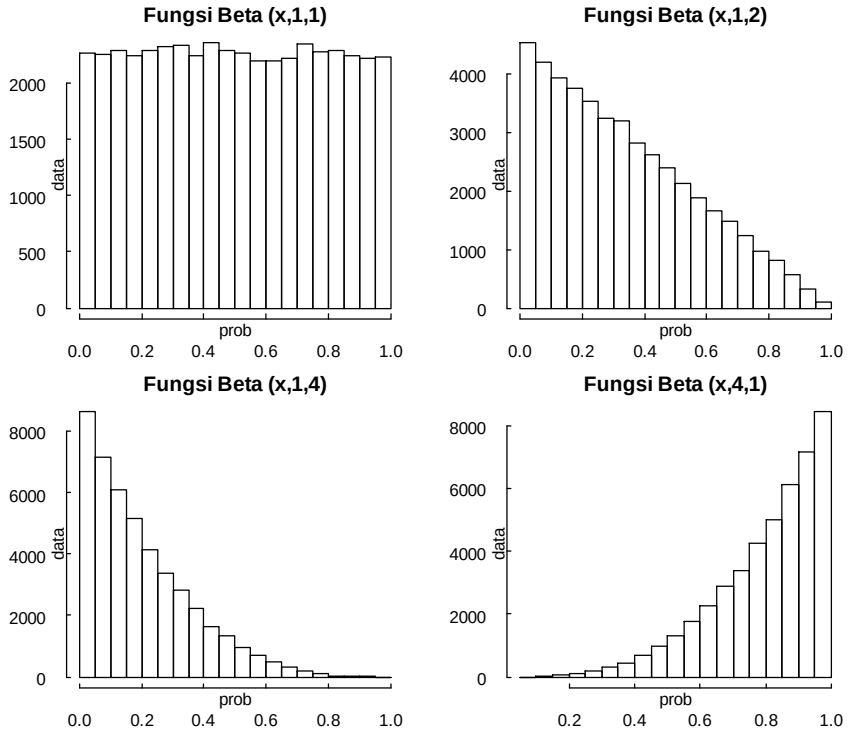
Variabel terikat pada kasus pelaku kecelakaan sepeda motor dibagi menjadi dua, yaitu luka-luka dan meninggal dunia yang didefinisikan ke dalam

variabel *y*. Data variabel terikat disimulasikan menggunakan percabangan dari nilai Distribusi Beta dengan parameter α = 1 dan β = 4 yang didefinisikan ke dalam variabel *pF*. Cara menentukan nilai y dengan menggunakan *pF* yaitu menggunakan syarat percabangan yaitu ketika $pF \leq 0.5$ maka *y* bernilai 0, sebaliknya jika $pF > 0.5$ maka *y* bernilai 1. Contoh penentuan nilai *y* dari nilai *pF* ditampilkan pada **Tabel 4.3.**

Tabel 4.3. Contoh Penentuan *y* dari Nilai *pF*

Distribusi Beta	*pF*	Penentuan	*y*
(x,1,1)	0.8965	> 0.5, maka	1
(x,1,2)	0.3487	≤ 0.5, maka	0
(x,1,4)	0.5523	> 0.5, maka	1

Untuk penentuan α dan β dapat dilihat pada **Gambar 4.1**.



Gambar 4.1. Pemilihan Nilai Parameter (x, α, β)

Pada data keputusan biner terdapat kecondongan nilai pada suatu kategori. Oleh karena itu α = 1 dan β = 1 tidak dipilih karena terdapat nilai probabilitas yang seragam sehingga ketika terdapat keputusan nilai probabiltas ≤ 0.5 masuk dalam kategori 0 dan sisanya masuk dalam kategori 1, maka akan menghasilkan

jumlah yang hampir sama. Kasus dengan jumlah data yang hampir sama pada tiap kategori jarang ditemui dikarenakan pokok permasalahan terdapat pada perbedaan nilai dari tiap kategori. α = 1 dan β = 4 dirasa cukup untuk membuat kondisi tersebut, dimana nilai probabiltas $\leq$ 0.5 untuk kategori 0 memiliki perbedaan signifikan dengan kategori 1. Untuk α = 4 dan β = 1 pada fungsi distribusi Beta juga memiliki pola data yang sama, namun kecondongan berada pada kategori yang sebaliknya yaitu kategori 0 sehingga akan memberikan hasil akurasi yang sama dengan α = 1 dan β = 4.

4.2 Variabel dan Definisi Operasional Variabel

Pada penelitian ini, himpunan data yang digunakan adalah data simulasi dari pelaku kecelakaan sepeda motor di provinsi bali, dengan definisi variabel-variabel yang digunakan seperti pada **Tabel 4.4**.

Tabel 4.4. Definisi Operasional Variabel

Varibel	Kode	Definisi Operasional Variabel
Penyebab Kecelakaan	x_1	Penyebab kecelakaan adalah suatu hal yang menjadi alasan bagi pelaku ketika menyebabkan kecelekaan sepeda motor.
Pendidikan	x_2	Pendidikan adalah tahapan jenjang pendidikan berstruktur yang ditempuh pelaku kecelakaan sepeda motor
Jenis Kelamin	x_3	Jenis kelamin adalah perbedaan biologis antara laki-laki dan perempuan
Waktu Kejadian	x_4	Waktu kejadian menandakan kapan kecelakaan terjadi, dalam hal ini terbagi menjadi 4 bagian dengan rentang waktu masing-masing selama 6 jam
Kategori Usia	x_5	Kategori usia merupakan sarana pembeda pelaku dengan kelompok umur tertentu
Penggunaan Helm	x_6	Penggunaan helm yaitu penggunaan pelindung kepala dalam mengendarai sepeda motor, baik menggunakan helm standar, tidak standar maupun tidak menggunakan helm.
Wilayah	r	Wilayah yaitu tempat dimana kecelakaan berlangsung, didefinisikan ke dalam 9 kategori yang merupakan kabupaten/kota dari provinsi Bali
Kategori Luka	y	Luka pelaku terbagi menjadi dua macam yaitu luka-luka dan meninggal dunia, korban yang meninggal dunia dimasukkan ke dalam kategori luka berat

4.3 Metode Analisis Data

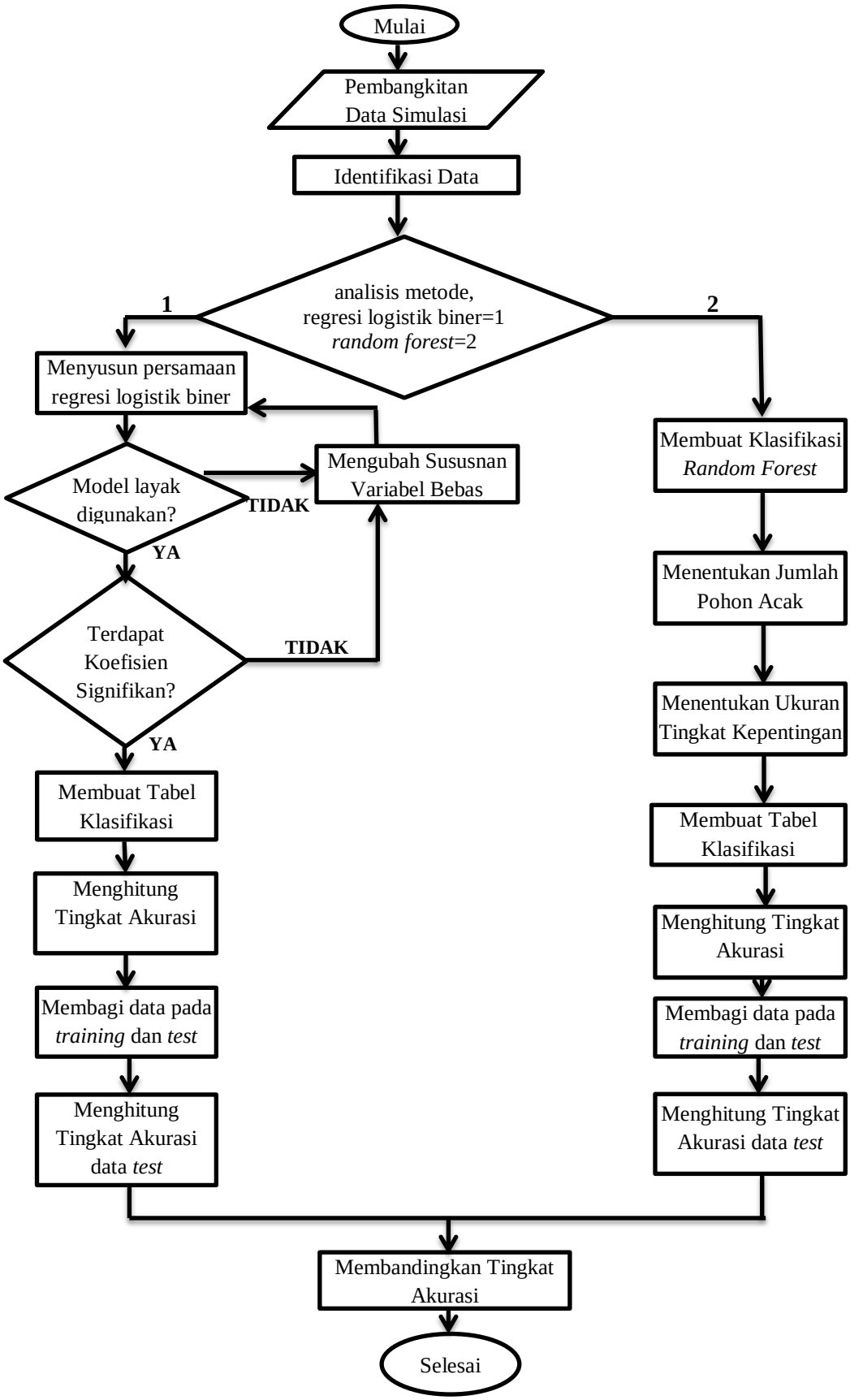
Penelitian ini menggunakan metode analisis Regresi Logistik Biner dan *Random Forest*. Proses analisis data dilakukan dengan bantuan perangkat lunak atau *software* R 3.2.4.

