

PERBANDINGAN KLASIFIKASI RISIKO BIAYA ASURANSI KESEHATAN MENGGUNAKAN *K-NEAREST NEIGHBOR* DAN REGRESI LOGISTIK MULTINOMIAL

FARAH FARHANI RAMADHANI



**PROGRAM STUDI AKTUARIA
SEKOLAH SAINS DATA, MATEMATIKA, DAN INFORMATIKA
INSTITUT PERTANIAN BOGOR
BOGOR
2026**



PERNYATAAN MENGENAI SKRIPSI DAN SUMBER INFORMASI SERTA PELIMPAHAN HAK CIPTA

Dengan ini saya menyatakan bahwa skripsi dengan judul “Perbandingan Klasifikasi Risiko Biaya Asuransi Kesehatan Menggunakan *K-Nearest Neighbor* dan Regresi Logistik Multinomial” 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, Juli 2026

Farah Farhani Ramadhani
G5402201008

ABSTRAK

FARAH FARHANI RAMADHANI. Perbandingan Klasifikasi Risiko Biaya Asuransi Kesehatan Menggunakan *K-Nearest Neighbor* dan Regresi Logistik Multinomial. Dibimbing oleh RUHIYAT dan ELIS KHATIZAH.

Penetapan premi asuransi kesehatan membutuhkan estimasi risiko biaya medis yang presisi. Penelitian ini bertujuan untuk membandingkan kinerja metode regresi logistik multinomial dan *K-nearest neighbor* (KNN) dalam mengklasifikasikan tingkat risiko biaya medis nasabah. Variabel prediktor yang dianalisis meliputi usia, indeks massa tubuh, status perokok, dan banyaknya anak. Variabel respons berupa biaya medis kontinu diubah menjadi tiga kategori risiko, yaitu rendah, sedang, dan tinggi, menggunakan persentil ke-33 dan ke-66. Evaluasi model dilakukan melalui teknik *5-fold cross validation*. Hasil penelitian menunjukkan bahwa pendekatan non-parametrik KNN dengan parameter $K = 7$ menghasilkan akurasi sebesar 89.13%, presisi 90.41%, *recall* 89.13%, dan *f1-score* 88.9%. Hasil-hasil ini lebih tinggi dibandingkan model regresi logistik multinomial yang hanya mencapai akurasi 81.27%. Dengan demikian, metode KNN relatif lebih handal dan stabil dalam mengklasifikasikan data risiko biaya medis yang memiliki variabilitas ekstrem.

Kata kunci: biaya medis, *k-fold cross validation*, klasifikasi, *K-nearest neighbor*, regresi logistik multinomial.

ABSTRACT

FARAH FARHANI RAMADHANI. Comparison of Health Insurance Cost Risk Classification Using *K-Nearest Neighbor* and Multinomial Logistic Regression. Supervised by RUHIYAT and ELIS KHATIZAH.

Accurate estimation of medical cost risk is essential in determining health insurance premium pricing. This study aims to compare the performance of the multinomial logistic regression and the *K-nearest neighbor* (KNN) methods in classifying policyholders' medical cost risk levels. The predictor variables considered in this study include age, body mass index, smoking status, and number of children. The continuous medical cost variable was transformed into three risk categories, low, medium, and high, based on the 33rd and 66th percentiles. Model evaluation was conducted using a *5-fold cross-validation* technique. The results demonstrate that the non-parametric KNN approach with $K = 7$ achieved an accuracy of 89.13%, precision of 90.41%, recall of 89.13%, and *f1-score* of 88.9%. These results are outperforming the multinomial logistic regression model which only achieved an accuracy of 81.27%. These findings indicate that the KNN method is relatively more reliable in classifying medical cost risk data characterized by high variability and extreme observations.

Keywords: classification, *k-fold cross validation*, *K-nearest neighbor*, medical cost, multinomial logistic regression



@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.

