

SKRIPSI

**KLASIFIKASI PENYAKIT STROKE MENGGUNAKAN MACHINE
LEARNING DENGAN IMBALANCED DATA**



ZHAFIRA HAJAR

**PROGRAM STUDI S1 SISTEM INFORMASI
DEPARTEMEN MATEMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS AIRLANGGA**

2022

SKRIPSI

**KLASIFIKASI PENYAKIT STROKE MENGGUNAKAN MACHINE
LEARNING DENGAN IMBALANCED DATA**



ZHAFIRA HAJAR

081811633035

**PROGRAM STUDI S1 SISTEM INFORMASI
DEPARTEMEN MATEMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS AIRLANGGA
2022**

**KLASIFIKASI PENYAKIT STROKE MENGGUNAKAN MACHINE
LEARNING DENGAN IMBALANCED DATA**

SKRIPSI

**Sebagai Salah Satu Syarat Untuk Memperoleh Gelar Sarjana Komputer
Bidang Sistem Informasi Pada Fakultas Sains Dan Teknologi Universitas
Airlangga**

Oleh :

Zhafira Hajar

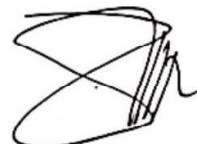
NIM : 081811633035

Disetujui Oleh :

Pembimbing I


Dr. Rimuljo Hendradi, S.Si., M.Si.
197102111997021001

Pembimbing II

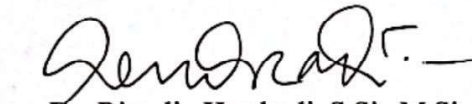

Drs. Eto Wuryanto, DEA.
196609281991021001

LEMBAR PENGESAHAN NASKAH SKRIPSI

Judul : Klasifikasi Penyakit *Stroke* Menggunakan *Machine Learning* dengan *Imbalanced Data*
Penyusun : Zhafira Hajar
NIM : 081811633035
Pembimbing I : Dr. Rimuljo Hendradi, S.Si., M.Si.
Pembimbing II : Drs. Eto Wuryanto, DEA.
Tanggal Ujian : 31 Agustus 2022

Disetujui Oleh,

Pembimbing I


Dr. Rimuljo Hendradi, S.Si., M.Si.
197102111997021001

Pembimbing II

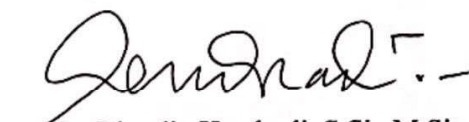

Drs. Eto Wuryanto, DEA.
196609281991021001

Mengetahui,

Ketua Departemen Matematika
Fakultas Sains dan Teknologi
Universitas Airlangga


Dr. Herry Suprajitno, S.Si., M.Si.
NIP. 196804041994031020

Koordinator Program Studi
S1 Sistem Informasi
Fakultas Sains dan Teknologi
Universitas Airlangga


Dr. Rimuljo Hendradi, S.Si., M.Si.
NIP. 197102111997021001

SURAT PERNYATAAN TENTANG ORISINALITAS

Yang bertanda tangan dibawah ini, saya:

Nama : Zhafira Hajar

NIM : 081811633035

Program Studi : S1 Sistem Informasi

Fakultas : Sains dan Teknologi

Jenjang : Sarjana (S1)

Menyatakan bahwa saya tidak melakukan kegiatan plagiat dalam penulisan skripsi saya yang berjudul :

“KLASIFIKASI PENYAKIT STROKE MENGGUNAKAN MACHINE LEARNING DENGAN IMBALANCED DATA”

Apabila suatu saat nanti terbukti melakukan tindakan plagiat, maka saya akan menerima sanksi yang telah ditetapkan.

Demikian surat pernyataan ini saya buat dengan sebenar-benarnya.

