TESIS

PREDIKSI STROKE MENGGUNAKAN METODE SYNTHETIC MINORITY OVER-SAMPLING TECHNIQUE DAN XTREME GRADIENT BOOSTING

Disusun oleh:

Nama : **Abd Mizwar A. Rahim**
NIM : **19.51.1274**
Konsentrasi : **Business Intelligence**

PROGRAM STUDI S2 TEKNIK INFORMATIKA
PROGRAM PASCASARJANA UNIVERSITAS AMIKOM YOGYAKARTA
YOGYAKARTA
2022

TESIS

PREDIKSI STROKE MENGGUNAKAN METODE SYNTHETIC MINORITY OVER-SAMPLING TECHNIQUE DAN XTREME GRADIENT BOOSTING

STROKE PREDICTION USING SYNTHETIC MINORITY OVER-SAMPLING TECHNIQUE AND XTREME GRADIENT BOOSTING METHOD

Diajukan untuk memenuhi salah satu syarat memperoleh derajat Magister

Disusun oleh:

Nama	**: Abd Mizwar A. Rahim**
NIM	**: 19.51.1274**
Konsentrasi	**: Business Intelligence**

PROGRAM STUDI S2 TEKNIK INFORMATIKA

PROGRAM PASCASARJANA UNIVERSITAS AMIKOM YOGYAKARTA

YOGYAKARTA

2022

HALAMAN PENGESAHAN

PREDIKSI STROKE MENGGUNAKAN METODE SYNTHETIC MINORITY OVER-SAMPLING TECHNIQUE DAN XTREME GRADIENT BOOSTING

STROKE PREDICTION USING SYNTHETIC MINORITY OVER-SAMPLING TECHNIQUE AND XTREME GRADIENT BOOSTING METHOD

Dipersiapkan dan Disusun oleh

Abd Mizwar A. Rahim

19.51.1274

Telah Diujikan dan Dipertahankan dalam Sidang Ujian Tesis
Program Studi S2 Teknik Informatika
Program Pascasarjana Universitas AMIKOM Yogyakarta
pada hari Kamis, 7 Juli 2022

Tesis ini telah diterima sebagai salah satu persyaratan
untuk memperoleh gelar Magister Komputer

Yogyakarta, 7 Juli 2022
Rektor

Prof. Dr. M. Suyanto, M.M.
NIK. 190302001

HALAMAN PERSETUJUAN

PREDIKSI STROKE MENGGUNAKAN METODE SYNTHETIC MINORITY OVER-SAMPLING TECHNIQUE DAN XTREME GRADIENT BOOSTING

STROKE PREDICTION USING SYNTHETIC MINORITY OVER-SAMPLING TECHNIQUE AND XTREME GRADIENT BOOSTING METHOD

Dipersiapkan dan Disusun oleh

Abd Mizwar A. Rahim

19.51.1274

Telah Diujikan dan Dipertahankan dalam Sidang Ujian Tesis
Program Studi S2 Teknik Informatika
Program Pascasarjana Universitas AMIKOM Yogyakarta
pada hari Kamis, 7 Juli 2022

Pembimbing Utama

Dr. Andi Sunyoto, M.Kom.
NIK. 190302052

Anggota Tim Penguji

Prof. Dr. Kusrini, M.Kom.
NIK. 190302106

Pembimbing Pendamping

M. Rudyanto Arief, M.T.
NIK. 190302098

Prof. Dr. Ema Utami, S.Si., M.Kom.
NIK. 190302037

Dr. Andi Sunyoto, M.Kom.
NIK. 190302052

Tesis ini telah diterima sebagai salah satu persyaratan
untuk memperoleh gelar Magister Komputer

Yogyakarta, 7 Juli 2022
Direktur Program Pascasarjana

Prof. Dr. Kusrini, M.Kom.
NIK. 190302106

HALAMAN PERNYATAAN KEASLIAN TESIS

Yang bertandatangan di bawah ini,

Nama mahasiswa : Abd Mizwar A. Rahim
NIM : 19.51.1274
Konsentrasi : Business Intelligence

Menyatakan bahwa Tesis dengan judul berikut:
Prediksi Stroke Menggunakan Metode Synthetic Minority Over-Sampling Technique Dan Xtreme Gradient Boosting

Dosen Pembimbing Utama : Dr. Andi Sunyoto, M.Kom.
Dosen Pembimbing Pendamping : M. Rudyanto Arief, M.T.

1. Karya tulis ini adalah benar-benar ASLI dan BELUM PERNAH diajukan untuk mendapatkan gelar akademik, baik di Universitas AMIKOM Yogyakarta maupun di Perguruan Tinggi lainnya
2. Karya tulis ini merupakan gagasan, rumusan dan penelitian SAYA sendiri, tanpa bantuan pihak lain kecuali arahan dari Tim Dosen Pembimbing
3. Dalam karya tulis ini tidak terdapat karya atau pendapat orang lain, kecuali secara tertulis dengan jelas dicantumkan sebagai acuan dalam naskah dengan disebutkan nama pengarang dan disebutkan dalam Daftar Pustaka pada karya tulis ini
4. Perangkat lunak yang digunakan dalam penelitian ini sepenuhnya menjadi tanggung jawab SAYA, bukan tanggung jawab Universitas AMIKOM Yogyakarta
5. Pernyataan ini SAYA buat dengan sesungguhnya, apabila di kemudian hari terdapat penyimpangan dan ketidakbenaran dalam pernyataan ini, maka SAYA bersedia menerima SANKSI AKADEMIK dengan pencabutan gelar yang sudah diperoleh, serta sanksi lainnya sesuai dengan norma yang berlaku di Perguruan Tinggi

Yogyakarta, 7 Juli 2022
Yang Menyatakan,

Abd Mizwar A. Rahim

v

HALAMAN PERSEMBAHAN

Dengan segala puji syukur kepada Allah Subhanahu Wata'ala dan atas dukungan dan doa dari orang-orang tercinta, akhirnya Thesis ini dapat saya selesaikan dengan baik. Oleh karena itu, dengan rasa bangga dan bahagia saya haturkan rasa syukur dan terima kasih kepada :

1. Allah Subhanau wata'ala, karena atas izin dan karuniaNya maka skripsi ini selesai pada waktunya.Puji syukur yang tak terhingga kepada Allah Subhanu Wata'ala yang telah meridhoi dan mengabulkan segala doa.
2. Ayah, Ibu, Kakak, Adek yang telah memberikan dukungan moril dan meteri serta doa yang tiada henti untuk kesuksesan saya, doa yang paling khusuk terucap dari kedua orangtua saya. Ucapan terima kasih tidak akan pernah cukup untuk membalas kebaikan kedua orangtua saya, karena itu terimalah persembahan bakti dan cintaku untuk kalian bapak ibuku.
3. Bapak Dr. Andi Sunyoto, M.Kom, dan Bapak M. Rudyanto Arief, M.T. selaku dosen pembimbing, serta Bapak Ibu dosen Amikom lainnya yang selama ini telah tulus ikhlas meluangkan waktunya untuk menuntun dan mengarahkan saya, memberikan bimbingan dan pelajaran yang tiada ternilai harganya. Terima kasih Bapak dan Ibu dosen, jasa kalian akan selalu terpatri dihati.

Terima kasih kepada Rekan-Rekan, serta pihak-pihak lain yang tidak bisa saya sebutkan satu persatu atas bantuannya yang telah memberikan dukungan dalam membantu dalam menyelesaikan Thesis ini.

HALAMAN MOTTO

("Sesungguhnya sesudah kesulitan itu ada kemudahan, sesungguhnya sesudah kesulitan itu ada kemudahan".
(Q.S Asy Syarh ayat 5-6)

"Catat apa yang kita mau, biar di tengah jalan, kita gak tergoda dengan pilihan yang lebih "murah dan mudah". Keep on chasing dreams. Harus bisa apa yang kamu impikan, atau Harus yang lebih baik dari itu. Maka dari itu fokus dan konsisten itu wajib."
(Alit Susanto – Motovlogger [shitlicious])

"Someone is sitting in the shade today because someone planted a tree a long time ago"
(Warren Buffett)

" Success is the sum of small efforts, repeated day-in and day-out."
(Robert Colller)

" Kamu tidak boleh bilang capek, karena kamu harus capek untuk bisa sukses."
(Raffi Ahmad)

KATA PENGANTAR

Puji syukur kehadirat Allah SWT, atas limpahan Rahmat dan Karunia-Nya, sehingga penulis dapat menyelesaikan Thesis dengan judul: Prediksi Stroke Menggunakan Metode Synthetic Minority Over-Sampling Technique Dan Xtreme Gradient Boosting. Ini untuk memenuhi salah satu syarat menyelesaikan studi serta dalam rangka memperoleh gelar Sarjana Pendidikan Strata Dua pada Program Studi Magister Informatika Fakultas Ilmu Komputer Universitas Amikom Yogyakarta.