4.4 Tahapan Penelitian

Penelitian ini menggunakan metode analisis Regresi Logistik Biner dan *Random Forest*. Proses analisis data dilakukan dengan bantuan perangkat lunak atau *software* R 3.2.4.

Adapun tahapan dalam melakukan perbandingan Regresi Logistik Biner denga *Random Forest* adalah sebagai berikut:

1. Membuat data simulasi dengan pembangkitan menggunakan *software* R.
2. Mengestimasikan data yang telah dibuat menggunakan model Regresi Logistik Biner.
3. Melakukan pengujian simultan dan parsial pada Regresi Logistik Biner.
4. Membuat tabel klasifikasi untuk melihat efektifitas pemodelan Regresi Logistik Biner.
5. Menghitung tingkat akurasi dari model Regresi Logistik Biner.
6. Mengestimasikan data yang telah dibuat menggunakan *Random Forest*.
7. Membuat plot *Mean Decrease Accuracy (MDA)* dan *Mean Decrease Gini (MDG)* untuk *Random Forest*.
8. Membuat tabel klasifikasi untuk melihat efektifitas pemodelan *Random Forest*.
9. Menghitung tingkat akurasi dari model *Random Forest*.
10. Menentukan *training* dan *test sample* pada data.
11. Membandingkan tingkat akurasi dari Regresi Logistik Biner dan *Random Forest* menggunakan plot.
12. Interpretasi perbandingan metode Regresi Logistik Biner dan *Random Forest*.



Gambar 4.2. Diagram Alir Penelitian

BAB V
HASIL DAN PEMBAHASAN

5.1. Pembangkitan Data

Untuk membandingkan Regresi Logistik Biner dengan *Random Forest* maka diperlukan suatu data untuk dianalisis. Dalam hal ini penulis menggunakan *software* R sebagai alat analisis, dimana *software* tersebut dapat digunakan untuk membangkitkan data atau membuat data dengan spesifikasi sesuai yang diinginkan. Data variabel bebas dibangkitkan menggunakan sistem perulangan pada R, untuk variabel terikat dibangkitkan menggunakan Distribusi Beta dengan nilai parameter $\alpha = 1$ dan $\beta = 4$.

```
> str(fD)
'data.frame':   45360 obs. of  8 variables:
 $ x1  : num  1 1 1 1 1 1 1 1 1 1 ...
 $ x2  : num  1 1 1 1 1 1 1 1 1 1 ...
 $ x3  : num  1 1 1 1 1 1 1 1 1 1 ...
 $ x4  : num  1 1 1 1 1 1 1 1 1 1 ...
 $ x5  : num  1 1 1 1 1 1 1 1 1 1 ...
 $ x6  : num  1 1 1 1 1 1 1 1 1 2 ...
 $ pF  : num  0.0672 0.0782 0.0403 0.0147 0.0287 ...
 $ y   : num  0 0 0 0 0 0 0 0 0 0 ...
```

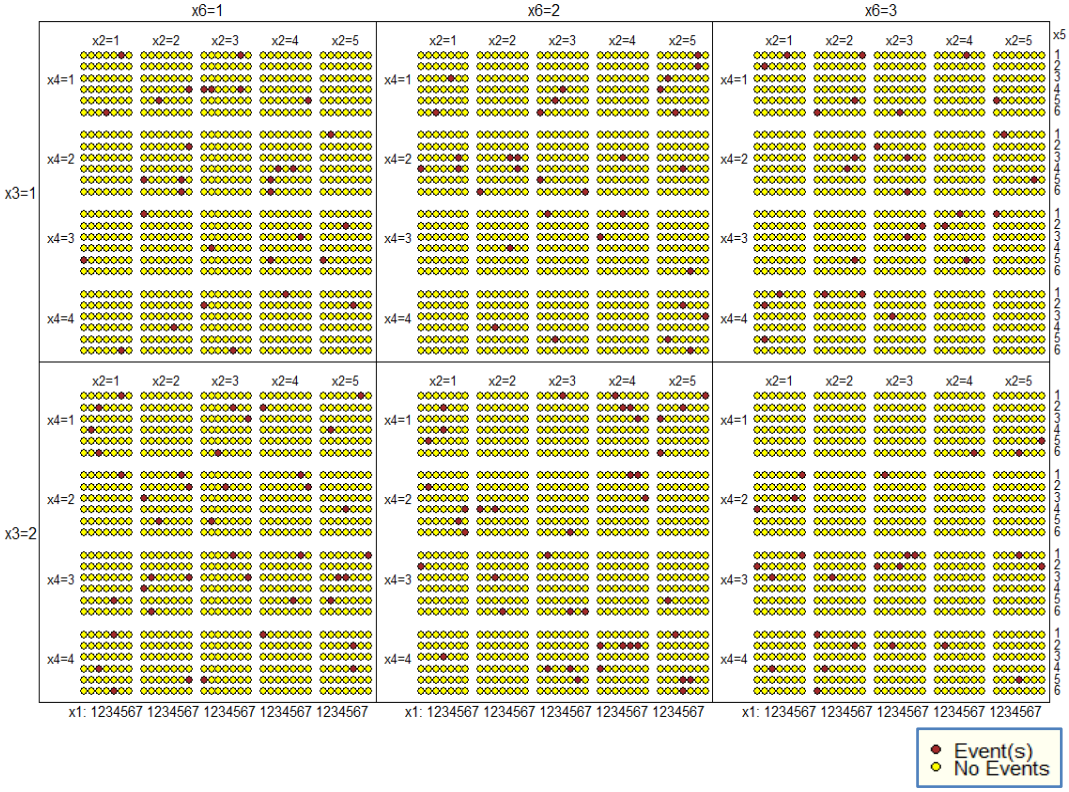
Gambar 5.1. Tipe Data Variabel-variabel yang Digunakan

Pada **Gambar 5.1** terdapat 45,360 observasi dan 8 variabel yang digunakan dalam analisis. Keseluruhan variabel yang digunakan memiliki tipe data numerik yaitu x_1, x_2, x_3, x_4, x_5, x_6, *pF*, dan *y*. Dengan keterangan x_1, x_2, x_3, x_4, x_5, x_6 sebagai variabel bebas, *pF* sebagai nilai data yang dibangkitkan menggunakan distribusi Beta pada $\alpha = 1$ dan $\beta = 4$ yang digunakan untuk mendapatkan nilai *y* dan *y* sebagai variabel terikat dengan kategori 0 dan 1.

```
> summary(fD)
       x1              x2              x3                  x4
 Min.   :1     Min.   :1   Min.   :1.0         Min.   :1.00

 Max.   :7     Max.   :5   Max.   :2.0         Max.   :4.00

       x5              x6              pF                  y
 Min.   :1.0   Min.   :1   Min.   :0.0000004   Min.   :0.00000

 Max.   :6.0   Max.   :3   Max.   :0.8794322   Max.   :1.00000
```

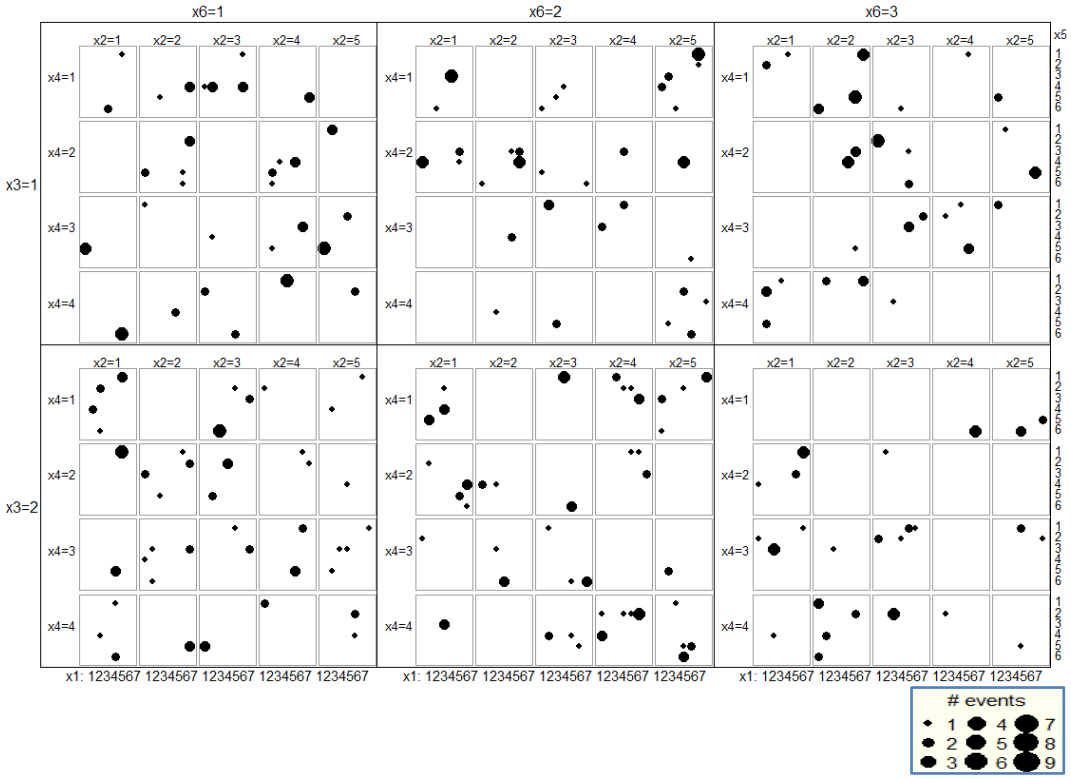
Gambar 5.2. Ringkasan Data

Kemudian dilakukan ringkasan data untuk mendeskripsikan data sesuai dengan kelasnya. Dari **Gambar 5.2** dapat dilihat bahwa x_1, x_2, x_3, x_4, x_5, x_6 sebagai variabel bebas dengan nilai minimum 1 sampai dengan maksimum jumlah kategori masing-masing yaitu 7, 5, 2, 4, 6, dan 3. kemudian *pF* yang merupakan data berdistribusi Beta yang dibangkitkan untuk mencapai kondisi variabel *y* dan terakhir yaitu *y* dengan hanya 2 kategori yang berasal dari pengkondisian data *pF* dengan ketentuan jika nilai $pF \leq 0.5$ akan bernilai 0 dan $pF > 0.5$ bernilai 1.



