

PERBANDINGAN METODE REGRESI LOGISTIK DAN *RANDOM FOREST* DALAM MEMPREDIKSI *CHURN* PADA PEMEGANG POLIS ASURANSI MOBIL

HANA ALYA



PROGRAM STUDI AKTUARIA
SEKOLAH SAINS DATA, MATEMATIKA, DAN INFORMATIKA
INSTITUT PERTANIAN BOGOR
BOGOR
2025

- Hak Cipta Dilindungi Undang-undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
 2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.



PERNYATAAN MENGENAI SKRIPSI DAN SUMBER INFORMASI SERTA PELIMPAHAN HAK CIPTA

Dengan ini saya menyatakan bahwa skripsi dengan judul “Perbandingan Metode Regresi Logistik dan *Random Forest* dalam Memprediksi *Churn* pada Pemegang Polis Asuransi Mobil” adalah karya saya dengan arahan dari dosen pembimbing dan belum diajukan dalam bentuk apa pun kepada perguruan tinggi mana pun. Sumber informasi yang berasal atau dikutip dari karya yang diterbitkan maupun tidak diterbitkan dari penulis lain telah disebutkan dalam teks dan dicantumkan dalam Daftar Pustaka di bagian akhir skripsi ini.

Dengan ini saya melimpahkan hak cipta dari karya tulis saya kepada Institut Pertanian Bogor.

Bogor, Juni 2025

Hana Alya
G5402211069

- Hak Cipta Dilindungi Undang-undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
 2. Dilarang menggunakan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.



ABSTRAK

HANA ALYA. Perbandingan Metode Regresi Logistik dan *Random Forest* dalam Memprediksi *Churn* pada Pemegang Polis Asuransi Mobil. Dibimbing oleh BUDI SUHARJO dan RUHIYAT.

Penelitian ini membandingkan regresi logistik dan *random forest* dalam memprediksi *churn* pada pemegang polis asuransi mobil dengan teknik *undersampling* proporsi 30%, 50%, dan 70%. Model *random forest* setelah *undersampling* 30% menunjukkan akurasi (99,07%) dan AUC tertinggi (99,91%), sementara sensitivitas tertinggi (98,85%) pada model *random forest* sebelum *undersampling*. F1-score tertinggi (99,29%) diperoleh dari model *random forest* setelah *undersampling* 70%. Variabel paling penting yang memengaruhi keputusan pemegang polis untuk melakukan *churn* dalam model regresi logistik berdasarkan nilai *odds ratio* adalah status pekerjaan (pensiun), jumlah pembayaran bulanan pemegang polis asuransi (kategori 500-600), dan tingkat pendidikan tertinggi pemegang polis asuransi (doktor). Pada model *random forest*, variabel paling penting berdasarkan nilai *mean decrease gini* adalah tipe penawaran perpanjangan polis, status pekerjaan, dan tingkat pendidikan terakhir pemegang polis asuransi. Berdasarkan analisis nilai SHAP, variabel seperti status pekerjaan (pensiun), tipe penawaran perpanjangan polis (tipe 4), dan jumlah pembayaran per bulan (500-600) secara konsisten meningkatkan peluang *churn*.

Kata Kunci: data *churn*, metode *supervised learning*, *random forest*, regresi logistik, *undersampling*

ABSTRACT

HANA ALYA. Comparison of Logistic Regression and Random Forest Methods in Predicting *Churn* Among Auto Insurance Policyholders. Supervised by BUDI SUHARJO and RUHIYAT.

This study compares logistic regression and random forest models in predicting *churn* among car insurance policyholders using undersampling techniques at proportions of 30%, 50%, and 70%. The random forest model with 30% undersampling achieved the highest accuracy (99.07%) and AUC (99.91%), while the highest sensitivity (98.85%) was observed in the random forest model before undersampling. The highest F1-score (99.29%) was obtained from the random forest model with 70% undersampling. The most influential variables affecting policyholders *churn* decisions in the logistic regression model, based on odds ratio values, were employment status (retired), monthly insurance payment amount (category 500–600), and highest level of education (doctoral degree). In the random forest model, the most important variables based on mean decrease in Gini were type of policy renewal offer, employment status, and highest level of education. Based on the SHAP value analysis, variables such as employment status (retired), policy renewal offer type (type 4), and monthly payment amount (500–600) consistently increased the likelihood of *churn*.