Thesis ini telah saya selesaikan dengan maksimal berkat kerjasama dan bantuan dari berbagai pihak,oleh karena itu saya ucapkan banyak terima kasih kepada segenap pihak yang telah berkontribusi dalam menyelesaikan skripsi ini

Diluar itu penulis menyadari sebagai manusia biasa bahwa masih banyak kekurangan dalam penulisan thesis ini, baik dari tata bahasa susunan kalimat maupun isi. Oleh sebab itu dengan segala kerendahan hati,saya menerima segala kritik dan saran yang membangun dari pembaca.

Demikian yang bisa saya sampaikan,semoga thesis ini dapat menambah wawasan ilmu pengetahuan dan memberi manfaat bagi masyarakat luas

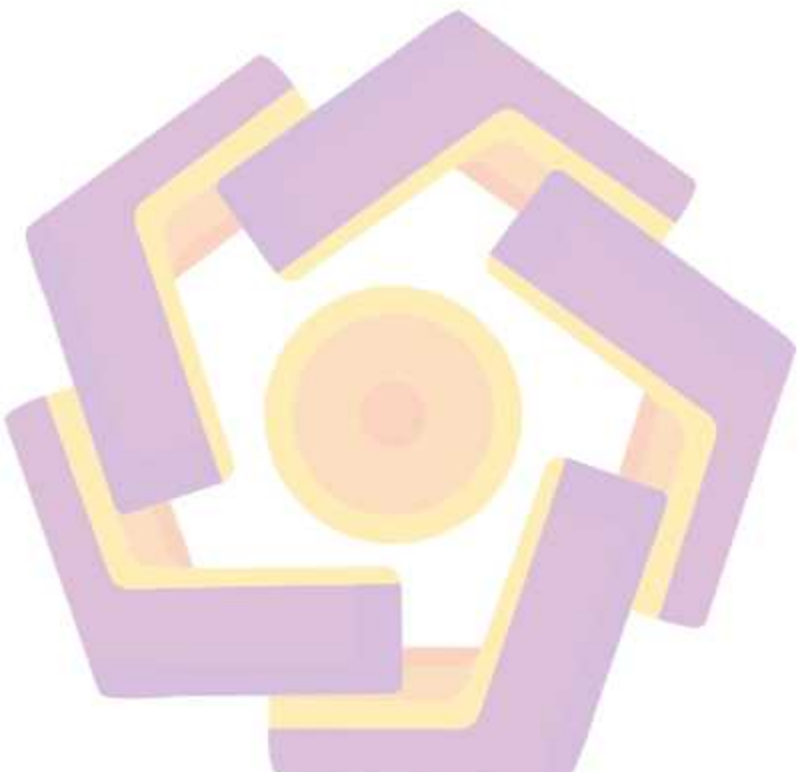
Yogyakarta, 7 Juli 2022

Penulis

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR



INTISARI

Stroke merupakan suatu penyakit atau gangguan fungsional otak berupa lumpuhnya saraf akibat hambatan yang terjadi yaitu aliran darah ke otak yang dapat mengakibatkan lumpuhnya organ tubuh secara menyeluruh atau sebagian, hingga menyebabkan kematian. Stroke merupakan alasan terjadinya kematian seseorang ke dua secara global, sekitar 11% dari total kematian. Terdapat banyak cara agar dapat membantu petugas kesehatan dalam dalam menemukan seseorang terindikasi penyakit stroke atau tidak agar ketika pasien yang mengalami stroke ini dapat ketahui dengan cepat, salah satunya ialah dengan penggunaan Machine learning. terdapat beberapa penelitian sebelumnya dengan study kasus yang sama yaitu prediksi penyakit stroke, dari penelitian yang ada menggunakan beberapa metode machine learning untuk dapat memprediksi seseorang terindikasi penyakit stroke yaitu random forest classifier, ann, svm, c4.5, Naïve bayes, support vector machine,dll sebagainya, hasil penelitian sebelumnya memiliki hasil akurasi yang paling baik adalah 96%, dengan hasil penelitian sebelumnya teknik-teknik yang belum diimplementasikan agar mendapatkan hasil yang optimal. Tujuan dari penelitian ini ialah meningkatkatkan hasil akurasi pada prediksi penyakit stroke menggunakan metode Smote dan machine learning dengan algoritma Xtreme Gradient Boosting untuk mendapatkan hasil akurasi yang lebih baik dari akurasi yang dihasilkan sebelumnya. penggunaan metode Xgboost tanpa menggunakan smote mampu menaikan 1% dari hasil terbaik penelitian sebelumnya yaitu sebesar 96% akurasi, lalu penggunaan metode Xgboost dan menggunakan smote mampu menaikan 3% dari hasil terbaik penelitian sebelumnya yaitu sebesar 99% akurasi.

Kata kunci: Machine Learning, Prediksi Stroke, Xtreme Gradient Boosting, Smote, Ensemble Learning.

ABSTRACT

Stroke is a disease or functional brain disorder in the form of nerve paralysis due to obstacles that occur, namely blood flow to the brain which can result in paralyzing the body's organs completely or partially, causing death. Stroke is the reason for a person's death globally, about 11% of total deaths. There are many ways to be able to assist health workers in finding someone with an indication of stroke or not so that when a patient has a stroke they can find out quickly, one of which is by using machine learning. there are several previous studies with the same case study, namely prediction of stroke, from existing research using several machine learning methods to be able to predict someone indicated by stroke, namely random forest classifier, ann, svm, c4.5, Naïve bayes, support vector machine, etc. etc., the results of previous studies have the best accuracy results is 96%, with the results of previous research techniques that have not been implemented in order to get optimal results. The purpose of this study is to increase the accuracy of stroke prediction using the Smote method and machine learning with the Xtreme Gradient Boosting algorithm to get better accuracy results than the accuracy previously generated. the use of the Xgboost method without using a smote was able to increase 1% of the best results of previous research, which was 96% accuracy, then the use of the Xgboost method and using a smote was able to increase 3% of the best results of previous studies, which was 99% accuracy.

Keyword: Machine Learning, Stroke Prediction, Xtreme Gradient Boosting, Smote, Ensemble Learning.

BAB I

PENDAHULUAN

1.1. Latar Belakang Masalah

Di seluruh dunia, terdapat lebih dari 13,7 juta orang yang mengalami stroke pada setiap tahun, dan dapat mencapai 5,8 juta bahkan lebih, orang-orang tersebut meninggal dunia karena mengalami atau terindikasi penyakit stroke (Pierot et al., 2018). Menurut World Health Organization (WHO) stroke merupakan alasan terjadinya kematian seseorang ke dua secara global, sekitar 11% dari total kematian (Fedesoriano, n.d.).

Stroke terjadi karena aliran darah hilang secara tiba-tiba ke area otak yang dapat mengakibatkan hilangnya fungsi neurologis (Phipps & Cronin, 2020). Biaya penanganan penyakit stroke dapat menghabiskan biaya untuk pelayanan kesehatan sebesar 2,56 triliun rupiah pada tahun 2018, dan terjadi peningkatan seseorang terindikasi penyakit tersebut yang dapat mengakibatkan biaya pelayanan kesehatan semakin meningkat (Kemenkes RI, 2018).

Terdapat beberapa riset terkait penyakit stroke ini dimana riset tersebut mengungkapkan bahwa penyakit stroke ialah salah satu penyakit yang perlu mendapatkan perhatian serius karena berdampak besar terhadap perkembangan ekonomi Negara, dan akibat jika tidak ada penanganan yang cepat dan tepat atau didiamkan jika terindikasi penyakit stroke, maka dapat memicu komplikasi, seperti demensia. Dementia atau demensia adalah penyakit yang dapat mengakibatkan penurunan daya ingat dan cara berpikir (dr. Sienny Agustin, 2021). di Indonesia tepatnya di Sulawesi utara terdapat 10,8%, Provinsi DKI

Jakarta 9,7%, dll hasil ini merupakan hasil prevelensi terjadinya stroke (Kemenkes RI, 2018), (Kementerian Kesehatan RI, 2018).

