

PERBANDINGAN METODE *MACHINE LEARNING* *RANDOM FOREST* DAN *XGBOOST* UNTUK MEMPREDIKSI RISIKO KEJADIAN *STROKE*

NAYLA JASMINE MAHIRADEWI



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



PERNYATAAN MENGENAI SKRIPSI DAN SUMBER INFORMASI SERTA PELIMPAHAN HAK CIPTA

Dengan ini saya menyatakan bahwa skripsi dengan judul “Perbandingan Metode *Machine Learning Random Forest* dan *XGBoost* untuk Memprediksi Risiko Kejadian *Stroke*” 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, Agustus 2025

Nayla Jasmine Mahiradewi
G5402211067

ABSTRAK

NAYLA JASMINE MAHIRADEWI. Perbandingan Metode *Machine Learning Random Forest* dan *XGBoost* untuk Memprediksi Risiko Kejadian *Stroke*. Dibimbing oleh RUHIYAT dan RAHMA ANISA.

Stroke merupakan kondisi medis serius yang dapat menyebabkan kecacatan permanen atau bahkan kematian jika tidak ditangani dengan cepat. Penelitian ini menggunakan data sekunder dari *dataset stroke* yang terdiri dari 5,110 observasi, mencakup informasi mengenai usia, jenis kelamin, riwayat hipertensi, penyakit jantung, kadar glukosa rata-rata, indeks massa tubuh, tipe tempat tinggal, tipe pekerjaan, dan riwayat merokok. Dua metode *machine learning*, *random forest* dan *XGBoost*, dipilih untuk memprediksi risiko *stroke*. Kedua metode tersebut dianggap cocok untuk data besar dan lebih fleksibel karena tidak memerlukan asumsi parametrik. Penelitian ini bertujuan untuk membandingkan kinerja kedua metode tersebut dalam memprediksi risiko *stroke*. Evaluasi model dilakukan menggunakan metrik *balanced accuracy*, sensitivitas, spesifisitas, dan *F1-score*. Hasil penelitian menunjukkan bahwa *XGBoost* unggul pada *balanced accuracy* dibandingkan *random forest*. Analisis *variable importance* pada model *XGBoost* mengidentifikasi usia, rata-rata kadar glukosa, dan indeks massa tubuh sebagai faktor paling berpengaruh terhadap risiko *stroke*. Penelitian ini diharapkan dapat membantu pengembangan sistem pendukung keputusan untuk deteksi dini *stroke* serta memberikan wawasan mengenai faktor risiko utama yang perlu diperhatikan.

Kata kunci: *machine learning*, *random forest*, *stroke*, *XGBoost*.



ABSTRACT

NAYLA JASMINE MAHIRADEWI. The Comparison of Random Forest and XGBoost Machine Learning Methods for Predicting Stroke Risk. Supervised by RUHIYAT and RAHMA ANISA.

Stroke is a serious medical condition that can lead to permanent disability or even death if not promptly managed. This study utilizes secondary data from a stroke dataset comprising 5,110 observations, including information on age, gender, history of hypertension, heart disease, average glucose level, body mass index, residence type, work type, and smoking history. Two machine learning methods, random forest and XGBoost, were chosen to predict stroke risk. Both methods are considered suitable for large datasets and are more flexible as they do not require parametric assumptions. This research aims to compare the performance of these two methods in predicting stroke risk. Model evaluation was conducted using balanced accuracy, sensitivity, specificity, and F1-score metrics. The results indicated that XGBoost performs superiorly in terms of balanced accuracy compared to random forest. Variable importance analysis from the XGBoost model identified age, average glucose level, and body mass index as the most influential factors contributing to stroke risk. This study is expected to contribute to the development of decision support systems for early stroke detection and provide insights into key risk factors that need attention.

Keywords: machine learning, random forest, stroke, XGBoost.



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 2025
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.

PERBANDINGAN METODE *MACHINE LEARNING* *RANDOM FOREST* DAN *XGBOOST* UNTUK MEMPREDIKSI RISIKO KEJADIAN *STROKE*

NAYLA JASMINE MAHIRADEWI

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
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:
1 Dr. Ir. Budi Suharjo, M.S.



Judul Skripsi : Perbandingan Metode *Machine Learning Random Forest* dan *XGBoost* untuk Memprediksi Risiko Kejadian *Stroke*

Nama : Nayla Jasmine Mahiradewi

NIM : G5402211067

@Hak cipta milik IPB University

Disetujui oleh

Pembimbing 1:

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

Pembimbing 2:

Rahma Anisa, S.Stat., M.Si., M.Act.Sc.

Diketahui oleh

Ketua Program Studi Aktuaria:

Dr. Ir. I Gusti Putu Purnaba, DEA.

NIP. 196512181990021001

Tanggal Ujian:
14 Juli 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. Tema yang dipilih dalam penelitian yang dilaksanakan sejak bulan Oktober 2024 sampai bulan Agustus 2025 ini ialah Analisis Model Prediktif dengan judul “Perbandingan Metode *Machine Learning Random Forest* dan *XGBoost* untuk Memprediksi Risiko Kejadian *Stroke*”.