Keywords: *churn* data, logistic regression, random forest, supervised learning method, undersampling



@Hak cipta milik IPB University

Hak Cipta Dilindungi Undang-undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.

© Hak Cipta milik IPB, tahun 2025¹
Hak Cipta dilindungi Undang-Undang

Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan atau menyebutkan sumbernya. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik, atau tinjauan suatu masalah, dan pengutipan tersebut tidak merugikan kepentingan IPB.

Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apa pun tanpa izin IPB.



- Hak Cipta Dilindungi Undang-undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
 2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.

PERBANDINGAN METODE REGRESI LOGISTIK DAN RANDOM FOREST DALAM MEMPREDIKSI *CHURN* PADA PEMEGANG POLIS ASURANSI MOBIL

HANA ALYA

**PROGRAM STUDI AKTUARIA
SEKOLAH SAINS DATA, MATEMATIKA, DAN INFORMATIKA
INSTITUT PERTANIAN BOGOR
BOGOR
2025**



@Hak cipta milik IPB University

Hak Cipta Dilindungi Undang-undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.

Tim Penguji pada Ujian Skripsi: Ir. N. K. Kutha Ardana, M.Sc.



Judul Skripsi : Perbandingan Metode Regresi Logistik dan *Random Forest* dalam Memprediksi *Churn* pada Pemegang Polis Asuransi Mobil

Nama : Hana Alya
NIM : G5402211069

@Hak cipta milik IPB University

Disetujui oleh

Pembimbing 1:

Dr. Ir. Budi Suharjo, M.S.

Pembimbing 2:

Ruhyat, S.Si., M.Si., M.Act.Sc.

Diketahui oleh

Ketua Program Studi:

Dr. Ir. I Gusti Putu Purnaba, DEA.
NIP. 196512181990021001

Tanggal Ujian:
27 Mei 2025

Tanggal Lulus:



PRAKATA

Puji dan syukur penulis panjatkan kepada Allah *Subhanaahu Wa Ta'ala* atas segala karunia-Nya sehingga karya ilmiah ini berhasil diselesaikan. Judul yang dipilih dalam penelitian yang dilaksanakan sejak bulan Agustus 2024 dengan judul “Perbandingan Metode Regresi Logistik dan *Random Forest* dalam Memprediksi *Churn* pada Pemegang Polis Asuransi Mobil”.

Terima kasih penulis ucapkan kepada para pembimbing, Dr. Ir. Budi Suharjo, M.S. dan Ruhiyat, S.Si., M.Si., M.Act.Sc. yang telah membimbing dan banyak memberi saran. Ucapan terima kasih juga disampaikan kepada pembimbing akademik, moderator seminar, dan penguji luar komisi pembimbing. Ungkapan terima kasih juga disampaikan kepada ayah, ibu, adik, serta seluruh keluarga yang telah memberikan dukungan, doa, dan kasih sayangnya yang selalu diberikan.

Semoga karya ilmiah ini bermanfaat bagi pihak yang membutuhkan dan bagi kemajuan ilmu pengetahuan.

Bogor, Juni 2025

Hana Alya



DAFTAR ISI

DAFTAR TABEL	ix
DAFTAR GAMBAR	ix
DAFTAR LAMPIRAN	x
PENDAHULUAN	1
1.1 Latar Belakang	1
1.2 Tujuan	2
TINJAUAN PUSTAKA	3
2.1 <i>Churn</i>	3
2.2 Pra-proses Data	4
2.3 Data Tidak Seimbang	6
2.4 Metode <i>Supervised Learning</i>	6
2.5 Evaluasi dan Validasi Model	8
2.6 Variabel Penting dalam Metode <i>Supervised Learning</i>	10
III METODE PENELITIAN	11
3.1 Sumber Data	11
3.2 Prosedur Analisis	12
3.3 Perangkat Lunak	13
IV HASIL DAN PEMBAHASAN	14
4.1 Gambaran Umum Data	14
4.2 Pembagian Data	16
4.3 Penanganan Data Tidak Seimbang	17
4.4 Metode <i>Supervised Learning</i> dengan Regresi Logistik	18
4.5 Metode <i>Supervised Learning</i> dengan <i>Random Forest</i>	26
4.6 Perbandingan Metode Regresi Logistik dengan <i>Random forest</i>	31
4.7 Faktor-Faktor yang Berpengaruh dalam Regresi Logistik dan <i>Random forest</i>	33
V SIMPULAN DAN SARAN	38
5.1 Simpulan	38
5.2 Saran	38
DAFTAR PUSTAKA	39
LAMPIRAN	41



Hak Cipta Dilindungi Undang-undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.