Banyak cara yang bisa dilakukan untuk dapat membantu para petugas medis dalam menemukan apakah seseorang terindikasi penyakit stroke atau tidak agar ketika pasien yang mengalami stroke ini dapat ketahui dengan cepat, salah satunya ialah dengan penggunaan Machine learning, dengan penggunaan ini terbukti mampu untuk dapat menyelesaikan topik klasifikasi, dan optimasi dalam pembuatan sebuah system penyedia layanan kesehatan (Cholissodin et al., 2017), (Shameer et al., 2018). Sebagai contoh menangani pasien yang terinfeksi penyakit jantung untuk dapat memprediksi dari data yang dihasilkan oleh industry kesehatan sehingga dapat membantu dan menyelematkan nyawa seseorang dalam jangka panjang, dan paling tidak dapat mempersingkat waktu untuk dapat mengetahui pasien terindikasi penyakit karena terbantukan dengan metode machine learning yang digunakan (Mohan et al., 2019).

Ada beberapa penelitian sebelumnya dengan study kasus yang sama yaitu prediksi penyakit stroke, dari penelitian yang ada menggunakan beberapa metode machine learning untuk dapat memprediksi seseorang terindikasi penyakit stroke yaitu random forest classifier, ann, svm, c4.5,Naïve bayes, support vector machine,dll sebagainya, hasil penelitian sebelumnya memiliki hasil akurasi yang paling baik adalah 96% dengan menggunakan teknik seperti pre-prosessing data, penentuan hyperparameter dll.

Berdasarkan hasil yang didapatkan pada proses indentifikasi masalah maka topik yang diangkat dalam penelitian ini adalah penyakit stroke

maka, data yang digunakan dalam pengolahan penelitian ini adalah stroke prediction dataset yang diambil dari kaggel, dan dataset yang dijadikan pengolahan pada proses klasifikasi ini mengalami imbalance data dimana jumlah kategori stroke dengan total 249 dan tidak stroke total 4861, imbalance kelas dapat mempengaruhi model saat klasifikasi, model hanya dapat menentukan kelas mayoritas dan kemungkinan besar kelas minoritas yang diprediksi, akan diprediksi sebagai kelas mayoritas. Dengan terjadinya imbalance pada dataset, maka diterapkan metode Synthetic Minority Over-Sampling Technique untuk dapat mengatasi dataset mempunyai masalah imbalance data, dimana kelas target memiliki jumlah yang lebih kecil dibandingkan dengan kelas target lain. (Bunkhumpornpat et al., 2012).

Penelitian ini yaitu dengan melakukan klasifikasi terhadap penyakit stroke dengan menggunakan algoritma Extreme Gradient Boosting. Algoritma tersebut merupakan algoritma dari metode boosting pada ensemble learning, yang dimana algoritma ini menggunakan prinsip dari ensemble yaitu membuat pohon yang lemah secara berurutan sehingga setiap pohon baru (atau pelajar) berfokus pada kelemahan (data yang salah diklasifikasikan) dari yang sebelumnya (Ichi.Pro, 2020). Dengan Xgboost mampu mengerjakan berbagai fungsi seperti regresi, klasifikasi, dan ranking. Kesuksesan yang pernah diraih dengan penggunaan Xgboost ini yaitu dapat dirterapkan dalam berbagai kasus pada machine learning. Xgboost awalnya dikenalkan dalam Higgs Boson Competition. Pada akhir dari kompetisi ini, metode yang paling banyak digunakan oleh sebagian besar tim dalam mengikuti kompetisi yaitu XGBoost. Juga

kompetisi yang diadakan oleh kaggle pada tahun 2015. 17 di antaranya menggunakan algoritme XGBoost, Dari 29 winning solution. 17 diantaranya menggunakan algoritme XGBoost. sedangkan sisanya mengombinasikan algoritme XGBoost dengan algoritme k-Nearest. Neighbor (k-NN) (Handayani et al., 2017).

Penelitian-penelitian sebelumnya dengan study kasus yang sama mendapatkan hasil akurasi yang paling baik yaitu 96%. Tujuan dari penelitian ini ialah meningkatkatkan hasil akurasi pada prediksi penyakit stroke menggunakan metode SMOTE dan machine learning dengan algoritma Xtreme Gradient Boosting untuk mendapatkan hasil akurasi yang lebih baik dari akurasi yang dihasilkan sebelumnya. Di dalam penelitian ini diusulkannya penggunaan algoritma Xtreme Gradient Boosting, metode SMOTE, dan Confusion matrix dimana Confusion matrix untuk dapat menilai performa dari metode yang di pakai pada proses prediksi penyakit stroke ini.

1.2. Rumusan Masalah

Berdasarkan uraian latar belakang diatas, maka dirumuskan suatu masalah yang akan dipecahkan/diselesaikan pada penelitian ini. Rumusan masalah yang diangkat sebagai berikut :

a. Apakah menggunakan metode *SMOTE* dapat mengatasi ketidakseimbangan class pada dataset penyakit stroke dan mampu meningkatkan akurasi metode?
b. Apakah penerapan machine learning dengan metode *Xgboost* untuk proses klasifikasi dapat meningkatkan akurasi?

1.3. Batasan Masalah

Batasan masalah dari penelitian ini sebagai berikut :

a. Data yang di gunakan dalam penelitian ini adalah data sekunder yang terdapat pada situs *kaggle*. Data tersebut merupakan data penyakit stroke yang dimiliki oleh ilmuwan di kaggle yaitu Fedesoriano.
b. Metode yang di gunakan pada proses klasifikasi ini adalah *Extreme Gradient Boosting* (XGBoost).
c. Metode yang digunakan untuk menangani imbalance data adalah metode *Synthetic Minority Over-Sampling Technique* (SMOTE).
d. Evaluasi metode yang digunakan Confusion Matrix.
e. Software yang digunakan dalam membantu proses klasifikasi ini adalah *Google Colabolatory* dan *Microsoft Excel*.

1.4. Tujuan Penelitian

Berikut ini merupakan tujuan dari penelitian yang dilakukan:

a. Penerapan metode machine learning dengan algoritma *Extreme Gradient Boosting* (XGBoost) pada proses klasifikasi untuk dapat meningkatkan akurasi.
b. Penerapan metode keseimbangan dataset *Synthetic Minority Over-Sampling Technique* (SMOTE) untuk mengatasi ketidakseimbangan class pada dataset dan mampu meningkatkan akurasi metode.

1.5. Manfaat Penelitian

Bagian ini memuat penjelasan tentang:

a. Dapat mengetahui hasil klasifikasi kemungkinan terindikasi penyakit stroke dari total pasien yang ada, dengan menggunakan metode machine learning yang dipakai.
b. Mengetahui penggunaan metode yang dipakai, untuk dapat menghasilkan akurasi yang lebih baik.
c. Dapat menjadikan bahan pertimbangan atau pemilihan metode yang tepat, saat pembuatan system pelayanan kesehatan, khususnya prediksi penyakit stroke.
d. Dapat mengembangkan technology dengan metode machine learning lainnya untuk mengoptimalkan prediksi dalam bidang kesehatan, tepatnya prediksi penyakit stroke.
e. Dapat di jadikan referensi pada penelitian selanjutnya terkait dengan prediksi penyakit stroke.

BAB II

TINJAUAN PUSTAKA

2.1. Tinjauan Pustaka

Beberapa penelitian mengenai prediksi penyakit stroke dengan menggunakan metode machine learning sebagai berikut :

Memprediksi stroke dengan menggunakan algoritma xgboost dilakukan oleh Rahim,A.M.A,dkk (A. Rahim M.A, et al., 2021), yang dihasilkan oleh penelitian ini adalah dengan menggunakan metode xgboost menghasilkan akurasi tertinggi adalah 96% dari lima pengujian yang dilakukan. Dataset yang digunakan dalam pengolahan ini mengalami ketidakseimbangan kelas pada dataset yang dapat memperoleh akurasi model yang kurang maksimal maka dengan itu diperlukan penerapan metode yang dapat mengatasi ketidakseimbangan dataset yang mampu meningkatkan akurasi model secara keseluruhan. Hal yang sama juga dilakukan dalam penelitian yaitu komparasi metode klasifikasi decision tree algoritma c4.5 dan random forest untuk prediksi penyakit stroke oleh Azizah, N.,dkk (Azizah.,dkk, 2021), penelitian ini menghasilkan akurasi decision tree 92.56% dan akurasi dari random forest 93.80%.