Surabaya, 15 Agustus 2022



Zhafira Hajar

NIM. 081811633035

PEDOMAN PENGGUNAAN SKRIPSI

Skripsi ini tidak dipublikasikan, namun tersedia di perpustakaan dalam lingkungan Universitas Airlangga, diperkenankan untuk dipakai sebagai referensi kepustakaan, tetapi pengutipan harus seizin penulis dan harus menyebutkan sumbernya sesuai kebiasaan ilmiah.

Dokumen skripsi ini merupakan hak milik Universitas Airlangga.

KATA PENGANTAR

Puji syukur berkat kehadiran Tuhan Yang Maha Esa, penyusun dapat menyelesaikan penyusunan proposal tugas akhir skripsi yang berjudul **Klasifikasi Penyakit *Stroke* Menggunakan *Machine Learning* dengan *Imbalanced Data***. Berkat rahmat-Nya, penyusun dapat mengerjakan penyusunan dengan lancar sampai pada tahap ini.

Terima kasih juga penyusun ucapkan kepada seluruh pihak yang telah mendukung maupun ikut membantu secara langsung dalam pengerjaan proposal ini hingga dapat terselesaikan dengan baik. Ucapan tersebut ditujukan oleh penyusun pada:

1. Tuhan YME yang senantiasa memberikan berkah dan rahmat-Nya sehingga penelitian ini dapat berjalan lancar dan memberi kekuatan pada penyusun untuk menyelesaikan skripsi ini.
2. Orang tua dan keluarga yang selalu mendukung dan memberi penyusun semangat serta fasilitas sehingga dapat menyelesaikan skripsi ini dengan baik.
3. Dr. Rimuljo Hendradi, S.Si., M.Si. selaku dosen pembimbing I sekaligus kepala Program Studi S1 Sistem Informasi Universitas Airlangga yang sudah memberikan arahan serta ilmu semenjak penyusun masih merupakan mahasiswa baru sampai dengan saat ini dan mendorong penyusun untuk berkembang menjadi mahasiswa yang baik.
4. Drs. Eto Wuryanto, DEA selaku dosen pembimbing II yang selalu memberi masukan dan arahan terkait skripsi anak bimbingnya hingga bisa menghasilkan penelitian yang baik, sehingga penyusun dapat menyelesaikan penelitian ini.
5. Seluruh dosen Program Studi S1 Sistem Informasi Universitas Airlangga yang telah memberi ilmu akademik maupun non-akademik yang bermanfaat selama perkuliahan.

6. Teman-teman mahasiswa Program Studi S1 Sistem Informasi Universitas Airlangga khususnya angkatan 2018 yang telah menemani penyusun dari awal perkuliahan hingga saat ini, yang telah berbagi cerita dan pengalaman serta berkembang bersama sebagai mahasiswa.
7. Kakak-kakak tingkat yang telah memberi ilmu serta saran kepada penyusun, serta adik-adik tingkat yang telah menerima penyusun sebagai senior dalam hal organisasi maupun sebagai mahasiswa.
8. Ahmad Ikhlil Nur, seluruh sahabat, serta teman-teman yang sudah menemani penyusun hingga saat ini serta berbagi banyak pengalaman suka dan duka.

Akhir kata, penyusun menyadari jika terdapat kekurangan pada skripsi ini, oleh karena itu, kritik dan saran akan sangat bermanfaat. Penyusun berharap, semoga penelitian ini kelak dapat selesai dengan baik dan menjadi sumber ilmu bagi seluruh praktisi maupun akademisi untuk masa depan yang lebih baik.

Surabaya, 15 Agustus 2022

Penyusun



Zhafira Hajar

Zhafira Hajar, 2022, **Klasifikasi Penyakit Stroke Menggunakan Machine Learning dengan Imbalanced Data**, Skripsi ini di bawah bimbingan Dr. Rimuljo Hendradi, S.Si., M.Si., dan Drs. Eto Wuryanto, DEA., Departemen Matematika, Fakultas Sains dan Teknologi, Universitas Airlangga, Surabaya.