PERBANDINGAN KLASIFIKASI RISIKO BIAYA ASURANSI KESEHATAN MENGGUNAKAN *K-NEAREST NEIGHBOR* DAN REGRESI LOGISTIK MULTINOMIAL

FARAH FARHANI RAMADHANI

Skripsi
sebagai salah satu syarat untuk memperoleh gelar
Sarjana pada
Program Studi Aktuaria

**PROGRAM STUDI AKTUARIA
SEKOLAH SAINS DATA, MATEMATIKA, DAN INFORMATIKA
INSTITUT PERTANIAN BOGOR
BOGOR
2026**



@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:
1 Dr. Ir. Retno Budiarti, M.S.



Judul Skripsi : Perbandingan Klasifikasi Risiko Biaya Asuransi Kesehatan Menggunakan *K-Nearest Neighbor* dan Regresi Logistik Multinomial

Nama : Farah Farhani Ramadhani

NIM : G5402201008

Disetujui oleh

Pembimbing 1:

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

Pembimbing 2:

Elis Khatizah, S.Si., M.Si., Ph.D.

Diketahui oleh

Ketua Program Studi:

Dr. Ir. I Gusti Putu Purnaba, DEA.

NIP. 196512181990021001

Tanggal Ujian: 19 Mei 2026

Tanggal Lulus:



PRAKATA

Puji dan syukur penulis panjatkan kepada Allah *Subhanaahu Wa Ta'ala* atas segala karunia-Nya sehingga karya ilmiah ini berhasil diselesaikan. Tema yang dipilih dalam penelitian yang dilaksanakan sejak bulan Februari 2026 sampai bulan Juli 2026 ini ialah klasifikasi, regresi logistik multinomial, dan *K-nearest neighbor* dengan judul “Perbandingan Klasifikasi Risiko Biaya Asuransi Kesehatan Menggunakan *K-Nearest Neighbor* dan Regresi Logistik Multinomial”.

Penyusunan dan penulisan karya ilmiah ini tidak lepas dari berbagai pihak. Oleh karena itu, penulis ingin menyampaikan ucapan terima kasih kepada:

1. Ibu Intan Kusriani, Ayah Fakhru Mubin, Adik Fakhri Naufal Ramadhan, Alm. Mbah Yatimah, serta seluruh keluarga besar yang telah memberikan dukungan moral, material, doa tiada henti, kasih sayang, dan kepercayaan tinggi untuk penulis selama hidup dan penulisan karya ilmiah ini.
2. Bapak Ruhayat, S.Si., M.Si., M.Act.Sc. selaku dosen pembimbing I dan Ibu Elis Khatizah, S.Si., M.Si., Ph.D. selaku dosen pembimbing II yang telah meluangkan waktu, pikiran, dan tenaga untuk memberikan bimbingan, arahan, serta masukan yang sangat berarti selama proses penulisan karya ilmiah.
3. Ibu Dr. Ir. Retno Budiarti, M.S. selaku penguji luar yang telah meluangkan waktu, pikiran, dan tenaga untuk memberikan saran dan arahan yang berharga sehingga karya ilmiah ini terselesaikan.
4. Seluruh dosen dan tenaga kependidikan Program Studi Aktuaria maupun dosen program studi lain yang telah membekali penulis dengan ilmu pengetahuan dan keterampilan selama masa studi.
5. Zahra Fitri Ardani selaku sahabat dekat penulis yang senantiasa menemani dalam suka dan duka, mendengarkan tanpa menghakimi, menjadi rekan yang selalu ada bagi penulis sejak masa SMP hingga saat ini.
6. Nazwa, Aliyya, Intan, Setiyani, Hilda, Syifa, Marthin, Wawan, Hanif, dan Krishna selaku teman-teman seperjuangan yang selalu memberikan bantuan, dukungan dan banyak petualangan selama masa studi di IPB *University*.
7. Gayatri, Tiana, Farida, Cholifah, Layla, Rosa, Eklecia selaku teman SMA yang konsisten menemani penulis sejak masa remaja hingga saat ini.
8. Orang-orang yang telah hadir di hidup penulis, teman-teman selama masa studi, rekan kerja selama magang, dan teman daring atas kebersamaan dan semangat perjuangan.