Penelitian berikutnya mengenai penanganan ketidakseimbangan dataset stroke dan mengklasifikasi kemungkinan penyakit stroke, dilakukan oleh Mutmainah,S.(Mutmainah, 2021), hasil penelitian yang dilakukan ialah, dengan penggunaan teknik random oversampling mendapatkan akurasi lebih tinggi dibandingkan dengan teknik random undersampling. Dengan teknik random oversampling mendapatkan 95% dan untuk teknik random undersampling

mendapatkan akurasi 76%. Dengan Pemakaian metode Random oversampling dan Random undersampling, dimana dengan penerapan teknik Random oversampling akan menyalin kelas minoritas untuk menyamakan kelas mayoritas namun sangat rentan terjadinya overvitting, dan Random undersampling beberapa pengamatan yang ada dibuang dimana data tersebut informasi yang berharga yang dapat digunakan untuk pembelajaran pada metode machine learning. Sama halnya dengan penelitian yang dilakukan dengan merancang machine learning terdistribusi berbasis Apache Spark untuk memprediksi penyakit stroke dilakukan oleh Ahmed.H.,dkk (Ahmed et al., 2019), dengan memakai beberapa metode machine learning diantaranya : logistic regression, random forest, decision tree, support vector machine dalam proses klasifikasi dan menggunakan metode random resample techniques untuk penanganan ketidakseimbangan data, hasil penelitian ini ialah dengan menggunakan Random Forest Classifier mendapatkan akurasi terbaik sebesar 90%. Dengan teknik penanganan data tidak seimbang yang dilakukan yaitu dengan metode Random oversampling dan Random undersampling sebagaimana Random undersampling beberapa pengamatan yang ada dibuang dimana data tersebut informasi yang berharga yang dapat digunakan untuk pembelajaran pada metode machine learning dan teknik Random oversampling akan menyalin kelas minoritas untuk menyamakan kelas mayoritas namun sangat rentan terjadinya overvitting. Dan juga sama halnya dengan penelitian Menganalisis performa prediksi stroke Menggunakan Algoritma Klasifikasi ML dilakukan oleh Sailasya.G & Aruna Kumari L.G. (Sailasya & Kumari, 2021) metode yang digunakan diantaranya : Logistics Regression (Lr),

Decision Tree Classifier (Dtc), Naive Bayes Algorithm, Random Forest Algorithm dll. Hasil dari penelitian ini adalah Dari semua algoritma yang dipilih, Klasifikasi dengan algoritma Naïve Bayes mempunyai performa terbaik dengan akurasi 82%. Dengan penggunaan dataset kaggle yaitu stroke prediction dataset, dataset yang digunakan mengalami ketidakseimbangan dataset dan dalam penelitian yang di lakukan menerapkan metode random undersampling.

Penelitian selanjutnya melakukan perbandingan metode fuzzy k-nearest neighbor dan neighbor weighted k-nearest neighbor dalam memprediksi stroke yang dilakukan oleh Nugroho,S.A., (Nugroho, 2020). hasil dari penelitian ini adalah hasil uji data seimbang mendapatkat akurasi 81.272% dan 81.814% dari metode Fuzzy KNearest Neighbor dan Neighbor Weighted KNearest Neighbor dan data yang tidak seimbang 82.45% dan 82.75%. pada proses pre-prosessing data tidak melakukan penghapusan data atau menggantikan dengan rata-rata ataupun dengan cara lainnya pada data yang bernilai kosong dalam sebuah atribut pada dataset yang dimiliki, sehingga nantinya Output yang dihasilkan tidak bias.

Terdapat banyak metode untuk melakukan klasifikasi dan keseimbangan kelas pada dataset. Metode yang dapat diterapkan untuk proses klasifikasi ini adalah Support Vector Machine, Decision Tree, Naïve Bayes, dll begitu juga dengan metode balancing dataset seperti Random Oversampling, Random Undersampling, Synthetic Minority Over-Sampling Technique (SMOTE), dll sebagainya. Pada penelitian kali ini penulis menggunakan metode Synthetic Minority Over-Sampling Technique (SMOTE) untuk mengatasi ketidakseimbangan data, dan menggunakan algoritma Xtreme Gradient Boosting

untuk proses klasifikasi pada dataset stroke prediction. Penelitian ini akan meningkatkan peforma akurasi model dari metode Xtreme Gradient Boosting pada prediksi kemungkinan penyakit stroke.

2.2. Keaslian Penelitian

Tabel 2. 1 Matriks literatur review dan posisi penelitian

Prediksi Stroke Menggunakan Metode Smote Dan Xtreme Gradient Boosting

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
1.	Stroke Prediction using Distributed Machine Learning Based on Apache Spark	Hager Ahmed, Sara F. Abd-el ghany, Eman M.G.Youn, Nahla F.Omran, Abdelmgeid A.Ali, International Journal of Advanced Science and Technology,2019	Perancangan machine learning terdistribusi berbasis Apache Spark untuk memprediksi penyakit stroke.	Terdiri dari lima tahap, yaitu memuat dataset stoke, pra-pemrosesan data, validasi silang dan penyetelan hyperparameter, pengklasifikasi, dan mengevaluasi pengklasifikasi. Hasil penelitian menunjukkan bahwa random forest classifier mencapai hasil akurasi terbaik sebesar 90%.	Random oversampling sebagai balancing dataset, dengan teknik Random oversampling sangat rentan terjadinya overvitting.	Dalam mengatasi balancing dataset penelitian ini menggunakan dengan teknik Random oversampling untuk mengatasi ketidakseimbangan dataset dengan cara mengambil secara acak pada kelas minoritas untuk menyamakan kelas mayoritas, sedangkan penelitian yang dilakukan menerapkan metode Synthetic Minority Over-Sampling Technique dalam mengatasi imbalance kelas dataset yaitu dengan menambah kelas minoritas agar sama dengan kelas mayoritas dengan cara menambahkan data buatan, sintesis.

Tabel 2. 1 Matriks literatur review dan posisi penelitian

Prediksi Stroke Menggunakan Metode Smote Dan Xtreme Gradient Boosting (lanjutan)

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
2.	Analyzing the Performance of Stroke Prediction using ML Classification Algorithms	Gangavarapu Sailasya, Gorli L Aruna Kumari, (IJACSA) International Journal of Advanced Computer Science and Applications. 2021	Membangun model pembelajaran mesin dapat membantu dalam prediksi awal stroke dan mengurangi dampak parah di masa depan.	Makalah ini menunjukkan kinerja berbagai algoritma pembelajaran mesin dalam memprediksi stroke dengan sukses berdasarkan beberapa atribut fisiologis. Dari semua algoritma yang dipilih, Klasifikasi Naïve Bayes memiliki performa terbaik dengan akurasi 82%.	Saran dari penelitian yang dilakukan yaitu mencoba dengan menerapkan beberapa metode balancing dataset, agar mendapatkan hasil yang maksimal terhadap model yang dibangun dalam melakukan klasifikasi.	Pada penelitian ini menggunakan metode undersampling untuk mengatasi ketidakseimbangan dataset yang terjadi dimana dengan penerapan metode tersebut menghapus beberapa data mayoritas untuk menyeimbangi kelas mayoritas tentunya proses klasifikasi tersebut kehilangan beberapa data yang berharga untuk dijadikan data latih pada algoritma machine learning. Sedangkan pada penelitian yang dilakukan menggunakan metode smote untuk mengatasi imbalance dataset.