ABSTRAK

Otak adalah organ terpenting pada tubuh manusia, dan salah satu penyakit yang dapat menyerang otak adalah *stroke*. Pencegahan *stroke* dapat dilakukan dengan edukasi terkait faktor risiko penyebab terjadinya penyakit *stroke*. Edukasi faktor risiko *stroke* dapat dilakukan dengan *feature importance*, yang merupakan salah satu keluaran klasifikasi menggunakan *machine learning*. Salah satu tantangan klasifikasi adalah ketidakseimbangan data. Tujuan penelitian ini adalah melakukan klasifikasi menggunakan *machine learning* dengan data yang tidak seimbang. Klasifikasi dilakukan dengan menggunakan metode *Support Vector Machine* (SVM), *Random Forest* (RF), *Extreme Gradient Boosting* (XGBoost). Untuk menangani ketidakseimbangan data, digunakan tiga metode resampling data, yaitu *Random Undersampling* (RUS), *Random Oversampling* (ROS), dan *Synthetic Minonrity Oversampling Technique* (SMOTE). Klasifikasi dilakukan dengan menggunakan tiga rasio training testing yang berbeda, yaitu 60% *training* 40% *testing*, 70% *training* & 30% *testing* dan 80% *training* 20% *testing*. Data yang digunakan dalam penelitian ini adalah *The International Stroke Trial Database* yang memiliki 19.435 data dan empat kelas, serta memiliki ketidakseimbangan antar kelasnya. Dari 13.230 data setelah *preprocessing*, terdapat 12.041 data yang termasuk kelas “DDIAGISC” atau *ischemic stroke*. Pertama-tama klasifikasi dilakukan dengan menggunakan *default parameter*. Kemudian dilakukan *5-fold cross validation* untuk menemukan nilai parameter terbaik untuk setiap metode klasifikasi. Setelah melakukan klasifikasi menggunakan parameter terbaik, hasil menunjukkan bahwa setelah dilakukan *hyper parameter tuning*, hasil paling optimal dengan nilai akurasi 44,728%, AUC 0,713, dan *recall / sensitivity* 40% didapatkan menggunakan metode klasifikasi RF yang diimplementasikan dengan menggunakan SMOTE untuk *resampling data*, dan rasio 60% *training* 40% *testing* sebagai rasio *training testing* nya. Hasil *feature importance* dari semua metode klasifikasi menjelaskan bahwa atribut “RVISINF” atau indikator apakah pasien memiliki *infarct* pada hasil CT scan merupakan atribut yang paling berpengaruh dalam melakukan klasifikasi *stroke* pada data.

Kata Kunci: *stroke*, klasifikasi, *machine learning*, *imbalanced data*, *resampling data*, *feature importance*

Zhafira Hajar, 2022, **Stroke Classification Using Machine Learning with Imbalanced Data**, This undergraduate thesis was supervised by Dr. Rimuljo Hendradi, S.Si., M.Si.,. and Drs. Eto Wuryanto, DEA., Departemen Matematika, Fakultas Sains dan Teknologi, Universitas Airlangga, Surabaya.

ABSTRACT

The brain is the most important organ in the human body, and one of the diseases that can attack the brain is stroke. Stroke prevention can be done with education related to risk factors that cause stroke. Stroke risk factor education can be done with feature importance, which is one of the classification outputs using machine learning. One of the classification challenges is data imbalance. The purpose of this research is to classify using machine learning with imbalanced data. Classification was carried out using the Support Vector Machine (SVM), Random Forest (RF), Extreme Gradient Boosting (XGBoost) methods. To deal with data imbalances, three data resampling methods were used, namely Random Undersampling (RUS), Random Oversampling (ROS), and Synthetic Minority Oversampling Technique (SMOTE). Classification is done using three different ratios of training testing, namely 60% training 40% testing, 70% training & 30% testing and 80% training 20% testing. The data used in this study is The International Stroke Trial Database which has 19,435 data and four classes, and has an imbalance between classes. Of the 13,230 data after preprocessing, there are 12,041 data belonging to the "DDIAGISC" or ischemic stroke class. First of all the classification is done by using the default parameters. Then, 5-fold cross validation was carried out to find the best parameter values for each classification method. After classifying using the best parameters, the results show that after hyper-parameter tuning, the most optimal results with an accuracy value of 44.728%, AUC 0.713, and recall / sensitivity 40% were obtained using the RF classification method implemented using SMOTE for resampling data, and the ratio 60% training 40% testing as its training testing ratio. The results of the feature importance of all classification methods explain that the attribute "RVISINF" or an indicator of whether the patient has an infarct on the CT scan results is the most influential attribute in classifying strokes on the data.