Gambar 5.3. Tabulasi Silang Data dengan Keterangan Kategori Variabel Terikat

Pada **Gambar 5.3** dapat dilihat gambar tabulasi silang data sebanyak 45360 observasi yang diambil dari kombinasi kategori pada tiap variabel bebas yaitu x_1, x_2, x_3, x_4, x_5, dan x_6. *Event(s)* dan *No Events* menandakan kategori pada variabel terikat *Event(s)* ketika bernilai 1 dan *No Events* jika bernilai 0. Pada **Gambar 5.1** *Event(s)* ditandai dengan data berwarna merah, sedangkan *No Events* ditandai dengan warna kuning.



Gambar 5.4. Sebaran *Event(s)* pada Tabulasi Silang Data

Sebaran Event(s) pada perulangan data yang dilakukan sebanyak 9 kali untuk masing-masing kategori pada variabel bebas dapat dilihat pada **Gambar 5.4**. Kondisi perulangan dibedakan dengan tanda bulatan hitam yang berbeda ukuran. Data perulangan pertama ditandai dengan bulatan hitam paling kecil, lalu perulangan kedua dan seterusnya sampai perulangan sembilan yang memiliki bulatan hitam paling besar. Dengan melihat data *Event(s)* pada **Gambar 5.4** yang menyebar keseluruh kategori dan perulangan maka dapat dikatakan bahwa pengembangan data sudah cukup baik.

5.2. Regresi Logistik Biner

Setelah pembangkitan data selesai dilakukan, maka data simulasi tersebut dapat dianalisis menggunakan metode Regresi Logistik Biner. Jika terdapat variabel bebas kategorik dalam analisis Regresi Logistik Biner, maka variabel tersebut akan dimasukkan kedalam model regresi dengan pengkondisian *dummy*.

```
> contrasts(fD$x1)
  2 3 4 5 6 7
1 0 0 0 0 0 0
2 1 0 0 0 0 0
3 0 1 0 0 0 0
> contrasts(fD$x2)
  2 3 4 5
1 0 0 0 0
2 1 0 0 0
3 0 1 0 0
> contrasts(fD$x3)
  2
1 0
2 1
```

```
4 0 0 1 0 0 0       4 0 0 1 0
5 0 0 0 1 0 0       5 0 0 0 1
6 0 0 0 0 1 0
7 0 0 0 0 0 1

> contrasts(fD$x4) > contrasts(fD$x5)  > contrasts(fD$x6)
  2 3 4              2 3 4 5 6           2 3
1 0 0 0            1 0 0 0 0 0         1 0 0
2 1 0 0            2 1 0 0 0 0         2 1 0
3 0 1 0            3 0 1 0 0 0         3 0 1
4 0 0 1            4 0 0 1 0 0
                   5 0 0 0 1 0
                   6 0 0 0 0 1
```

Gambar 5.5. Pengkondisian *Dummy*

Pada **Gambar 5.5** dapat dilihat bahwa kategori pertama untuk masing-masing variabel bebas dijadikan sebagai kategori basis oleh *software* R. Variabel basis adalah variabel yang dijadikan acuan dalam *dummy*. Ketika mengolah data menggunakan *Software* R, maka kategori 1 dari tiap variabel otomatis akan terpilih sebagai variabel basis. Dalam Regresi Logistik Biner, sisa kategori yang tidak dijadikan basis (k-1) akan dimasukkan sebagai koefisien dalam model Regresi. Sehingga koefisien-koefisien regresi yang dihasilkan yaitu x_{12}, x_{13}, x_{14}, x_{15}, x_{16}, x_{17}, x_{22}, x_{23}, x_{24}, x_{25}, x_{32}, x_{42}, x_{43}, x_{44}, x_{52}, x_{53}, x_{54}, x_{55}, x_{56}, x_{62}, dan x_{63}.

```
> library(aod)
> wald.test(b = coef(fD.lr), Sigma = vcov(fD.lr), Terms =
1:6)
Wald test:
----------

Chi-squared test:
X2 = 1494.2, df = 6, P(> X2) = 0.0
```

Gambar 5.6. Uji Simultan

Dari **Gambar 5.6** terdapat pengujian simultan pada Regresi Logistik Biner. Menggunakan *packages* aod pada R, dapat dilihat nilai chi-square (χ^2) dengan derajat bebas yaitu df (*degree of freedom*) dan tingkat signifikasinya (*p-value*). Uji simultan berfungsi untuk menguji keseluruhan model dari Regresi Logistik Biner yang digunakan. Pengujian hipotesisnya dilakukan sebagai berikut:

a. Hipotesis

   $H_0 : \forall \delta_i = 0$ (Model tidak layak)

   $H_1 : \exists \delta_i \neq 0$ (Model layak digunakan)

b. Tingkat Signifikasi

γ = 5%

c. Statistik Uji

P-value = 0,0

d. Daerah Kritis

H_0 ditolak jika *P-value* < γ

e. Keputusan

P-value (0,0) < γ (0,05) sehingga H_0 ditolak

f. Kesimpulan

Dengan tingkat signifikansi γ = 5% maka $\exists \delta_i \neq 0$ dan dapat disimpulkan bahwa model sesuai.

Lalu dilakukan pengujian pengujian parsial pada Regresi Logistik Biner. Pengujian dilakukan untuk melihat apakah tiap parameter layak digunakan dalam model. Dengan pengujian hipotesis sebagai berikut:

a. Hipotesis

$H_0 : \delta_i = 0$ (parameter δ_i tidak layak dalam model)

$H_1 : \delta_i \neq 0$ (parameter δ_i layak dalam model)

b. Tingkat Signifikasi

γ = 5% = 0,05

c. Statistik Uji

P-value pada **Tabel 5.1**

d. Daerah Kritis

H_0 ditolak jika *P-value* < γ

e. Keputusan

Tabel 5.1. Keputusan Uji Parsial

Koefisien	*P-value*	< / ≥ γ = 0.05	Keputusan
Intercept	0.0000	< 0.05	Tolak H_0
x_{12}	0.6550	≥ 0.05	Gagal tolak H_0
x_{13}	0.0278	< 0.05	Tolak H_0
x_{14}	0.8597	≥ 0.05	Gagal tolak H_0
x_{15}	1.0000	≥ 0.05	Gagal tolak H_0
x_{16}	0.0262	< 0.05	Tolak H_0
x_{17}	0.9298	≥ 0.05	Gagal tolak H_0
x_{22}	0.3900	≥ 0.05	Gagal tolak H_0
x_{23}	0.2472	≥ 0.05	Gagal tolak H_0

Koefisien	*P-value*	$< / \geq \gamma = 0.05$	Keputusan
x_{24}	0.0647	≥ 0.05	Gagal tolak H_0
x_{25}	0.3506	≥ 0.05	Gagal tolak H_0
x_{32}	0.8878	≥ 0.05	Gagal tolak H_0
x_{42}	0..4565	≥ 0.05	Gagal tolak H_0
x_{43}	0.0000	< 0.05	Tolak H_0
x_{44}	0.0000	< 0.05	Tolak H_0
x_{52}	0.0451	< 0.05	Tolak H_0
x_{53}	0.0123	< 0.05	Tolak H_0
x_{54}	0.0786	≥ 0.05	Gagal tolak H_0
x_{55}	0.0370	< 0.05	Tolak H_0
x_{56}	0.1101	≥ 0.05	Gagal tolak H_0
x_{62}	0.0000	< 0.05	Tolak H_0
x_{63}	0.0000	< 0.05	Tolak H_0

f. Kesimpulan

Dengan tingkat signifikansi γ = 5% dapat disimpulkan bahwa terdapat parameter δ_i yang layak maupun tidak layak dalam model.

Berdasarkan pengujian parsial yang telah dilakukan, terdapat beberapa koefisien Regresi yang tidak signifikan. Umumnya, koefisien model dalam Regresi mewakili masing-masing variabel bebas. Ketika terdapat kasus pada pengujian parsial dimana salah satu koefisien tidak signifikan, maka akan dilakukan pengujian ulang dengan tidak mengikutsertakan variabel yang diwakilinya.

Namun untuk kasus *dummy*, koefisien model Regresi yang dihasilkan mewakili masing-masing kategori kecuali kategori yang dijadikan basis dengan *dummy*-nya pada masing-masing variabel. Jika penulis ingin menghilangkan salah satu koefisien yang tidak signifikan, maka koefisien lain yang dihasilkan pada variabel yang sama juga harus dibuang.

Untuk kasus pengujian parsial, pada **Tabel 5.1** dapat dilihat bahwa variabel yang kesemua koefisiennya signifikan hanyalah x_6. Sehingga akan sangat mengurangi informasi dari data jika variabel lainnya tidak dimasukkan dalam penelitian. Selain alasan tersebut, pengolahan data menggunakan Regresi Logistik Biner tersebut juga akan dibandingkan dengan *Random Forest*. Untuk membandingkan, tentu data pada dua metode tersebut lebih baik tetap menggunakan variabel yang sama, selain itu data yang digunakan yaitu data

simulasi yang tidak berbasis data *real*. Maka dari itu, Pemenuhan signifikansi variabel tidak wajib terpenuhi, maka analisis tetap dilanjutkan pada analisis Regresi Logistik Biner dengan 6 variabel bebas.

```
> addmargins(cTab.lr)

pred.fD.lr     0     1   Sum
       0   27438   147 27585
       1   17463   312 17775
       Sum 44901   459 45360
```

Gambar 5.7. Hasil Klasifikasi Regresi Logistik Biner

Hasil klasifikasi dapat dilihat pada **Gambar s5.7** dengan angka banyaknya observasi pada masing-masing kategori variabel terikat. Pada awalnya data dengan variabel terikat berkategori 0 ada sebanyak 44,901 observasi. Sedangkan untuk kategori 1 ada sebanyak 459 observasi. Namun hasil pada Regresi Logistik Biner memprediksikan ada sebanyak 27,585 pada kategori 0 dan 17,775 pada kategori 1. Hal tersebut menyatakan adanya kesalahan prediksi pada observasi yang diteliti.