Tabel 2. 1 Matriks literatur review dan posisi penelitian

Prediksi Stroke Menggunakan Metode Smote Dan Xtreme Gradient Boosting (lanjutan)

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
3.	Stroke Prediction Using Machine Learning Method with Extreme Gradient Boosting Algorithm	Abd Mizwar A. Rahim, Andi Sunyoto, Muhammad Rudyanto Arief, Matrik : Jurnal Manajemen, Teknik Informatika, dan Rekayasa Komputer,2022	Meningkatkan akurasi pada kasus prediksi penyakit stroke,dan mendapatkan akurasi yang lebih baik dari penelitian yang pernah dilakukan sebelumnya.	Hasil akurasi yang didapatkan ini, mendapatkan hasil lebih baik dari pada penelitian yang dilakukan sebelumnya dengan menggunakan pola dataset yang sama, dataset tersebut adalah Stroke Prediction Dataset. penelitian menunjukan bahwa klasifikasi menggunakan XGBoost (Xtreme Gradient Boosting) mencapai hasil akurasi terbaik sebesar 96%.	Dalam dataset stroke prediksi ini terdapat ketidakseimbangan kelas pada dataset, dan dalam penelitian yang dilakukan metode penyeimbangan kelas pada dataset tidak di gunakan.	Dalam penelitian ini belum adanya penerapan metode untuk dapat menangani imbalance kelas pada dataset, tentunya dengan belum adanya penerapan metode untuk mengatasi permasalahan tersebut maka akan dapat mempengaruhi hasil klasifikasi yaitu hasil yang didapatkan menjadi bias dll, artinya model klasifikasi akan tidak dapat memperoleh hasil yang baik ketika mengklasikasikan kelas minoritas. Dalam penilitian proposal penelitian ini terdapat metode Synthetic Minority Over-Sampling Technique untuk dapat mengatasi imbalance dataset agar model dapat mengklasikasikan kelas minoritas dengan baik dan juga kelas mayoritas.

Tabel 2. 1 Matriks literatur review dan posisi penelitian

Prediksi Stroke Menggunakan Metode Smote Dan Xtreme Gradient Boosting (lanjutan)

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
4.	Exploratory Data Analysis and Machine Learning Algorithms to Classifying Stroke Disease	Prismahardi Aji Riyantoko, Tresna Maulana Fahrudin, Kartika Maulida Hindrayani, Mohammad Idhom, ijconsist, 2021(Riyantoko, 2021)	Menggunakan beberapa metode machine learning dalam mengklasifikasikan data penyakit stroke, dan membandingkan metode machine learning yang mendapatkan hasil akurasi terbaik.	Algoritma machine learning untuk klasifikasi menunjukkan bahwa regresi logistik dan stochastic gradient descent memberikan akurasi tertinggi sebesar 94,61%. Hampir semua model dapat mencapai lebih dari 90%, namun Naïve Bayes hanya memiliki akurasi sekitar 20,74%.	Menggambarkan pola data dari dataset yang digunakan dalam bentuk grap, melakukan pergantian terhadap missing value yang ada pada dataset dengan rata-rata guna tidak menghilangkan informasi yang ada, dan menerapkan metode balancing dataset	Prismahardi Aji Riyantoko,dkk melakukan penghapusan pada missing value yang ada pada dataset. Sedangkan pada penelitian ini dalam melakukan pre-prosessing data menggantikan missing value dengan rata-rata yang berhubungan.

Tabel 2. 1 Matriks literatur review dan posisi penelitian

Prediksi Stroke Menggunakan Metode Smote Dan Xtreme Gradient Boosting (lanjutan)

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
5.	Penanganan Imbalance Data Pada Klasifikasi Kemungkinan Penyakit Stroke	Siti Mutmainah, https://journal.uii.ac.id/ ,2021	Menagani ketidakseimbangan data, dan klasifikasi penyakit stroke	hasil penelitian yang dilakukan ialah, dengan penggunaan teknik Random Oversampling mendapatkan akurasi lebih tinggi dibandingkan dengan teknik random undersampling. Dengan teknik RandomOversampling mendapatkan 95% dan untuk teknik Random Undersampling mendapatkan akurasi 76%.	Pemakaian metode Random oversampling dan Random undersampling dimana dengan teknik Random oversampling akan menyalin kelas minoritas untuk menyamakan kelas mayoritas namun sangat rentan terjadinya overvitting, dan Random undersampling beberapa pengamatan yang ada dibuang dimana data tersebut informasi yang berharga.	Penelitian ini memakai teknik Random undersampling dan teknik Random oversampling untuk mengatasi imbalance kelas pada dataset terdapat beberapa hasil yang perlu diperhatikan yaitu kehilangan data pada dataset, terjadinya overfitting, dan high variance. sendangkan penelitian yang dilakukan menggunakan Metode Synthetic Minority Over-Sampling Technique untuk mengatasi imbalance dataset dengan cara menambahkan data kelas minoritas dengan cara menambahkan data buatan atau sintesis tersebut di buat berdasarkan *k-tetangga* terdekat.

Tabel 2. 1 Matriks literatur review dan posisi penelitian

Prediksi Stroke Menggunakan Metode Smote Dan Xtreme Gradient Boosting (lanjutan)

No	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
6.	Perbandingan Metode Fuzzy K-Nearest Neighbor Dan Neighbor Weighted K-Nearest Neighbor Untuk Deteksi Penyakit Stroke	Syamsul Aji Nugroho, http://eprints.uty.ac.id/4840/ , 2020	Mengetahui metode yang terbaik yaitu dengan metode Fuzzy K-Nearest Neighbor Dan Neighbor Weighted K-Nearest Neighbor untuk mendeteksi penyakit stroke.	Dengan menggunakan metode Fuzzy K-Nearest Neighbor dan Neighbor Weighted K- Nearest Neighbor secara berurutan rata-rata akurasi akurasi 81.272% dan 81.814% untuk data seimbang, sedangkan untuk hasil pengujian data uji tidak seimbang dengan beragam rasio jumlah kelas di dapatkan hasil secata berturut-turut 82.45% dan 82.75%.	Tidak melakukan penghapusan data yang bernai kosong dalam sebuah dataset yang dimiliki, tetapi menggantikan dengan rata-rata atau apapun yang berhubungan, sehingga data tersebut tidak hilang dari proses klasifikasi.	Penelitian yang dilakukan oleh Syamsul Aji Nugroho, melakukan penghapusan data yang bernilai kosong dalam dataset,dan melakukan perbandingan metode Fuzzy K-Nearest Neighbor Dan Neighbor Weighted K-Nearest Neighbor untuk mendeteksi penyakit stroke. Sedangkan penelitian yang dilakukan dalam proposal ini adalah menggantikan data yang bernilai kosong dengan rata-rata dari data yang bernilai kosong tersebut, dan untuk proses klasifikasi menggunakan Metode Synthetic Minority Over-Sampling Technique Dan Xtreme Gradient Boosting

Pada Table 2.1, dipaparkan matriks literatur review dan posisi penelitian. Dengan topik penelitian yang dilakukan oleh beberapa peneliti sebelumnya yaitu memprediksi penyakit stroke dengan penggunaan beberapa metode machine learning yang berbeda-beda (Svm,Decicion three,naïve bayes,dll), begitu juga dengan penanganan imbalance data dengan penerapan beberapa metode untuk dapat mengatasinya seperti Random oversampling, dan Random undersampling yang dilakukan oleh Mutmainah,S.(Mutmainah, 2021), Ahmed,H.,dkk (Ahmed et al., 2019),dan Sailasya.G & Aruna Kumari L.G. (Sailasya & Kumari, 2021) dengan penggunaan kedua metode tersebut dapat menyebabkan terjadinya overfitting dan juga kehilangan beberapa informasi dalam dataset. Adapun penelitian yang belum mengatasi balancing kelas pada dataset, dengan ini tentunya dapat mempengaruhi model saat klasifikasi, model hanya dapat menentukan kelas mayoritas dan kemungkinan besar kelas minoritas yang diprediksi, akan diprediksi sebagai kelas mayoritas, ini yang pernah dilakukan oleh Rahim,A.M.A,dkk (A. Rahim M.A, et al., 2021). Berikutnya pada tahap preprosessing data yang dilakukan oleh Nugroho,S.A., (Nugroho, 2020) pada proses pre-prosessing data tidak melakukan penghapusan data atau menggantikan dengan rata-rata ataupun dengan cara lainnya pada data yang bernilai kosong pada atribut di dataset yang dimiliki, sehingga nantinya Output yang dihasilkan dari klasifikasi penyakit stroke ini tidak bias. Adapun melakukan preprosessing data dengan melakukan penghapusan data pada data bernilai kosong di sebuah atribut tentunya dengan ini akan terjadi kehilangan informasi sebanyak 202 data pada dataset, teknik tersebut dilakukan oleh (Riyantoko, 2021).