Keyword: *stroke, classification, machine learning, imbalanced data, resampling data, feature importance*

DAFTAR ISI

SKRIPSI.....	ii
LEMBAR PENGESAHAN NASKAH SKRIPSI	iv
SURAT PERNYATAAN TENTANG ORISINALITAS	v
PEDOMAN PENGGUNAAN SKRIPSI	vi
KATA PENGANTAR	vii
ABSTRAK	ix
ABSTRACT	x
DAFTAR ISI	xi
DAFTAR TABEL	xiii
DAFTAR GAMBAR	xiv
DAFTAR LAMPIRAN	xvii
BAB I PENDAHULUAN	18
1.1 Latar Belakang	18
1.2 Rumusan Masalah	22
1.3 Tujuan Penelitian.....	22
1.4 Manfaat Penelitian.....	22
1.5 Batasan Penelitian	23
BAB II TINJAUAN PUSTAKA	24
2.1 Anatomi dan Fisiologi Otak.....	24
2.2 Stroke.....	25
2.3 Machine Learning.....	27
2.4 Klasifikasi	27
2.5 Imbalanced Data.....	31
2.6 Evaluasi Sistem	33
2.7 Penelitian Sebelumnya.....	38
BAB III METODE PENELITIAN	43
3.1 Pengumpulan Data.....	44
3.2 Pemilihan Atribut	44
3.3 Preprocessing Data	46
3.4 Resampling Data	47
3.5 Klasifikasi	48

3.6 Evaluasi Sistem	48
BAB IV HASIL PENELITIAN DAN PEMBAHASAN	50
4.1 Hasil Penelitian	50
4.2 Pembahasan.....	69
BAB V KESIMPULAN DAN SARAN.....	80
5.1 Kesimpulan	80
5.2 Saran	80
LAMPIRAN	86

DAFTAR TABEL

Nomor	Judul Tabel	Halaman
Tabel 2.1	Tabel ilustrasi <i>confussion matrix</i>	35
Tabel 2.2	Perbandingan penelitian sebelumnya dengan dan penelitian ini.	41
Tabel 3.1	Atribut <i>predictor</i> masukan sistem.	45
Tabel 3.2	<i>Class target</i> masukan sistem.....	46
Tabel 4.1	Jumlah <i>missing value</i> pada setiap atribut.....	50
Tabel 4.2	Distribusi <i>training data</i> sebelum dan sesudah dilakukan <i>resampling</i> menggunakan RUS pada semua rasio <i>training testing data</i>	55
Tabel 4.3	Distribusi <i>training data</i> sebelum dan sesudah dilakukan <i>resampling</i> menggunakan ROS pada semua rasio <i>training testing data</i>	56
Tabel 4.4	Distribusi <i>training data</i> sebelum dan sesudah dilakukan <i>resampling</i> menggunakan SMOTE pada semua rasio <i>training testing data</i>	58
Tabel 4.5	Hasil dari 5-Fold Cross Validation pada metode SVM.	61
Tabel 4.6	Hasil dari 5-Fold Cross Validation pada metode RF.	61
Tabel 4.7	Hasil dari 5-Fold Cross Validation pada metode XGBoost.....	63
Tabel 4.8	Spesifikasi <i>hardware</i> yang digunakan.....	74
Tabel 4.9	waktu komputasi untuk menemukan parameter terbaik menggunakan 5-fold cross validation.....	74
Tabel 4.10	Perbandingan hasil akhir klasifikasi saat menggunakan <i>default parameter</i> dan menggunakan parameter hasil <i>hyper parameter tuning</i>	77
Tabel 4.11	Perbandingan nilai akurasi, AUC dan <i>recall</i> sebelum dan sesudah dilakukan <i>resampling</i>	78