Regresi Logistik Biner berhasil memprediksi kategori 0 pada 27,438 observasi. Sisanya yaitu 17,463 observasi salah diprediksikan sebagai observasi 1. Untuk data dengan kategori 1 dengan jumlah observasi 459 berhasil memprediksikan 312 observasi. Sedangkan sisanya sebanyak 147 salah diperkirakan sebagai data dengan kategori 0.

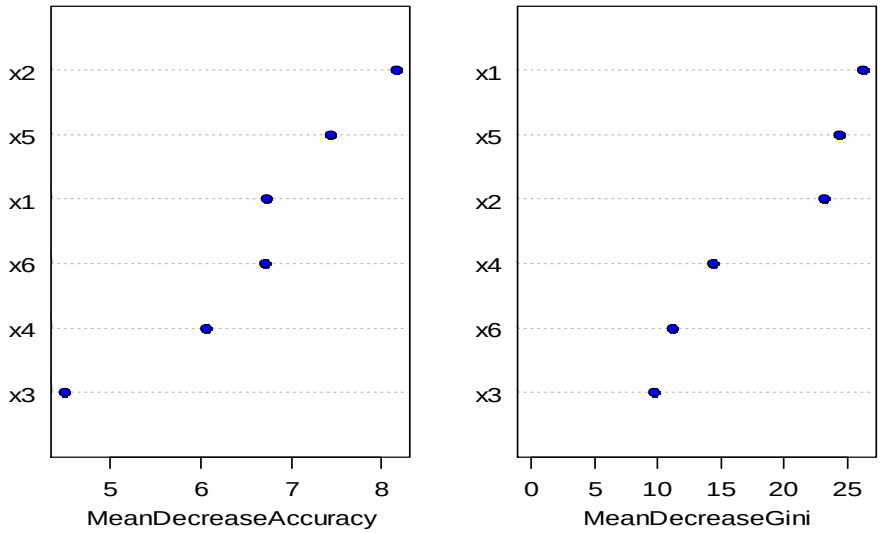
```
> pa.fD.lr <- 100*sum(diag(cTab.lr))/sum(cTab.lr)
> pa.fD.lr
[1] 61.17725
```

Gambar 5.8. Tingkat Akurasi Regresi Logistik Biner

Tingkat akurasi pada **Gambar 5.8** berasal dari hasil klasifikasi yang tertera pada tabel klasifikasi yang dihitung pada **Gambar 5.7**. Nilai presentase akurasi dinyatakan sebanyak 61.17725 yang berarti ketepatan model pada analisis Regresi Logistik Biner yang diteliti adalah sebesar 61.17725%.

5.3. *Random Forest*

Setelah *output* dari Regresi Logistik Biner telah dihasilkan, maka data kembali diolah dengan alat analisis lainnya yaitu *Random Forest*.



Gambar 5.9. *Mean Decrease Accuracy (MDA)* dan *Mean Decrease Gini (MDG)*

Pada penerapannya, *Random Forest* menghasilkan ukuran tingkat kepentingan (*variable importance*) pada masing-masing variabel bebas yaitu *Mean Decrease Accuracy (*MDA) dan *Mean Decrease Gini* (MDG). Plot dari kedua ukuran tersebut ditampilkan pada **Gambar 5.9**.

MDA menampilkan seberapa besar tambahan observasi yang mengalami misklasifikasi jika satu persatu variabel bebas tidak diikutsertakan kedalam pengujian. Dalam hal ini, x_2 memiliki penurunan terbanyak mencapai 8 observasi. Yang artinya variabel x_2 memiliki peran penting sebagai variabel bebas yang memengaruhi variabel terikat pada *Random Forest*. Sedangkan pada peringkat terakhir, ada x_3 dengan misklasifikasi paling kecil diantara variabel lain, yang menandakan bahwa tingkat kepentingan variabel x_3 dalam pengujian sangatlah kecil. mendekati nilai 0. sangat dianjurkan untuk digunakan dalam analisis.

Ukuran kepentingan lainnya yaitu MDG. Ukuran tersebut digunakan untuk melihat kestabilan variabel bebas baik x_1 hingga x_6. Tingkat kepentingan disusun

berdasarkan peringkat. Variabel bebas yang mempunyai tingkat kepentingan paling tinggi secara berurutan adalah $x_{1,}$ lalu disusul x_5, x_2 dan seterusnya sampai variabel terakhir yaitu x_3. Ketiga variabel dengan nilai tertinggi memiliki selisih yang tidak jauh berbeda yakni pada rentang 20-25. Sehingga dapat disimpilkan bahwa berdasarkan MDG, x_1 lalu x_5 dan x_2 merupakan variabel bebas paling stabil dan penting dalam pengujian *Random Forest*.

```
> addmargins(cTab.rf)

pred.fD.rf     0     1   Sum
       0   43748   234 43982
       1    1153   225  1378
       Sum 44901   459 45360
```

Gambar 5.10. Hasil Klasifikasi *Random Forest*

Hasil klasifikasi menggunakan *Random Forest* dapat dilihat pada **Gambar 5.10**. Terdapat 44,901 observasi yang termasuk dalam data berkategori 0. Sedangkan untuk kategori 1 ada sebanyak 459 observasi. Namun hasil pada pengujian memprediksikan ada sebanyak 43,982 pada kategori 0 dan 1,378 pada kategori 1. Hal tersebut menyatakan adanya kesalahan prediksi pada observasi yang diteliti.

Kategori 0 berhasil diprediksi pada 43,748 observasi. Sisanya yaitu 1,153 observasi salah diprediksikan sebagai observasi 1. Untuk data dengan kategori 1 dengan jumlah observasi 459 berhasil memprediksikan 225 observasi. Sedangkan sisanya sebanyak 234 salah diperkirakan sebagai data dengan kategori 0.

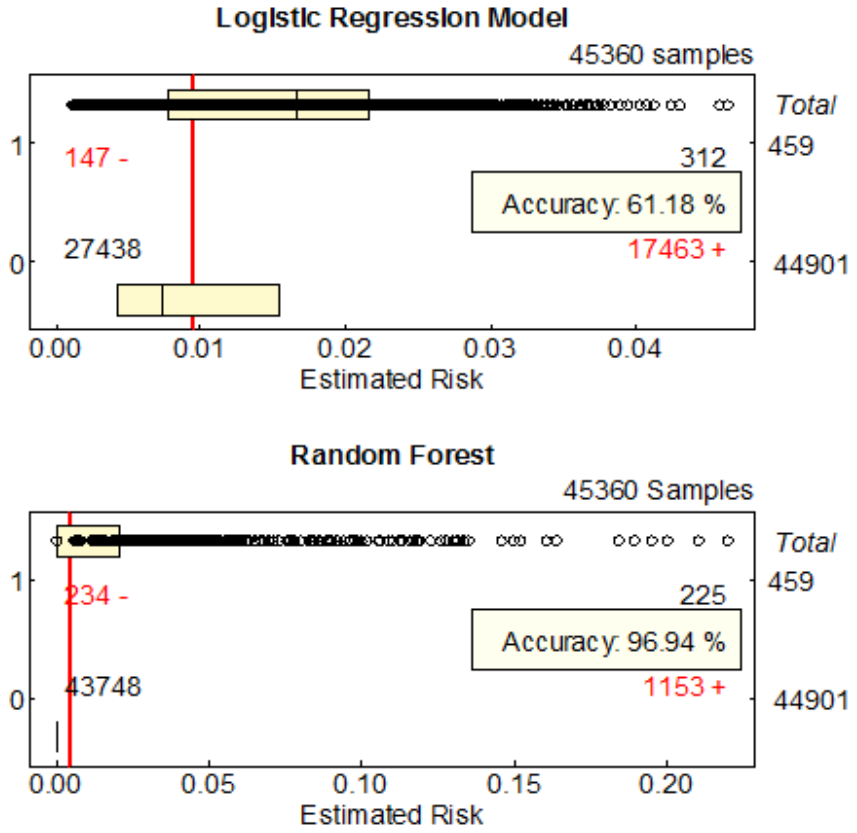
```
> pa.fD.rf <- 100*sum(diag(cTab.rf))/sum(cTab.rf)
> pa.fD.rf
[1] 96.94224
```

Gambar 5.11. Tingkat Akurasi *Random Forest*

Tingkat akurasi pada hasil klasifikasi *Random Forest* dimunculkan pada **Gambar 5.11**. Nilai presentase akurasi dinyatakan sebanyak 96.94224 yang berarti ketepatan model pada klasifikasi *Random Forest* adalah sebesar 96.94224%.

5.4. Perbandingan Regresi Logistik Biner dengan *Random Forest*

Berdasarkan pengujian yang telah dilakukan, terdapat nilai akurasi dari masing-masing metode. nilai tersebut kemudian digunakan untuk membandingkan metode Regresi Logistik Biner dengan *Random Forest*.