Pada penelitian ini, penulis akan menyempurnakan penelitian yang sedang dilakukan, berdasarkan kekurangan-kekurangan dari penelitian sebelumnya yaitu memperbaiki sebelum masuk dalam tahap pengujian dengan mengatur pada tahapan preprosessing data dengan menggantikan data yang bernilai kosong agar data tersebut tidak dihapus agar digunakan untuk pembelajaran metode machine learning, juga metode yang digunakan berupa penyetelan parameter, dll hingga mendapatkan peforma akurasi model yang digunakan pada proses klasifikasi penyakit stroke ini menjadi lebih baik yaitu penelitian mengenai Prediksi Stroke Menggunakan Metode Synthetic Minority Over-Sampling Technique Dan Xtreme Gradient Boosting.

2.3. Landasan Teori

2.3.1 Penyakit Stroke

Stroke merupakan suatu penyakit atau gangguan fungsional otak berupa lumpuhnya saraf akibat hambatan yang terjadi yaitu aliran darah ke otak, (Junaidi, 2011). Stroke secara garis besar mempunyai dua jenis *ischemic stroke* atau *Non hemorrhagic* (penyumbatan) dan *hemorrhagic stroke* (pendarahan), kedua jenis ini dapat mengakibatkan lumpuhnya organ tubuh secara menyeluruh atau sebagian, hingga menyebabkan kematian (Aji Seto Arifianto, Moechammad Sarosa, 2014). Seseorang yang menderita diabetes melitus, hipertensi, dan seseorang yang memiliki riwayat kedua penyakit tersebut biasanya terjadi pada kasus *Ischemic stroke* atau *Stroke Non Hemoragik* (Permatasari, 2020).

Beberapa faktor dan gejala secara umum dalam kehidupan sehari hari yang dapat menyebabkan terjadinya penyakit stroke (SAPUTRI, 2021):

a. Merokok

Dengan kebiasaan merokok dapat mengakibatkan penyumbatan pembuluh darah sehingga dapat mengganggu aliran darah yang membahayakan kesehatan tubuh seseorang. Dengan merokok biasanya dilakukan oleh laki-laki dibanding dengan perempuan, sehingga penyakit stroke lebih berpeluang pada laki-laki.

b. Hipertensi

Kerusakan pada otak karena tekanan yang terjadi pada *intra kranial* meningkat dan *embolus* yang terlepas dari pembuluh non otak dapat menyebabkan hipertensi, pencegahan penyakit hypertensi dapat melakukan pola hidup sehat dengan mengurangi makanan berlemak,serta olahraga yang teratur. Jika tidak mengendalikan hipertensi maka berpeluang untuk terkena stroke.

c. Diabetes Melitus

Peningkatan kadar gula darah dalam tubuh yang disebabkan oleh konsumsi makanan yang memiliki kandungan gula darahnya berlebihan.

d. Usia

Berdasarkan data kesehatan bahwa usia sekitar 15-64 tahun, usia tersebut sangat berpotensi terserang stroke.

e. Jenis Kelamin

Kebanyakan jenis kelamin pria menderita penyakit stroke, karena dengan pola gaya hidup merokok yang dapat meyumbat pembuluh darah pada tubuh.

2.4 Data Mining

Data mining menurut Garner Group adalah sebuah proses analisis menggunakan teknik matematika atau statistic pada data yang diolah sehingga dapat menemukan hubungan baru atau informasi baru yang berarti bagi pengguna (Mardi, 2017). Dan data mining meliputi kegiatan yang meliputi pengumpulan, pemakaian data dari objek yang dianalisis guna menemukan pola,dan hubungan pada data yang berukuran besar (Saleh, 2015). Dari data yang di analisis yang dapat menemukan informasi keputusan masa depan bagi instansi tertentu sehingga informasi tersebut dapat digunakan pada proses strategi pada objek kedepan (S Mujiasih, 2011).

Dalam mengaplikasikan data mining adapun proses Knowledge Discovery in Database (KDD) yang dilakukan agar pengolahan data menghasil data yang berkualitas.

Adapun proses KDD sebagai berikut (Angga Ginanjar Mabrur, 2012):

a. *Data Selection*

   Pemilihan data pada sekumpulan dataset yang digunakan.

b. *Preprocessing*

   Sebelum masuk pada pengolahan data menggunakan data mining maka dilakukannya proses cleaning data,pemeriksaan data yang tidak konsisten,dan memperbaiki kesalahan yang terdapat pada data.

c. *Transformation*

   Yaitu proses pemindahan data yang telah dipilih atau yang sudah melewati proses sebelumnya pada proses coding untuk dapat masuk pada tahap selanjutnya yaitu pada tahap data mining.

d. *Data mining*

Proses mencari pola pada data yang dipakai dengan menggunakan teknik atau metode tertentu.

e. *Interpretation / Evaluation*

Tahapan ini merupakan pemeriksaan dari pola yang di hasilkan dengan fakta pada data yang ada. Apakah yang dihasilkan bertantangan yang ada sebelumnya atau tidak.

2.5 *Synthetic Minority Over-Sampling Technique (SMOTE)*

Synthetic Minority Over-Sampling Technique (SMOTE) merupakan salah satu turunan dari oversampling, metode SMOTE ini pertama kali diperkenalkan oleh Nithes V. Chawla untuk dapat mengatasi ketidakseimbangan kelas dari suatu data *(Kovács et al., 2020)*. Metode SMOTE menambah kelas minoritas agar sama dengan kelas mayoritas dengan cara menambahkan data buatan, data buatan atau sintesis tersebut di buat berdasarkan *k-tetangga* terdekat (Sofyan,S., 2013).

Cara kerja dari *SMOTE* ini ialah pertama mengambil selisih antara vector dari fitur kelas minoritas dan nilai dari Nearest Neighbor dari kelas minoritas lalu kalikan nilai tersebut dengan angka acak antara 0 sampai 1, Berikutnya hasil dari penjumlahan tersebut di tambahkan dengan vector lainnya sehingga mendapatkan hasil baru dari vector, sebagaimana didefinisikan kedalam persamaan berikut (Kasanah et al., 2019):

$$X_{sya} = X_i + (\boldsymbol{x}_{knn} - \boldsymbol{x}_i) \times \delta \qquad (2.1)$$

Keterangan :

X_{sya} = data synthesis yang akan diciptakan

X_i = data yang akan di replikasi

$\boldsymbol{x}_{knn}$ = data yang memiliki jarak dari data X_i

δ = angka acak antara 0 sampai 1

2.6 Machine Learning

Machine learning adalah teknologi yang diciptakan untuk mampu meniru kecerdasan manusia, yang mengacu dari berbagai ilmu seperti kecerdasan buatan, statistik, ilmu komputer,teori informasi, psikologi, teori kontrol, dan filsafat. Ternologi yang diciptakan tersebut dapat diterapkan di berbagai bidang seperti visi computer, teknik pesawat ruang angkasa, keuangan, hiburan, ekologi, biologi komputasi,dan aplikasi biomedis dan medis (Naqa & Murphy, 2015). Machine learning juga berhubungan robotika dan kecerdasan otomatisasi, machine learning juga menyediakan berbagai alat dan metode untuk dapat menganalisis dan menarik kesimpulan dari data yang di analisis tersebut (Statistika et al., 2020).

Terdapat tiga cabang utama dalam machine learning diantaranya (Royal Society of Great Britain, 2017):

a. *Suvervised machine learning*

   Suatu system yang dilatih sesuai dengan data yang telah diberi label dari awal. Kemudian dari label tersebut membagikan setiap titik data dalam satu atau beberapa kelompok. Dari data yang ada,lalu system mempelajari bagaimana data tersebut dipelajari atau dikenal sebagai data training terstruktur, dan dari data training itulah system memprediksi atau mengklasifikasikan data *test* atau data *uji*.

b. *Unsupervised learning*

   Kebalikan dari pembelajaran sebelumnya adalah *Suvervised machine learning*, dimana *Unsupervised learning* ini adalah pembelajaran tanpa pengawasan dengan artinya analisis yang dilakukan pada data yang

diangkat dimana data tersebut tanpa label, ini bertujuan untuk mendeteksi karakteristik dari suatu titik data yang kurang lebih serupa antara satu sama lain, misalnya meng-*cluster* data dan menetapkan data dari *cluster* tersebut.

c. *Reinforcement learning*

Dimana pembelajaran yang memperkuat belajar berdasarkan dengan pengalaman, yang berada dalam pembelajaran *supervised* dan *unsupervised*.