Sebagai penutup, penulis juga ingin menyampaikan apresiasi untuk diri sendiri. Penghormatan yang terdalem karena selalu memilih dan menghargai diri sendiri. Terima kasih atas daya juang, dedikasi, dan komitmen selama penyusunan tugas akhir. Terima kasih telah bertahan dan tidak menyerah melewati segala kesulitan sehingga naskah ini berhasil diselesaikan dengan baik. Semoga karya ilmiah ini bermanfaat bagi pihak yang membutuhkan dan bagi kemajuan ilmu pengetahuan.

Bogor, Juli 2026

Farah Farhani Ramadhani



DAFTAR ISI

DAFTAR TABEL	viii
DAFTAR GAMBAR	viii
DAFTAR LAMPIRAN	viii
I PENDAHULUAN	1
1.1 Latar Belakang	1
1.2 Tujuan	2
II TINJAUAN PUSTAKA	3
2.1 Asuransi Kesehatan dan Risiko Biaya Medis	3
2.2 Faktor Risiko Biaya Medis	3
2.3 Tahapan Pra-Pemrosesan Data	4
2.4 <i>Machine Learning</i> untuk Klasifikasi	7
2.5 Evaluasi Model Klasifikasi	10
III METODE PENELITIAN	12
3.1 Studi Literatur	12
3.2 Pengumpulan Data	12
3.3 Alat Analisis	12
3.4 Langkah-Langkah Analisis	12
IV HASIL DAN PEMBAHASAN	14
4.1 Eksplorasi Data	14
4.2 Persiapan Data	16
4.3 Uji Multikolinearitas	17
4.4 Regresi Logistik Multinomial	17
4.5 <i>K-Nearest Neighbor</i>	19
4.6 Perbandingan Evaluasi Model	20
4.7 Teknik <i>k-Fold Cross Validation</i>	21
V SIMPULAN DAN SARAN	23
5.1 Simpulan	23
5.2 Saran	23
DAFTAR PUSTAKA	24
LAMPIRAN	28
RIWAYAT HIDUP	39



DAFTAR TABEL

1	<i>Confusion matrix</i> pada tiga kelas	10
2	Nilai VIF	17
3	Uji kelayakan model	18
4	<i>Confusion matrix</i> regresi logistik multinomial	18
5	Evaluasi metrik regresi logistik multinomial	19
6	<i>Confusion matrix K-nearest neighbor</i>	20
7	Evaluasi metrik <i>K-nearest neighbor</i>	20
8	Perbandingan metrik evaluasi antara regresi logistik multinomial dengan <i>K-nearest neighbor</i>	21
9	Hasil rata-rata <i>k-fold cross validation</i> ($k = 5$)	22

DAFTAR GAMBAR

1	Skema <i>5-fold cross validation</i>	10
2	<i>Boxplot</i> distribusi usia berdasarkan tingkat risiko	14
3	<i>Boxplot</i> distribusi indeks massa tubuh berdasarkan tingkat risiko	15
4	<i>Boxplot</i> distribusi banyaknya anak berdasarkan tingkat risiko	15
5	<i>Bar chart</i> status merokok berdasarkan tingkat risiko	16
6	Grafik pemilihan nilai K optimal	20
7	Perbandingan kinerja model regresi multinomial dengan model KNN	21
8	Perbandingan kinerja <i>k-fold cross validation</i> ($k = 5$) antara metode regresi logistik multinomial dan <i>K-nearest neighbor</i>	22

DAFTAR LAMPIRAN

1	Program RStudio untuk <i>install package</i> , <i>input</i> data, dan pembuatan variabel risiko	29
2	Program RStudio untuk visualisasi data	30
3	Program RStudio untuk mengubah variabel kategori menjadi kategori faktor, <i>split</i> data <i>testing</i> dan data <i>training</i>	31
4	Program RStudio untuk uji multikolinearitas, uji kelayakan model, <i>confusion matrix</i> regresi logistik multinomial	32
5	Program RStudio untuk <i>encoding</i> variabel <i>smoker</i> , <i>scaling</i> data, pencarian nilai K optimal, dan <i>confusion matrix K-nearest neighbor</i>	33
6	Program RStudio untuk evaluasi perbandingan model regresi logistik multinomial dan <i>K-nearest neighbor</i>	35
7	Program RStudio untuk <i>k-fold cross validation</i> model regresi logistik multinomial dan <i>K-nearest neighbor</i>	36