Gambar 5.12. Perbandingan Hasil Akurasi Regresi Logistik Biner dan *Random Forest*

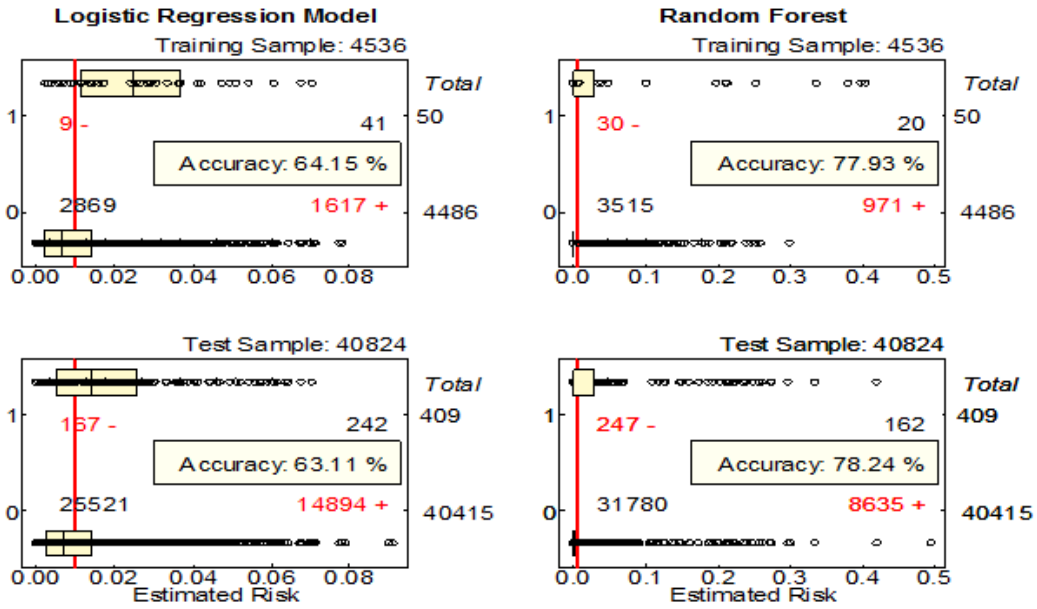
Gambar 5.12 menampilkan perbandingan tingkat akurasi dari metode Regresi Logistik Biner dengan *Random Forest*. Pada data yang sama yaitu sebanyak 45,360 observasi, 459 merupakan data dengan kategori 1 dan 44,901 observasi untuk kategori 0. Kedua metode menghasilkan hasil klasifikasi yang berbeda. Data yang digunakan terbagi menjadi 2 kategori dimana kategori 0 memiliki jumlah observasi yang lebih mendominasi daripada kategori 1 sehingga faktor tersebut juga dapat menjadi alasan mengapa tingkat akurasi tidak begitu mewakili.

Jika kedua hasil hanya dibandingkan dengan melihat kategori 1, maka Regresi Logistik Biner memprediksi lebih baik yaitu tepat sebanyak 312 observasi

dibandingkan *Random Forest* yang hanya 225 observasi dari total 459. Namun jika hanya dibandingkan berdasarkan kategori 0, maka *Random Forest* jauh lebih unggul dengan kesalahan prediksi sebanyak 1,153 observasi dibandingkan Regresi Logistik Biner yang tidak tepat memprediksikan kategori 0 pada 17,463 observasi.

Tingkat akurasi diambil dari jumlah data yang berhasil diprediksi sesuai kondisi aslinya. Dikarenakan ketimpangan jumlah pada data kategori 0 dan 1 dengan selisih maka metode yang lebih berhasil memprediksi kategori 0 akan mendominasi tingkat akurasi. Berdasarkan nilai akurasi yang ditampilkan pada **Gambar 5.12** yaitu 61.18% untuk Regresi Logistik Biner dan 96.94% untuk *Random Forest*. Maka *Random Forest* dinyatakan sebaga metode yang memprediksikan lebih baik daripada Regresi Logistik Biner pada kasus ini.

Namun mempertimbangkan ketimpangan data observasi yang menjadi salah satu faktor keraguan dalam menilai tingkat akurasi tersebut serta untuk melakukan prediksi yang lebih baik, maka data akan diolah kembali dengan membagi data menjadi dua bagian yaitu data *training* dan data *test*. Data *training* digunakan dalam analisis dan hasilnya kemudian diterapkan pada data *test*.



Gambar 5.13. Perbandingan Regresi Logistik Biner dan *Random Forest* dalam Data *Training* dan *Test*

Data dibagi menjadi dua bagian yaitu data *training* dan data *test*. Kondisi yang diterapkan pada pembagian yang terlihat pada **Gambar 5.13** yaitu dengan proporsi 10:90. 10% data untuk *training* dan sisanya 90% dijadikan data *test*. Perbandingan proporsi didasarkan pada selisih tingkat akurasi *Regresi Logistik Biner* dan *Random Forest* pada data empiris yang cukup besar yaitu 35.76%. Maka dipilih data *training* yang sangat kecil yaitu 10% dari data empiris untuk mempertimbangkan ketimpangan jumlah observasi.

Data *training* dan *test* hasil selisih data *train* yang pemilihan **Gambar 5.13** menampilkan perbandingan tingkat akurasi dari metode Regresi Logistik Biner dengan *Random Forest* untuk masing-masing data *training* dan *test*.

Pada data training, Regresi Logistik Biner menampilkan tingkat akurasi sebesar 64.15% yang kemudian turun dengan selisih yang kecil pada data *test* yaitu akurat sebesar 63.11%. Tingkat akurasi pada data *training Random Forest* adalah 77.93% yang kemudian naik dengan selisih yang kecil yaitu sebesar 78.24% pada data *test*.

Pada data *training* dengan 4,536 observasi, 50 merupakan data dengan kategori 1 dan 4486 observasi untuk kategori 0. Berdasarkan nilai akurasi yang ditampilkan pada data *test* **Gambar 5.13** yaitu 63.11% untuk Regresi Logistik Biner dan 78.24% untuk *Random Forest*. Jika dibandingkan dengan hasil pada **Gambar 5.12** data dengan kategori 1 masih mendominasi, namun tingkat akurasi untuk *Random Forest* memiliki hasil yang sangat berbeda yaitu dari 96.94% menjadi 78.24%.

BAB VI
PENUTUP

6.1 Kesimpulan

Dari berbagai hal yang telah dilakukan oleh peneliti, maka dapat disimpulkan beberapa hal sebagai berikut:

1. Penerapan Regresi Logistik Biner serta *Random Forest* pada penelitian ini dilakukan pada data observasi yang dibangkitkan menggunakan *software* R. Dilakukan uji simultan, uji parsial, serta pengklasifikasian pada Regresi Logistik Biner dengan tingkat akurasi sebesar 61.18%. *Random Forest* dilakukan dengan pertimbangan MDA dan MDG lalu kemudian dilakukan pengklasifikasian dengan tingkat akurasi sebesar 96.94%
2. Perbandingan kemampuan Regresi Logistik Biner dengan *Random Forest* setelah dianalisis dilihat pada tingkat akurasi. Terdapat selisih persentase yang besar yaitu 35.76% yang diungguli oleh *Random Forest*. Setelah dianalisis kembali dengan pembagian data ke dalam data *training* dan data *test* didapatkan tingkat akurasi data *test* untuk Regresi Logistik Biner sebesar 63.11% dan 78.24% untuk *Random Forest*. Pada analisis ini *Random Forest* kembali mengungguli dengan selisih 15.13%. Maka dari itu dapat disimpulkan bahwa *Random Forest* merupakan metode yang lebih baik ketika dihadapkan pada *Big Data* pada kasus ini.

6.2 Saran

Pada penelitian ini pengujian parsial Regresi Logistik Biner menyatakan terdapat beberapa koefisien Regresi yang tidak signifikan, untuk penelitian selanjutnya diharapkan semua asumsi yang dilibatkan dapat terpenuhi. Selain itu pada penelitian selanjutnya Regresi Logistik Biner dan *Random Forest* dapat dikembangkan juga pada *Big Data* yang sesungguhnya untuk menganalisis secara menyeluruh dan melihat bagaimana pengaruh penggunaan masing-masing metode. Ada baiknya pula untuk membandingkan metode-metode lain dalam Statistika maupun *Machine Learning* untuk melihat efektifitas metode dari kedua disiplin ilmu tersebut.

DAFTAR PUSTAKA

Agresti, Alan. 2007. *An Introduction to Categorical Data Analysis, 2nd Edition*, New York: John Wiley & Sons.

Anova, Nikolas. 2013. *Statistical Thinking di Era Big Data*. Ponorogo – Jawa Timur. https://www.kompasiana.com/nikolas_anova/statistical-thinking-di-era-big-data_5528b144f17e6191788b45cf. Diakses tanggal 18 Desember 2017 jam 15:20 WIB.

Badan Pusat Statistik. 2000. *Statistik Indonesia Tahun 2000*. Jakarta Pusat : Badan Pusat Statistik.

Biau, Gerard. 2012. *Analysis of a random forests model, Journal of Machine Learning Research*, Vol. 13, pp. 1063-1095.

Breiman, Leo. 2000. *Statistical modeling: The two cultures (with comments and a rejoinder by the author)*. Statistical science, 16(3):199–231.

Dewi, Nariswari Karina., Syafitri, Utami Dyah., dan Mulyadi, Soni Yadi. 2011. *Penerapan Metode Random Forest dalam Driver Analysis*. Bogor: Institut Pertanian Bogor.

Diaprina, Sistya Rosi dan Suhartono. 2014. *Analisis Klasifikasi Kredit Menggunakan Regresi Logistik Biner dan Radial Basis Function Network di Bank "X" Cabang Kediri*. Surabaya: Institut Teknologi Sepuluh November.

Direktorat Lalu Lintas Kepolisisan Negara Republik Indonesia Daerah DI Yogyakarta. 2018. *Kasus Kecelakaan Sepeda Motor DIY 2017*. Yogyakarta: Ditlantas Polda DIY.

Dumbill, E. 2012. *Big Data Now: 2012 Edition*. O'Reilly Media Inc.

Eaton, C., Dirk, D., Tom, D., George, L., & Paul, Z. 2012. *Understanding Big Data*. Mc Graw Hill.

Fathony, Rizal. 2015. *Statistika dan Machine Learning: Satu Ilmu Dua Wajah*. Chicago : University of Illinois Chicago.

Ghozali, Imam. 2005. *Aplikasi Analisis Multivariate Dengan Program SPSS*. Semarang: Penerbit Universitas Dipenegoro.

Hasan, M. Iqbal. 2002. *Pokok-pokok Materi Metodologi Penelitian dan Aplikasinya*. Bogor: Ghalia Indonesia.

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

Ilham, Fajar P., Wahidah, Mardiana Nur., Khoiriyah, Qudhrotul Zahro'., dan Fauzianizahra, Anindya. 2017. *Aplikasi Big Data pada Airline On-time Performance 2005 dengan Regresi Logistik Biner*. Yogyakarta: Universitas Gadjah Mada.