Dengan rasa syukur yang mendalam, penulis menyadari bahwa penyelesaian tugas akhir ini tidak terlepas dari bimbingan, doa, dan dukungan yang diberikan oleh berbagai pihak. Oleh karena itu, penulis ingin menyampaikan rasa terima kasih yang sebesar-besarnya kepada:

1. Ruhiyat, S.Si., M.Si., M.Act.Sc. selaku dosen pembimbing 1 yang dengan penuh kesabaran dan ketulusan memberikan bimbingan, arahan, dan dukungan yang sangat berarti dalam setiap tahap penelitian ini.
2. Rahma Anisa, S.Stat., M.Si., M.Act.Sc. selaku dosen pembimbing 2 yang senantiasa memberikan saran konstruktif dan masukan berharga.
3. Dr. Ir. Budi Suharjo, M.S. selaku dosen penguji yang dengan keahlian dan ketelitian memberikan masukan yang sangat berguna dalam penyempurnaan tugas akhir ini.
4. Seluruh Dosen dan Staf Program Studi Aktuaria SSMI IPB, yang telah memberikan pendidikan dan pengetahuan yang sangat bermanfaat selama masa perkuliahan.
5. Alm. Ayah, Mama, Kakak, Adik, Molly, dan Coco yang senantiasa memberikan dukungan moril dan doa yang tidak pernah putus. Tanpa kasih sayang dan semangat yang diberikan, penulis tidak akan sampai pada titik ini. Mereka adalah sumber kekuatan dan inspirasi terbesar dalam perjalanan akademik ini.
6. Dea, Devina, dan Paulina, yang telah menemani penulis dalam suka dan duka dari awal hingga akhir perjalanan ini. Dukungan mereka memberikan rasa tenang dan kebahagiaan yang tak ternilai.
7. Fikah, Tamara, Lita, sahabat-sahabat “*crepes*” yang selalu memberi semangat, kebahagiaan, dan dukungan moral selama masa perkuliahan. Terima kasih telah menjadi bagian penting dalam perjalanan ini.
8. Teman-teman Aktuaria 58 dan KKN, yang telah menjadi sahabat dalam perjalanan akademik ini, memberikan dukungan, semangat, dan kebahagiaan yang sangat berarti selama masa kuliah.
9. Diri sendiri, yang telah berjuang keras untuk menyelesaikan tugas akhir ini. Terima kasih kepada diri sendiri yang telah melalui banyak tantangan dan rintangan dengan tekad yang kuat untuk menyelesaikan pendidikan ini.

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

Bogor, Agustus 2025

Nayla Jasmine Mahiradewi



DAFTAR ISI

DAFTAR TABEL	x
DAFTAR GAMBAR	x
DAFTAR LAMPIRAN	x
PENDAHULUAN	1
1.1 Latar Belakang	1
1.2 Rumusan Masalah	3
1.3 Tujuan	3
1.4 Manfaat	3
TINJAUAN PUSTAKA	4
2.1 <i>Stroke</i>	4
2.2 <i>Machine Learning</i>	4
2.3 <i>Random Forest</i>	5
2.4 <i>Extreme Gradient Boosting</i>	6
2.5 <i>K-Fold Cross Validation</i>	7
2.6 <i>Hyperparameter Tuning</i>	8
2.7 Evaluasi Kinerja Model	9
2.8 <i>Receiver Operating Characteristic Curve</i>	10
III METODE	12
3.1 Sumber Data	12
3.2 Prosedur Analisis Data	13
IV HASIL DAN PEMBAHASAN	14
4.1 Prapemrosesan Data	14
4.2 <i>Exploratory Data Analysis</i>	14
4.3 <i>Hyperparameter Tuning</i>	17
4.4 Penentuan Metode Terbaik	18
4.5 Penerapan Model Terbaik pada Data Uji	20
V SIMPULAN DAN SARAN	24
5.1 Simpulan	24
5.2 Saran	24
DAFTAR PUSTAKA	25
LAMPIRAN	29
RIWAYAT HIDUP	41

DAFTAR TABEL

1	<i>Confusion matrix</i>	9
2	Deskripsi <i>dataset</i>	12
3	<i>Hyperparameter random forest</i>	17
4	<i>Hyperparameter XGBoost</i>	17
5	Kombinasi <i>hyperparameter</i> terbaik menurut <i>random forest</i>	18
6	Kombinasi <i>hyperparameter</i> terbaik menurut <i>XGBoost</i>	18
7	<i>Confusion matrix</i> dari <i>XGBoost</i>	20
8	Hasil metrik evaluasi model <i>XGBoost</i>	20

DAFTAR GAMBAR

1	Ilustrasi <i>5-fold cross validation</i> (Hastie <i>et al.</i> 2009)	7
2	Ilustrasi dari <i>k-fold cross validation</i> (Sevinç 2022)	8
3	Contoh dari <i>ROC curves</i> (Brownlee 2020)	11
4	Proporsi peubah target	14
5	Visualisasi hubungan peubah prediktor numerik dengan peubah target	15
6	Visualisasi hubungan peubah-peubah prediktor kategorik dengan peubah target	16
7	<i>Boxplot</i> perbandingan model	19
8	Kurva ROC untuk model final <i>XGBoost</i>	21
9	<i>Variable importance XGBoost</i>	21

DAFTAR LAMPIRAN

1	Hasil dari <i>feature importance</i> model akhir <i>XGBoost</i>	30
2	Kode R untuk <i>hyperparameter tuning</i> risiko <i>stroke</i>	31
3	Kode R untuk prediksi risiko <i>stroke</i>	35