DAFTAR TABEL

1	<i>Confusion matrix</i>	9
2	Tabulasi kategori setiap variabel	11
3	Presentase pemegang polis asuransi melakukan <i>churn</i> untuk setiap variabel prediktor	15
4	Pembagian data berdasarkan variabel <i>response</i>	17
5	Pembagian data berdasarkan variabel <i>response</i>	17
6	Koefisien estimasi, <i>odds ratio</i> , dan <i>p-value</i> variabel prediktor sebelum <i>random undersampling</i>	18
7	Koefisien estimasi, <i>odds ratio</i> , dan <i>p-value</i> variabel prediktor setelah 30% <i>random undersampling</i>	20
8	Koefisien estimasi, <i>odds ratio</i> , dan <i>p-value</i> variabel prediktor setelah 50% <i>random undersampling</i>	22
9	Koefisien estimasi, <i>odds ratio</i> , dan <i>p-value</i> variabel prediktor setelah 70% <i>random undersampling</i>	24
10	Kombinasi parameter optimal dan nilai AUC model <i>random forest</i>	31
11	Indikator ketepatan metode regresi logistik	32
12	Indikator ketepatan metode <i>random forest</i>	32
13	Nilai <i>odds ratio</i> masing-masing variabel sebelum dan setelah <i>random undersampling</i>	33
14	Nilai SHAP pada prediksi <i>churn</i>	36

DAFTAR GAMBAR

1	Ilustrasi algoritma <i>random forest</i> (Trivusi 2022)	8
2	Presentase <i>churn</i>	14
3	Presentase <i>churn</i> berdasarkan status pekerjaan	14
4	Kombinasi <i>mtry</i> dan <i>ntree</i> sebelum diterapkan <i>random undersampling</i> berdasarkan perubahan <i>mtry</i> (a) dan berdasarkan perubahan <i>ntree</i> (b)	27
5	Kombinasi <i>mtry</i> dan <i>ntree</i> setelah 30% <i>random undersampling</i> berdasarkan perubahan <i>mtry</i> (a) dan berdasarkan perubahan <i>ntree</i> (b)	28
6	Kombinasi <i>mtry</i> dan <i>ntree</i> setelah 50% <i>random undersampling</i> berdasarkan perubahan <i>mtry</i> (a) dan berdasarkan perubahan <i>ntree</i> (b)	29
7	Kombinasi <i>mtry</i> dan <i>ntree</i> setelah <i>random undersampling</i> 70% berdasarkan perubahan <i>mtry</i> (a) dan berdasarkan perubahan <i>ntree</i> (b)	30
8	Nilai <i>mean decrease gini</i> metode <i>random forest</i>	35



DAFTAR LAMPIRAN

1	Visualisasi variabel kategorik	42
2	Nilai AUC kombinasi t dan m sebelum <i>random undersampling</i> (diurutkan dari nilai AUC tertinggi)	44
3	Nilai AUC kombinasi t dan m dengan 30% <i>random undersampling</i> (diurutkan dari nilai AUC tertinggi)	45
4	Nilai AUC kombinasi t dan m dengan 50% <i>random undersampling</i> (diurutkan dari nilai AUC tertinggi)	46
5	Nilai AUC kombinasi t dan m dengan 70% <i>random undersampling</i> (diurutkan dari nilai AUC tertinggi)	47
6	Nilai <i>Mean Decrease Gini</i> (diurutkan dari nilai MDG tertinggi)	48
7	Kode yang digunakan pada RStudio	50

@Hak cipta milik IPB University

Hak Cipta Dilindungi Undang-undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.