Mambang dan Byna, Agus. 2017. *Analisis Perbandingan Algoritma C4.5, Random Forest Dengan CHAID Decision Tree Untuk Klasifikasi Tingkat Kecemasan Ibu Hamil*. Yogyakarta: Universitas Amikom Yogyakarta.

Manggala, Yudha. 2017. Kecelakaan Lalu Lintas Bali Meningkat 14 Persen. Denpasar: Republika. http://www.republika.co.id/berita/nasional/daerah/17/12/28/p1nuql284-kecelakaan-lalu-lintas-bali-meningkat-14-persen. Diakses tanggal 20 Maret 2018 jam 06:25 WIB.

Nugroho, Yusuf Sulistyo dan Emiliyawati, Nova. 2017. *Sistem Klasifikasi Variabel Tingkat Penerimaan Konsumen Terhadap Mobil Menggunakan Metode Random Forest*. Surakarta: Universitas Muhammadiyah Surakarta.

Parama, Satya. 2017. *Angka Korban Tewas Kecelakaan Lalu Lintas di Bali Mengkhawatirkan, Jasa Raharja Lakukan Ini*. Bali: Tribun Bali. http://bali.tribunnews.com/2017/11/15/angka-korban-tewas-kecelakaan-lalu-lintas-di-bali-mengkhawatirkan-jasa-raharja-lakukan-ini?page=all. Diakses tanggal 20 Maret 2018 jam 07:45 WIB.

Peraturan Kepala Kepolisian Negara Republik Indonesia Nomor 15. 2013. *Tata Cara Penanganan Kecelakaan Lalu Lintas*. Jakarta: Kapolri.

Permana, Yana. 2016. *Mengenal Big Data*. https://www.codepolitan.com/mengenal-big-data. Diakses tanggal 20 Maret 2018 jam 06:09 WIB.

Rumaendra, Wella. 2016. *Perbandingan Klasifikasi Penyakit Hipertensi Menggunakan Regresi Logistik Biner dan Algoritma C4.5*. Semarang: Universitas Dipenegoro.

Tampil, Yumira Adriani., Komalig, Hanny., dan Langi, Yohanis. *Analisis Regresi Logistik Untuk Menentukan Faktor-faktor Yang Mempengaruhi Indeks Prestasi Mahasiswa Kumulatif (IPK) Mahasiswa FMIPA Universitas Sam Ratulangi Manado*. Manado: Universitas Sam Ratulangi.

Utami, Komang. 2015. *Kajian Masalah Antrian pada Sistem Pengumpulan Tol Konvensional terhadap Rancangan Sistem Pengumpulan Tol Elektronik*. Bali: Universitas Udayana.

Walpole, Ronald E. 1993. *Pengantar Statistika Edisi ke-3*. Jakarta: PT Gramedia Pustaka Utama.

Widarjono, Agus. 2010. *Analisis Statistika Multivariat Terapan*. Yogyakarta: UPP STIM YKPN.

Wiener, JL dan J, Tang. 2005. *Multicollinearity in Customer Satisfaction Research*. Ipsos Loyalty.

Yendra, Rado. 2008. *Teori Probabilitas*. Pekanbaru: Suska Press.

LAMPIRAN

LAMPIRAN

LAMPIRAN 1 Sintaks Pembangkitan Data

```
> # numbers of levels for predictors
> n1 <- 7; n2 <- 5; n3 <- 2; n4 <- 4; n5 <- 6; n6 <- 3
> nCells <- n1*n2*n3*n4*n5*n6
> nR <- 9          # repeated cells
> n <- nR*nCells   # sample size
> n
> a <- 1; b <- 4   # parameters in Beta distribution
> fD <- NULL       # initiate data table
> id <- 0          # initiate record IDs
> for (i1 in c(1:n1)) {
+  for (i2 in c(1:n2)) {
+   for (i3 in c(1:n3)) {
+    for (i4 in c(1:n4)) {
+     for (i5 in c(1:n5)) {
+      for (i6 in c(1:n6)) {
+       pMax <- rbeta(shape1=a,shape2=b,n=1)
+       ids <- id+c(1:nR)
+       x <- c(i1,i2,i3,i4,i5,i6)
+       dim(x) <- c(1,6)
+       xs <- rbind(x,x,x,x,x,x,x,x,x)
+       pFs <- rep(pMax,nR)*runif(nR)     # outcome
probabilities
+       Fs <- ifelse(pFs>0.5,1,0)       # final outcomes
+       fDs <- cbind(ids,xs,pFs,Fs)
+       fD <- rbind(fD,fDs)
+       id <- id+nR
+      }
+     }
+    }
+   }
+  }
+ }
> fD <- as.data.frame(fD)
> names(fD) <-
c("rID","x1","x2","x3","x4","x5","x6","pF","y")
```

LAMPIRAN 2 Sintaks Tabulasi Silang Data dengan Keterangan Kategori Variabel Terikat

```
> #plot summary events for each family
> windows(11.7,8)
> par(mfrow=c(2,3),las=1,oma=c(3,4,4,2),
  mar=c(0,0,0,0),mgp=c(1.1,0.1,0),tcl=0.2)
> xmax <- (n1+1)*n2
> ymax <- (n5+1)*n4
> xlm <- c(-5,xmax)
> ylm <- c(0,ymax+2)
> for (i3 in c(1:n3)) {
+  for (i6 in c(1:n6)) {
+   fd <- subset(fD,x3==i3 & x6==i6)
+   plot(1,type="n",xlim=xlm,ylim=ylm,ylab="",xlab="",
+   xaxt="n",yaxt="n",xaxs="i",yaxs="i")
+   if (i3==1) mtext(side=3,adj=0.5,line=0.3,paste("x6=",i6,sep=""))
+   if (i3==n3) mtext(side=1,adj=1,line=0.3,bl,cex=0.9)
+   if (i6==1) mtext(side=2,adj=0.5,line=1.5,paste("x3=",i3,sep=""))
+   if (i6==n6) axis(side=4,at=(1:((n5+1)*n4)+0.25),lab=ylab4,
+   tcl=0,hadj=-0.5,cex.axis=1.2)
+   if (i6==n6 & i3==1) axis(side=4,at=ymax+1,lab="x5",tcl=0,
+   hadj=-0.15,cex.axis=1.2)
+   for (i4 in c(1:n4)) {
+    for (i2 in c(1:n2)) {
+     if (i4==1) text((n4+4)*i2-4,ymax+0.5,paste("x2=",i2,sep=""),cex=1.2)
+     if (i2==1) text(-2,ymax-(n2+2)*i4+4,paste("x4=",i4,sep=""),cex=1.2)
+     fdd <- subset(fd,x4==i4 & x2==i2)
+     for (i5 in c(1:n5)) {
+      for (i1 in c(1:n1)) {
+       fddd <- subset(fdd,x5==i5 & x1==i1)
+       x0 <- mean(fddd$x1); y0 <- mean(fddd$x5); z <- mean(fddd$F)
+       clr <- ifelse(z>0,"brown",7)
+       x <- x0+(i2-1)*(n1+1); y <- ymax-(y0+(i4-1)*(n5+1))
+       points(x,y,cex=1.6,pch=21,bg=clr)
+      }
+     }
+    }
+   }
+  }
+ }
> lg <- c("Event(s)","No Events")
> legend("bottomright",inset=c(0.02,0.2),leg=lg,pch=21,pt.bg=c("brown",7),
+ y.intersp=0.7,cex=1.4,bg="ivory")
```

LAMPIRAN 3 Sintaks Sebaran *Event(s)* pada Tabulasi Silang Data

```
> #plot individual events for each family
> windows(11.7,8)
> par(mfrow=c(2,3),las=1,oma=c(3,5,4,2),
+ mar=c(0,0,0,0),mgp=c(1.1,0.1,0),tcl=0.2)
>
> for (i3 in c(1:n3)) {
+  for (i6 in c(1:n6)) {
+   fd <- subset(fD,x3==i3 & x6==i6)
+   plot(1,type="n",xlim=xlm,ylim=ylm,ylab="",xlab="",
+   xaxt="n",yaxt="n",xaxs="i",yaxs="i")
+   if (i3==1) mtext(side=3,adj=0.5,line=0.3,paste("x6=",i6,sep=""))
+   if (i3==n3) mtext(side=1,adj=1,line=0.3,bl,cex=0.9)
+   if (i6==1) mtext(side=2,adj=0.5,line=1.5,paste("x3=",i3,sep=""))
+   if (i6==n6) axis(side=4,at=(1:((n5+1)*n4)+0.25),lab=ylab4,
+   tcl=0,hadj=-0.5,cex.axis=1.2)
+   if (i6==n6 & i3==1) axis(side=4,at=ymax+1,lab="x5",tcl=0,
+   hadj=-0.15,cex.axis=1.2)
+   for (i4 in c(1:n4)) {
+    for (i2 in c(1:n2)) {
+     if (i4==1) text((n4+4)*i2-4,ymax+0.5,paste("x2=",i2,sep=""),cex=1.2)
+     if (i2==1) text(-2,ymax-(n2+2)*i4+4,paste("x4=",i4,sep=""),cex=1.2)
+     fdd <- subset(fd,x4==i4 & x2==i2)
+     x1 <- 999; x2 <- 0; y1 <- 999; y2 <- 0
+     for (i5 in c(1:n5)) {
+      for (i1 in c(1:n1)) {
+       fddd <- subset(fdd,x5==i5 & x1==i1)
+       x0 <- mean(fddd$x1); y0 <- mean(fddd$x5)
+       siz <- 2*sqrt(sum(fddd$F))
+       x <- x0+(i2-1)*(n1+1); y <- ymax-(y0+(i4-1)*(n5+1))
+       points(x,y,cex=siz,pch=20)
+       x1 <- min(x1,x); x2 <- max(x2,x)
+       y1 <- min(y1,y); y2 <- max(y2,y)
+      }
+     }
+     h <- 0.8
+     polygon(c(x1-h,x2+h,x2+h,x1-h,x1-h),c(y1-h,y1-h,y2+h,y2+h,y1-h),
+ border="grey60")
+    }
+   }
+  }
+ }
> lg <- c(1:9)
> legend("bottomright",inset=c(0.02,0.2),leg=lg,pch=20,pt.cex=2*sqrt(1:9),
+ title="# events",y.intersp=1,cex=1.4,bg="ivory",ncol=3)
```