2.7 Klasifikasi

Klasifikasi adalah suatu proses untuk memasukan objek data atau mengkategorikan sebuah data kedalam kelas tertentu berdasarkan ketersediaan kelas yang ada pada dataset. Klasifikasi membangun sebuah model berdasarkan pada data latih yang diangkat, lalu data tersebut di klasifikasikan berdasarkan rancangan model yang telah dibuat (Utomo & Mesran, 2020). Dalam klasifikasi terdapat sebuah varibel target atau kelas target, dimana sebuah data yang di ujikan akan di tentukan bahwa data tersebut akan masuk pada kategori kelas mana lewat proses klasifikasi tersebut (Septian, 2009).

Adapun langkah-langkah dalam proses klasifikasi yaitu (Kaur et al., 2015):

a. Membangun model dari data training yang ada, dan data tersebut sudah mempunyai label yang sudah diketahui sebelumnya. algoritma klasifikasi diterapkan untuk membuat model berdasarkan data *training*.

b. Melakukan evaluasi terhadap model yang di hasilkan, untuk mengetahui seberapa baik kinerja dari metode yang dipakai pada model tersebut.

2.8 Boosting

Pada tahun 1998 Robert E. Schapire memperkenalkan *Boosting* yang merupakan salah satu metode *ensemble* dari klasifikasi yang meningkatkan beberapa klasifikasi lemah menjadi klasifikasi yang kuat. Teknik dari metode ini dilihat dari metode rata-rata yang dibangun untuk metode klasifikasi tetapi dapat juga diimplementasikan ke metode regresi (Syarif et al., 2012).

Sama halnya dengan metode *bagging* yang memanfaatkan voting pada tujuan perhitungan rata-rata numerik dalam *output* model individu tunggal. Persamaan berikutnya yaitu menggabungkan model yang mempunyai jenis yang sama, seperti *decision tree*. *Bagging* memakai model individu yang dirancang secara terpisah, untuk *boosting* menggunakan model yang dibangun sebelumnya untuk mempengaruhi model baru (Yaman & Subasi, 2019).

Langkah-langkah teknik *gradient boosting*, di awali dengan memuat model kedalam data yang didefinisikan kedalam persamaan berikut:

$$F1(x) = y \tag{2.1}$$

Berikutnya menghitung *residual* pada proses sebelumnya, perhitungan residual tersebut didefinisikan kedalam persamaan berikut:

$$h1(x) = \mathrm{y} - \mathrm{F1(x)} \tag{2.2}$$

Selanjutnya membuat model yang baru dengan persamaan:

$$F2(x) = \mathrm{F1(x)} - \mathrm{h1(x)} \tag{2.3}$$

Dari proses diatas mendapatkan akhir model yang dibuat dari kumpulan-kumpulan model sebanyak n iterasi sampai menghasilkan *error* terkecil dari *residual*. Final dari model pada metode *boosting* didefinisikan dengan persamaan berikut:

$$F(x) = F_1(x) \rightarrow F_2(x) = F_1(x) + h_1(x) \ldots \rightarrow F_M(x) = F_{M-1}(x) + h_{M-1}(x) \quad (2.3)$$

Akhir model yang didapatkan untuk metode *boosting* didefinisikan dengan persamaan berikut:

$$f(x) = \sum_{M-1}^{M} fm(x) \quad (2.4)$$

Atau didefinisikan dengan persamaan berikut:

$$f(x) = y0 \sum_{M-1}^{M} ym\, hm(x) \quad (2.5)$$

Dimana $f(x) = y0$ dan $fm(x) = ym\, hm(x)$ untuk m = 1,2,3 ……., M dengan nilai $hm(x) \in \{-1,1\}$. $\gamma_m(x)$ merupakan klasifikasi lemah sedangkan γ_m adalah bobot pada setiap klasifikasi (Syahrani et al., 2019).

2.9 *Extreme Gradient Boosting*

Extreme Gradient Boosting (XGBoost) sebuah teknik dalam *machine learning* untuk analisa regresi dan klasifikasi berdasarkan *Gradient Boosting Decision Tree* (GBDT). Friedman, 2019 pertama kali memperkenalkan Metode *XGBoost*, dalam penelitian yang di lakukan yaitu membangun Gradient Boosting Machine (GBM) yang menghubungkan antara boosting dan optimasi. Memprediksi error dari model yang di bangun sebelumya dilakukan dalam membangun model baru yang digunakan dalam metode boosting. Model baru yang ditambahkan melakukan perbaikan hingga tidak ada lagi perbaikan dari *error* yang dilakukan (Mo et al., 2019).

Dari langkah t yang diprediksi diumpamakan $\hat{y}^{(t)}$ dengan :

$$\hat{y}^{(t)} = \sum_{k=0}^{n} fk(x)$$

dengan $fk(x)$ menggambarkan model pohon. dan untuk yi didapatkan dengan perhitungan berikut:

$$\hat{y}i = 0$$
$$\hat{y}^{(1)} = f1(x) = \hat{y}^{(0)} + f1(x1)$$
$$\hat{y}^{(2)} = f1(x) + f2(x) = \hat{y}^{(1)} + f2(x2)$$
$$\vdots$$
$$\hat{y}^{(t)} = \hat{y}i^{(t-1)} + ft(xi)$$
$$\hat{y}i^{(t)} = \sum_{k=1}^{t} fk(xi)$$

dimana :

- $\hat{y}i^{(t)}$ = *Final tree* model
- $ft(xi)$ = Model baru yang dibangun
- $\hat{y}i^{(t-1)}$ = Model pohon yang dihasilkan sebelumnya
- t = Jumlah total model dari *base tree models*

Jumlah pohon dan dept merupakan hal yang penting dalam penentuan dari algoritma Extreme Gradient Boosting. Pencarian klasifikasi baru yang dapat mengurangi *loss function*, merupakan sebuah permasalahan dalam menemukan algoritma yang optimum, dengan target fungsi kerugian pada persamaan (2.6) berikut:

$$Obj^{(t)} = \sum_{i=1}^{t} l(yi, \hat{y}^{(t)}) + \sum_{i=1}^{t} \Omega(fi) \qquad (2.6)$$

Dimana :

- $\hat{y}i^{(t)}$ = Nilai prediksi
- yi = Nilai Aktual

- $l(\hat{y}^{(t)}, yi)$ = *loss function*
- $\Omega(fi)$ = istilah regularisasi

Pada persamaan (2.6) merupakan sebuah fungsi yang parameter dan tidak dapat dioptimalkan pada metode pengoptimalan tradisional di ruang *Euclidean*. Sehingga model latih digantikan dengan cara aditif, $\hat{y}i^{(t)}$ digunakan pada prediksi ke-i dan iterasi ke-t (Agarwal et al., 1994). Menambahkan f_i dalam meminimalkan *loss function* sehingga persamaan (2.7) sebagai berikut :

$$Obj^{(t)} = \sum_{i=1}^{t} l(yi, \hat{y}^{(i-1)} + (f_t(x_i)) + \Omega(fi) + Constant \quad (2.7)$$

Target akhir dari *loss function* diubah menjadi persamaan (2.8), kemudian dilatih dengan target *loss function* berikut:

$$Obj^{(t)} = \sum_{i=1}^{t} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(ft) \quad (2.8)$$

Dimana :

- $g_i = \partial_{\hat{y}i}{}^{(t-1)} l(yi, \hat{y}i^{(t-1)})$
- $h_i = \partial^2_{\hat{y}i\ (t-1)} l(yi, \hat{y}i^{(t-1)})$

urutan pertama dan kedua di statistik gradient pada loss function ialah g_i dan h_i

istilah regularisasi $\Omega(f_i)$ dihitung dengan menggunakan persamaan (2.9) perhitungan tersebut digunakan untuk mengurangi kompleksitas model dan dapat meningkatkan kegunaan pada dataset lainnya.

$$\Omega(f_i) = \gamma T + \frac{1}{2}\lambda\|\omega\|^2 \quad (2.9)$$

Dimana :

- T = Jumlah *leaf*
- λ dan γ = Koefisien, dengan nilai *default* ditetapkan untuk λ=1 dan γ=0
- ω = Bobot *Leaf*

2.10 *hyperparameter*

Nilai parameter diatur guna mendapatkan model yang optimal, pengaturan dari parameter ini terdapat pada beberapa metode machine learning biasanya disebut *hyperparameter* (Putatunda & Rama, 2018). *Hyperparameter* dipakai untuk mengatur macam-macam aspek dalam meachine learning yang dapat mempengaruhi performa dan model yang dihasilkan. Percarian *hyperparameter* dilakukan secara manual atau dengan menguji kumpulan *hyperparameter* (Claesen & De Moor, 2015).