DAFTAR GAMBAR

Nomor	Judul Gambar	Halaman
Gambar 2.1	Anatomi otak manusia beserta fungsinya (Martini et al., 2015).	24
Gambar 2.2	<i>Internal carotid arteries</i> dan <i>vertebral arteries</i> (Kerry & Taylor, 2010).	25
Gambar 2.3	Perbedaan <i>ischemic stroke</i> dan <i>hemorrhagic stroke</i> (Crouch, 2013).	26
Gambar 2.4	Gambaran <i>binary classification</i> (Herrera et al., 2016).	28
Gambar 2.5	Gambaran <i>multiclass classification</i> (Herrera et al., 2016).....	28
Gambar 2.6	Gambaran <i>multilabel classification</i> (Herrera et al., 2016).....	29
Gambar 2.7	<i>Hyperplane</i> dengan margin maksimal (Manning et al., 2012).....	31
Gambar 2.8	Gambaran <i>imbalanced data</i> pada <i>medical dataset</i> (Liu et al., 2019)	32
Gambar 2.9	SMOTE dalam melakukan <i>resampling data</i> (Chen et al., 2021).	33
Gambar 2.10	Partisi testing dan training data dengan K=5 (Xu et al., 2019).	34
Gambar 2.11	Hasil dari setiap kernel yang diuji pada penelitian (R. S. & Kumar Sukesh, 2016)	40
Gambar 2.12	Hasil <i>feature importance</i> dari penelitian (Liu et al., 2019).....	40
Gambar 3.1	Diagram alur penelitian.....	43
Gambar 4.1	Contoh data yang memiliki <i>unkown data</i> pada atributnya.....	51
Gambar 4.2	Contoh data yang memiliki <i>unkown data</i> pada <i>target class</i>	51
Gambar 4.3	Perbedaan <i>target class</i> sebelum dan sesudah dilakukan modifikasi.	52
Gambar 4.4	Distribusi data sebelum dilakukan <i>resampling data</i>	52
Gambar 4.5	Hasil klasifikasi sebelum dilakukan <i>resampling data</i>	53