LAMPIRAN 4 Sintaks Regresi Logistik Biner

```
> #logistic regression
> fD$x1 <- as.factor(fD$x1)
> fD$x2 <- as.factor(fD$x2)
> fD$x3 <- as.factor(fD$x3)
> fD$x4 <- as.factor(fD$x4)
> fD$x5 <- as.factor(fD$x5)
> fD$x6 <- as.factor(fD$x6)
> contrasts(fD$x1)
> contrasts(fD$x2)
> contrasts(fD$x3)
> contrasts(fD$x4)
> contrasts(fD$x5)
> contrasts(fD$x6)
> options(scipen=8)
> fD.lr <- glm(family=binomial,data=fD,y~x1+x2+x3+x4+x5+x6)
> summary(fD.lr)
> drop1(fD.lr,test="Chisq")
> library(aod)
> wald.test(b = coef(fD.lr), Sigma = vcov(fD.lr), Terms = 1:6)
> prob.fD.lr <- predict(fD.lr,type="response")
> co.lr <- 0.0095
> pred.fD.lr <- ifelse(prob.fD.lr>co.lr,1,0)
> cTab.lr <- table(pred.fD.lr,fD$y)
> addmargins(cTab.lr)
> pa.fD.lr <- 100*sum(diag(cTab.lr))/sum(cTab.lr)
> pa.fD.lr
```

LAMPIRAN 5 Sintaks *Random Forest*

```
> #random forest
> windows(6,6)
> library(randomForest)
> fD$y <- as.factor(fD$Y)
> fD.rf <- randomForest(data=fD,y~x1+x2+x3+x4+x5+x6,
+ importance=T)
> varImpPlot(fD.rf,pch=21,bg=4)
> mtext(line=0,adj=1,
+ paste("Sample size: ",nrow(fD),sep=""))
> pred.rf <- predict(fD.rf)
> prob.fD.rf <- predict(fD.rf,type="prob")[,2]
> co.rf <- 0.004
> pred.fD.rf <- ifelse(prob.fD.rf>co.rf,1,0)
> cTab.rf <- table(pred.fD.rf,fD$y)
> addmargins(cTab.rf)
> pa.fD.rf <- 100*sum(diag(cTab.rf))/sum(cTab.rf)
> pa.fD.rf
```

LAMPIRAN 6 Sintaks Plot Perbandingan Regresi Logistik Biner dengan *Random Forest*

```
> #plot results comparing methods for regression model and random forest
> windows(8,8)
> par(mfrow=c(2,1),oma=c(0,0,0,0),mar=c(3,3,3,4),las=1,mgp=c(1.1,0.1,0),tcl=0.2)
> ylm <- c(-0.1,1.1); xlm <- c(0,max(prob.fD.lr))
> xlb <- "Estimated Risk"
> plot(1,type="n",ylab="",xlab=xlb,ylim=ylm,xlim=xlm,yaxt="n")
> abline(v=co.lr,col=2,lwd=2)
> fraud <- prob.fD.lr[fD$y==1]
> legit <- prob.fD.lr[fD$y==0]
> q1 <- quantile(fraud,probs=c(0:4)/4)
> q0 <- quantile(legit,probs=c(0:4)/4)
> wd <- 0.08; clr <- "lemonchiffon"
> polygon(c(q1[2],q1[4],q1[4],q1[2],q1[2]),1+wd*c(-1,-1,1,1,-1),col=clr)
> points(c(q1[3],q1[3]),1+wd*c(-1,1),type="l")
> polygon(c(q0[2],q0[4],q0[4],q0[2],q0[2]),wd*c(-1,-1,1,1,-1),col=clr)
> points(c(q0[3],q0[3]),wd*c(-1,1),type="l")
> points(prob.fD.lr,fD$y)
> axis(side=2,at=c(0.2,0.8),lab=c("No","Yes"))
> axis(side=4,at=1,lab="Total",font=3,hadj=-0.25,tcl=0)
> mtext(side=3,adj=0.5,line=1.4,"Logistic Regression Model",font=2)
> tSum <- table(fD$y)
> axis(side=4,at=c(0.2,0.8),lab=tSum,hadj=-0.2)
> legend("bottomleft",inset=c(-0.02,0.2),bty="n",leg=cTab.lr[1,1])
> lg1 <- paste(cTab.lr[2,1],"+")
> legend("bottomright",inset=c(0.02,0.2),bty="n",leg=lg1,text.col=2)
> lg2 <- paste(cTab.lr[1,2],"-")
> legend("topleft",inset=c(-0.02,0.2),bty="n",leg=lg2,text.col=2)
> legend("topright",inset=c(0.02,0.2),bty="n",leg=cTab.lr[2,2])
> lg <- paste("Accuracy:",round(pa.fD.lr,2),"%")
> legend("right",inset=0.02,leg=lg,x.intersp=0.2,bg="ivory")
>
> ylm <- c(-0.1,1.1); xlm <- c(0,max(prob.fD.rf))
> xlb <- "Estimated Risk"
> plot(1,type="n",ylab="",xlab=xlb,ylim=ylm,xlim=xlm,yaxt="n")
> abline(v=co.rf,col=2,lwd=2)
> fraud <- prob.fD.rf[fD$y==1]
> legit <- prob.fD.rf[fD$y==0]
> q1 <- quantile(fraud,probs=c(0:4)/4)
> q0 <- quantile(legit,probs=c(0:4)/4)
> wd <- 0.08; clr <- "lemonchiffon"
> polygon(c(q1[2],q1[4],q1[4],q1[2],q1[2]),1+wd*c(-1,-1,1,1,-1),col=clr)
> points(c(q1[3],q1[3]),1+wd*c(-1,1),type="l")
> polygon(c(q0[2],q0[4],q0[4],q0[2],q0[2]),wd*c(-1,-1,1,1,-1),col=clr)
> points(c(q0[3],q0[3]),wd*c(-1,1),type="l")
> points(prob.fD.rf,fD$y)
> axis(side=2,at=c(0.2,0.8),lab=c("No","Yes"))
> axis(side=4,at=1,lab="Total",font=3,hadj=-0.25,tcl=0)
> mtext(side=3,adj=0.5,line=1.4,"Random Forest",font=2)
> tSum <- table(fD$y)
> axis(side=4,at=c(0.2,0.8),lab=tSum,hadj=-0.2)
> legend("bottomleft",inset=c(-0.02,0.2),bty="n",leg=cTab.rf[1,1])
> lg1 <- paste(cTab.rf[2,1],"+")
> legend("bottomright",inset=c(0.02,0.2),bty="n",leg=lg1,text.col=2)
> lg2 <- paste(cTab.rf[1,2],"-")
> legend("topleft",inset=c(-0.02,0.2),bty="n",leg=lg2,text.col=2)
> legend("topright",inset=c(0.02,0.2),bty="n",leg=cTab.rf[2,2])
> lg <- paste("Accuracy:",round(pa.fD.rf,2),"%")
> legend("right",inset=0.02,leg=lg,x.intersp=0.2,bg="ivory")
> #-----------------------------------------
```

LAMPIRAN 7 Sintaks Analisis Ulang untuk *Training* dan *Test* Data pada Regresi Logistik Biner dan *Random Forest*

```
> #split sample into training set and test set
>
> set.seed(1357)
> N <- nrow(fD)
> N1 <- floor(N/10)      # training data sample size
> N2 <- N-N1             # test data sample size
> trIDs <- sample(c(1:N),
+ replace=F,size=N1)     # training record IDs
> teIDs <- c(1:N)[-trIDs]
> fTr <- fD[trIDs,]      # training sample
> fTe <- fD[teIDs,]      # test sample
> fTr.lr <- glm(family=binomial,data=fTr,y~x1+x2+x3+x4+x5+x6)
> prob.fTr.lr <- predict(fTr.lr,type="response")
> pred.fTr.lr <- ifelse(prob.fTr.lr>co.lr,1,0)
> cTab.fTr.lr <- table(pred.fTr.lr,fTr$y)
> addmargins(cTab.fTr.lr)
> pa.fTr.lr <- 100*sum(diag(cTab.fTr.lr))/sum(cTab.fTr.lr)
> prob.fTe.lr <- predict(fTr.lr,fTe,type="response")
> pred.fTe.lr <- ifelse(prob.fTe.lr>co.lr,1,0)
> cTab.fTe.lr <- table(pred.fTe.lr,fTe$y)
> pa.fTe.lr <- 100*sum(diag(cTab.fTe.lr))/sum(cTab.fTe.lr)
>
>
> fTr$y <- as.factor(fTr$y)
> fTe$y <- as.factor(fTe$y)# factor outcome for rf
>
> fTr.rf <- randomForest(data=fTr,y~x1+x2+x3+x4+x5+x6,
+ importance=T)
> pred.rf <- predict(fTr.rf)
> prob.fTr.rf <- predict(fTr.rf,type="prob")[,2]
> pred.fTr.rf <- ifelse(prob.fTr.rf>co.rf,1,0)
> cTab.fTr.rf <- table(pred.fTr.rf,fTr$y)
> addmargins(cTab.fTr.rf)
> pa.fTr.rf <- 100*sum(diag(cTab.fTr.rf))/sum(cTab.fTr.rf)
> prob.fTe.rf <- predict(fTr.rf,fTe,type="prob")[,2]
> pred.fTe.rf <- ifelse(prob.fTe.rf>co.rf,1,0)
> cTab.fTe.rf <- table(pred.fTe.rf,fTe$y)
> addmargins(cTab.fTe.rf)
> pa.fTe.rf <- 100*sum(diag(cTab.fTe.rf))/sum(cTab.fTe.rf)
> pa.fTe.rf
```

LAMPIRAN 8 Sintaks Plot Perbandingan Regresi Logistik Biner dan *Random Forest* pada *Training* dan *Test* Data

```
> #compare methods for training and test sets
>
> fTr$y <- as.character(fTr$y)
> fTr$y <- as.integer(fTr$y)# convert back
> fTe$y <- as.character(fTe$y)# to integer outcome
> fTe$y <- as.integer(fTe$y)
> windows(10,8)
> par(mfrow=c(2,2),oma=c(1,1,1,0),mar=c(2,2,3,4),
+ las=1,mgp=c(1.1,0.1,0),tcl=0.2)
> xlb <- "Estimated Risk"
> wd <- 0.08
> clr <- "lemonchiffon"
> ylm <- c(-0.1,1.1)
> xlm.lr <- c(0,max(prob.fTr.lr,prob.fTe.lr))
> xlm.rf <- c(0,max(prob.fTr.rf,prob.fTe.rf))
>
> plot(1,type="n",ylab="",xlab="",ylim=ylm,xlim=xlm.lr,yaxt="n",cex.axis=1.2)
> abline(v=co.lr,col=2,lwd=2)
> fraud <- prob.fTr.lr[fTr$y==1]
> legit <- prob.fTr.lr[fTr$y==0]
> q1 <- quantile(fraud,probs=c(0:4)/4)
> q0 <- quantile(legit,probs=c(0:4)/4)
> clr <- "lemonchiffon"
> polygon(c(q1[2],q1[4],q1[4],q1[2],q1[2]),1+wd*c(-1,-1,1,1,-1),col=clr)
> points(c(q1[3],q1[3]),1+wd*c(-1,1),type="l")
> polygon(c(q0[2],q0[4],q0[4],q0[2],q0[2]),wd*c(-1,-1,1,1,-1),col=clr)
> points(c(q0[3],q0[3]),wd*c(-1,1),type="l")
> points(prob.fTr.lr,fTr$y)
> axis(side=2,at=c(0.2,0.8),lab=c("0","1"),cex.axis=1.2)
> mtext(side=3,adj=1,line=0.2,paste("Training Sample:",N1))
> axis(side=4,at=1,lab="Total",font=3,hadj=-0.25,tcl=0,cex.axis=1.2)
> mtext(side=3,adj=0.5,line=1.8,"Logistic Regression Model",font=2)
> tSum <- table(fTr$y)
> axis(side=4,at=c(0.2,0.8),lab=tSum,hadj=-0.2,cex.axis=1.2)
> legend("bottomleft",inset=c(-0.02,0.2),bty="n",leg=cTab.fTr.lr[1,1],cex=1.2)
> lg1 <- paste(cTab.fTr.lr[2,1],"+")
> legend("bottomright",inset=c(0.02,0.2),bty="n",leg=lg1,text.col=2,cex=1.2)
> lg2 <- paste(cTab.fTr.lr[1,2],"-")
> legend("topleft",inset=c(-0.02,0.2),bty="n",leg=lg2,text.col=2,cex=1.2)
> legend("topright",inset=c(0.02,0.2),bty="n",leg=cTab.fTr.lr[2,2],cex=1.2)
> lg <- paste("Accuracy:",round(pa.fTr.lr,2),"%")
> legend("right",inset=0.02,leg=lg,x.intersp=0.2,bg="ivory",cex=1.2)
>
> plot(1,type="n",ylab="",xlab="",ylim=ylm,xlim=xlm.rf,yaxt="n",cex.axis=1.2)
> abline(v=co.rf,col=2,lwd=2)
> fraud <- prob.fTr.rf[fTr$y==1]
> legit <- prob.fTr.rf[fTr$y==0]
> q1 <- quantile(fraud,probs=c(0:4)/4)
> q0 <- quantile(legit,probs=c(0:4)/4)
> polygon(c(q1[2],q1[4],q1[4],q1[2],q1[2]),1+wd*c(-1,-1,1,1,-1),col=clr)
> points(c(q1[3],q1[3]),1+wd*c(-1,1),type="l")
> polygon(c(q0[2],q0[4],q0[4],q0[2],q0[2]),wd*c(-1,-1,1,1,-1),col=clr)
> points(c(q0[3],q0[3]),wd*c(-1,1),type="l")
> points(prob.fTr.rf,fTr$y)
> axis(side=2,at=c(0.2,0.8),lab=c("0","1"),cex.axis=1.2)
> mtext(side=3,adj=1,line=0.2,paste("Training Sample:",N1))
> axis(side=4,at=1,lab="Total",font=3,hadj=-0.25,tcl=0,cex.axis=1.2)
> mtext(side=3,adj=0.5,line=1.8,"Random Forest",font=2)
> tSum <- table(fTr$y)
> axis(side=4,at=c(0.2,0.8),lab=tSum,hadj=-0.2,cex.axis=1.2)
> legend("bottomleft",inset=c(-0.02,0.2),bty="n",leg=cTab.fTr.rf[1,1],cex=1.2)
```

```
> lg1 <- paste(cTab.fTr.rf[2,1],"+")
> legend("bottomright",inset=c(0.02,0.2),bty="n",leg=lg1,text.col=2,cex=1.2)
> lg2 <- paste(cTab.fTr.rf[1,2],"-")
> legend("topleft",inset=c(-0.02,0.2),bty="n",leg=lg2,text.col=2,cex=1.2)
> legend("topright",inset=c(0.02,0.2),bty="n",leg=cTab.fTr.rf[2,2],cex=1.2)
> lg <- paste("Accuracy:",round(pa.fTr.rf,2),"%")
> legend("right",inset=0.02,leg=lg,x.intersp=0.2,bg="ivory",cex=1.2)
>
> plot(1,type="n",ylab="",xlab=xlb,ylim=ylm,xlim=xlm.lr,
+ yaxt="n",cex.axis=1.2,cex.lab=1.2)
> abline(v=co.lr,col=2,lwd=2)
> fraud <- prob.fTe.lr[fTe$y==1]
> legit <- prob.fTe.lr[fTe$y==0]
> q1 <- quantile(fraud,probs=c(0:4)/4)
> q0 <- quantile(legit,probs=c(0:4)/4)
> polygon(c(q1[2],q1[4],q1[4],q1[2],q1[2]),1+wd*c(-1,-1,1,1,-1),col=clr)
> points(c(q1[3],q1[3]),1+wd*c(-1,1),type="l")
> polygon(c(q0[2],q0[4],q0[4],q0[2],q0[2]),wd*c(-1,-1,1,1,-1),col=clr)
> points(c(q0[3],q0[3]),wd*c(-1,1),type="l")
> points(prob.fTe.lr,fTe$y)
> axis(side=2,at=c(0.2,0.8),lab=c("0","1"),cex.axis=1.2)
> mtext(side=3,adj=1,line=0.2,paste("Test Sample:",N2))
> axis(side=4,at=1,lab="Total",font=3,hadj=-0.25,tcl=0,cex.axis=1.2)
> tSum <- table(fTe$y)
> axis(side=4,at=c(0.2,0.8),lab=tSum,hadj=-0.2,cex.axis=1.2)
> legend("bottomleft",inset=c(-0.02,0.2),bty="n",leg=cTab.fTe.lr[1,1],cex=1.2)
> lg1 <- paste(cTab.fTe.lr[2,1],"+")
> legend("bottomright",inset=c(0.02,0.2),bty="n",leg=lg1,text.col=2,cex=1.2)
> lg2 <- paste(cTab.fTe.lr[1,2],"-")
> legend("topleft",inset=c(-0.02,0.2),bty="n",leg=lg2,text.col=2,cex=1.2)
> legend("topright",inset=c(0.02,0.2),bty="n",leg=cTab.fTe.lr[2,2],cex=1.2)
> lg <- paste("Accuracy:",round(pa.fTe.lr,2),"%")
> legend("right",inset=0.02,leg=lg,x.intersp=0.2,bg="ivory",cex=1.2)
>
> plot(1,type="n",ylab="",xlab=xlb,ylim=ylm,xlim=xlm.rf,yaxt="n",
+ cex.lab=1.2,cex.axis=1.2)
> abline(v=co.rf,col=2,lwd=2)
> fraud <- prob.fTe.rf[fTe$y==1]
> legit <- prob.fTe.rf[fTe$y==0]
> q1 <- quantile(fraud,probs=c(0:4)/4)
> q0 <- quantile(legit,probs=c(0:4)/4)
> polygon(c(q1[2],q1[4],q1[4],q1[2],q1[2]),1+wd*c(-1,-1,1,1,-1),col=clr)
> points(c(q1[3],q1[3]),1+wd*c(-1,1),type="l")
> polygon(c(q0[2],q0[4],q0[4],q0[2],q0[2]),wd*c(-1,-1,1,1,-1),col=clr)
> points(c(q0[3],q0[3]),wd*c(-1,1),type="l")
> points(prob.fTe.rf,fTe$y)
> axis(side=2,at=c(0.2,0.8),lab=c("0","1"),cex.axis=1.2)
> mtext(side=3,adj=1,line=0.2,paste("Test Sample:",N2))
> axis(side=4,at=1,lab="Total",font=3,hadj=-0.25,tcl=0,cex.axis=1.2)
> tSum <- table(fTe$y)
> axis(side=4,at=c(0.2,0.8),lab=tSum,hadj=-0.2,cex.axis=1.2)
> legend("bottomleft",inset=c(-0.02,0.2),bty="n",leg=cTab.fTe.rf[1,1],cex=1.2)
> lg1 <- paste(cTab.fTe.rf[2,1],"+")
> legend("bottomright",inset=c(0.02,0.2),bty="n",leg=lg1,text.col=2,cex=1.2)
> lg2 <- paste(cTab.fTe.rf[1,2],"-")
> legend("topleft",inset=c(-0.02,0.2),bty="n",leg=lg2,text.col=2,cex=1.2)
> legend("topright",inset=c(0.02,0.2),bty="n",leg=cTab.fTe.rf[2,2],cex=1.2)
> lg <- paste("Accuracy:",round(pa.fTe.rf,2),"%")
> legend("right",inset=0.02,leg=lg,x.intersp=0.2,bg="ivory",cex=1.2)
```