Gambar 4.6 Distribusi kelas pada <i>training data</i> menggunakan rasio 60% <i>training</i> dan 40% <i>testing</i> sebelum dilakukan <i>resampling</i> RUS.....	54
Gambar 4.7 Distribusi kelas pada <i>training data</i> menggunakan rasio 60% <i>training</i> dan 40% <i>testing</i> setelah dilakukan <i>resampling</i> RUS.....	54
Gambar 4.8 Distribusi kelas pada <i>training data</i> menggunakan rasio 60% <i>training</i> dan 40% <i>testing</i> sebelum dilakukan <i>resampling</i> ROS.....	55
Gambar 4.9 Distribusi kelas pada <i>training data</i> menggunakan rasio 60% <i>training</i> dan 40% <i>testing</i> setelah dilakukan <i>resampling</i> ROS.....	56
Gambar 4.10 Distribusi kelas pada <i>training data</i> menggunakan rasio 60% <i>training</i> dan 40% <i>testing</i> sebelum dilakukan <i>resampling</i> SMOTE.....	57
Gambar 4.11 Distribusi kelas pada <i>training data</i> menggunakan rasio 60% <i>training</i> dan 40% <i>testing</i> setelah dilakukan <i>resampling</i> SMOTE	57
Gambar 4.12 Hasil <i>accuracy</i> menggunakan <i>default parameter</i>	59
Gambar 4.13 Hasil AUC menggunakan <i>default parameter</i>	60
Gambar 4.14 Hasil <i>recall/sensitivity</i> menggunakan default parameter.....	60
Gambar 4.15 Nilai <i>accuracy</i> sebelum dan sesudah dilakukan <i>5-Fold Cross Validation</i> pada rasio 60% <i>training</i> dan 40% <i>testing</i>	64
Gambar 4.16 Nilai AUC sebelum dan sesudah dilakukan <i>5-Fold Cross Validation</i> pada rasio 60% <i>training</i> dan 40% <i>testing</i>	64
Gambar 4.17 Nilai <i>recall/sensitivity</i> sebelum dan sesudah dilakukan <i>5-Fold Cross Validation</i> pada rasio 60% <i>training</i> dan 40% <i>testing</i>	65
Gambar 4.18 Perbedaan <i>confusion matrix</i> sebelum dan sesudah dilakukan <i>5-Fold Cross Validation</i> pada metode RF menggunakan rasio 60% <i>training</i> 40% <i>testing</i> dan metode SMOTE.....	66
Gambar 4.19 Hasil <i>accuracy</i> menggunakan parameter yang sudah ditentukan menggunakan <i>5-fold cross validation</i>	66

Gambar 4.20 Hasil AUC menggunakan parameter yang sudah ditentukan menggunakan <i>5-fold cross validation</i>	67
Gambar 4.21 Hasil <i>recall/sensitivity</i> menggunakan parameter yang sudah ditentukan menggunakan <i>5-fold cross validation</i>	67
Gambar 4.22 <i>Feature importance</i> hasil klasifikasi menggunakan <i>Random Forest</i>	68
Gambar 4.23 <i>Feature importance</i> hasil klasifikasi menggunakan XGBoost.....	68
Gambar 4.24 <i>Feature importance</i> hasil klasifikasi menggunakan SVM.....	69
Gambar 4.25 Ilustrasi SVM dalam menangani kasus <i>multiclass classification</i> menggunakan pendekatan OVO atau OVR/OVA (Haixiang et al., 2017c).....	72
Gambar 4.26 perbandingan jumlah <i>evaluation metric</i> atau metode evaluasi, yang digunakan oleh penelitian sebelumnya (Haixiang et al., 2017c).....	73

DAFTAR LAMPIRAN

Nomor	Judul Lampiran	Halaman
Lampiran 1.	Hasil <i>accuracy</i> , AUC, dan <i>recall</i> keseluruhan iterasi menggunakan <i>default parameter</i>	86
Lampiran 2.	Hasil <i>accuracy</i> , AUC, dan <i>recall</i> seluruh iterasi menggunakan parameter hasil <i>hyper parameter tuning</i>	90
Lampiran 3.	<i>Confusion matrix</i> hasil klasifikasi menggunakan <i>default parameter</i> menggunakan metode SVM	95
Lampiran 4.	<i>Confusion matrix</i> hasil klasifikasi menggunakan <i>default parameter</i> menggunakan metode RF.....	105
Lampiran 5.	<i>Confusion matrix</i> hasil klasifikasi menggunakan <i>default parameter</i> menggunakan metode XGBoost.....	115
Lampiran 6.	<i>Confusion matrix</i> hasil klasifikasi menggunakan parameter hasil <i>hyper parameter tuning</i> menggunakan metode SVM	124
Lampiran 7.	<i>Confusion matrix</i> hasil klasifikasi menggunakan parameter hasil <i>hyper parameter tuning</i> menggunakan metode RF	134
Lampiran 8.	<i>Confusion matrix</i> hasil klasifikasi menggunakan parameter hasil <i>hyper parameter tuning</i> menggunakan metode XGBoost	145
Lampiran 9.	<i>Link</i> Google Drive.....	155