Tabel 2.2 Parameter pada Metode *XGBoost*

Sumber: (Xgboost.readthedocs.io)

Parameter	Keterangan
n_estimators	Jumlah pohon pada *tree*
learning_rate	Penyusutan ukuran yang digunakan untuk mencegah overfitting
gamma	Pengurangan kerugian minimum, Semakin besar gamma, semakin konservatif algoritmanya.
max_depth	Kedalaman maksimum pohon.
min_child_weight	Jumlah bobot minimum *child node*
subsample	Pengambilan sampel secara acak dari data pelatihan sebelum menanam pohon
base_score	Skor awal semua instance

2.11 Evaluasi Metode

Setelah proses analisis dilakukan dan mendapatkan hasil prediksi. Berikutnya melakukan penilaian nilai prediksi yang paling baik. Secara umum pengukuran kinerja metode klasifikasi dilakukan yaitu dengan membandingkan antara prediksi yang dihasilkan dengan variable data *testing* sebagai data sebenarnya.

2.11.1 *Confusion Matrix*

Confusion matrix memberikan perincian terkait kesalahan pada hasil klasifikasi dengan metode yang di gunakan. Confusion matrix adalah tabel yang berisikan perhitungan yang didasari pada evaluasi model klasifikasi berdasarkan jumlah study kasus yang di klasikasikan yang diprediksi benar dan salah (Gorunescu, 2011).

Tabel 2. 3 Confusion Matrix
Sumber: (Doreswamy & Hemanth, 2011)

Classification	*Predicted Positive*	*Predicted Negative*
Actual Positive	*True Positive (TP)*	*False Negative (FN)*
Actual Negative	*False Positive (FP)*	*True Negative (TN)*

Dalam pengukuruan kinerja menggunakan *confusion matrix* terdapat empat bagian untuk mengidentifikasi suatu prediksi, berikut diantaranya (Doreswamy & Hemanth, 2011) :

a. TP (True Positive) adalah jumlah data dengan nilai aktual positif dan nilai prediksi positif
b. TN (True Negative) adalah jumlah data dengan nilai actual positif dan nilai prediksi negatif

c. FP (False Positive) adalah jumlah data dengan nilai actual negatif dan nilai prediksi positif

d. FN (False Negative) adalah jumlah data dengan nilai actual negatif dan nilai prediksi negatif

Terdapat beberapa nilai evaluasi yang sering di pakai pada klasifikasi biner. Dapat dilihat berdasarkan nilai *confusion matrix* (Sokolova & Lapalme, 2009):

a. *Accuracy* (ACC) adalah efektivitas dari hasil yang didapatkan dalam proses klasifikasi

$$Accuracy\ (\%) = \frac{(TP+TN)}{(TP+TN+FP+FN)} \quad (2.6)$$

b. *Precision* (PREC) adalah persentase dari label data dengan label positif yang dihasil dari proses klasifikasi.

$$Precision\ (\%) = \frac{(TP)}{(TP+FP)} \quad (2.7)$$

c. *Recall* (REC) atau *sensitivity* adalah efektivitas dari pengklasifikasi dalam mengidentifikasi label positif

$$Recall\ (\%) = \frac{(TP)}{(TP+FN)} \quad (2.8)$$

d. *F1-Score* adalah perbandingan rata-rata presisi dan recall yang dibobotkan

$$F1\text{-}Score = 2 \times \frac{(\text{Precision} \times \text{Recall})}{(\text{Precision} + \text{Recall})} \quad (2.9)$$

BAB III

METODE PENELITIAN

3.1. Jenis, Sifat, dan Pendekatan Penelitian

Jenis penelitian yang digunakan yaitu dengan melakukan eksperimen terhadap dataset untuk proses klasifikasi kemungkinan terindikasi penyakit stroke. Dengan varibel independen adalah id, gender, age, hypertension, heart disease, ever married, work type, Residence type, avg glucose_level, body mass index, smoking status. dan untuk variabel dependent ialah stroke. Proses experiment yang dilakukan adalah mengklasifikasikan data dari stroke prediction dengan menggunakan metode SMOTE dan machine learning dengan algoritma Xtreme Gradient Boosting.

Pada penelitian ini dataset yang digunakan adalah stroke prediction dari kaggle (Fedesoriano, n.d.). setelah melakukan download dataset akan dilakukan pre-prosessing pada dataset yaitu LabelEncoder dapat mengubah fitur bukan numerik menjadi fitur numerik, Replace missing values menggantikan atribut yang bernilai kosong dengan rata-rata bmi mengalami stroke dan bmi tidak mengalami stroke, dan melakukan drop terhadap variabel id pada dataset karena variabel tersebut tidak berhubungan dengan varibel independen.

Langkah selanjutnya yaitu membagikan dataset menjadi data training dan data testing pembagian pada penelitian ini dibagikan data training dan data testing menjadi 70/30, setelah dilakukan split data selanjutnya melakukan balancing dataset dengan metode Synthetic Minority Over-Sampling Technique sehingga

jumlah kelas minoritas akan seimbang dengan kelas mayoritas, dengan ini maka jumlah dari dataset bertambah.

Selanjutnya melakukan Klasifikasi menggunakan Algortima Xtreme Gradient Boosting, tahapan ini merupakan tahapan yang di kerjakan oleh metode Xgboost. langkah utama yang dilakukan adalah tuning parameter yang mengatur beberapa parameter pengujian untuk mendapatkan hasil yang optimal dari proses klasifikasi ini. tuning parameter yang diujikan yaitu max_depth, learning_rate, min_child_weight, n_estimator, random_state dll,sebagainya.

Setelah proses klasfikasi dengan metode yang digunakan selesai, kemudian dilakukan proses evaluasi, untuk menilai peforma dari metode yang digunakan pada proses klasifikasi ini. Metode dengan penyetelan parameter yang mendapatkan peforma yang paling bagus akan dipakai dalam penelitian ini.

3.2. Metode Pengumpulan Data

Dataset yang digunakan pada penelitian ini diambil dari kaggle yaitu Stroke Prediction Dataset. Dataset ini merupakan dataset publik digunakan untuk memprediksi kemungkinan pasien terkena stroke berdasarkan parameter input seperti jenis kelamin, usia, berbagai penyakit, dan status merokok. Setiap baris dalam data memberikan informasi yang relevan tentang pasien.

Dataset stroke prediction ini mempunyai 12 atribut. Dengan varibel independen adalah id, gender, age, hypertension, heart disease, ever married, work type, Residence type, avg glucose_level, body mass index, smoking status. dan untuk variabel dependent ialah stroke. kategori pasien pada variabel dependent berjumlah 2 kategori yaitu pasien terindikasi stroke dan tidak.

Tabel 3.1 Dataset menggambarkan informasi fitur pada dataset yang digunakan.

Tabel 3.1 Dataset.

No	Fitur	Keterangan
1	Gender	Female Male
2	Age	Age
3	Hypertensi	hypertension
4	heart_disease	1 Have heart disease 0 does not have heart disease
5	ever_married	1 means Married 0 means Not married
6	work_type	Children Personal Never work government work entrepreneur
7	Residence_type	Rural Urban
8	avg_glucose_level	Average glucose level
9	bmi	body mass index
10	smoking_status	Never smoked Used to smoke

3.3. Metode Analisis Data

Dalam menunjang penelitian ini, penelitian menggunakan beberapa alat bantu berupa seperangkat komputer dan *software* sebagai berikut :

a. Processor intel core i5
b. Ram 4 GB
c. Harddisk 1 TB
d. SSD 240 GB
e. Microsoft Excel
f. Google Colaboratory

3.4. Alur Penelitian

Alur penelitian dapat dilihat pada gambar 1. Ada beberapa tahapan yang dilakukan dalam penelitian ini diantaranya :

a. Identifikasi masalah

   Proses ini merupakan tahap dimana penulis mencari permasalahan yang ada dengan mencari sumber informasi permasalahan berupa artikel terkait dan jurnal peneltian terkait dengan sumber yang terpercaya.

b. Penyusunan proposal thesis

   Penyusunan rancangan penelitian berdasarkan topik permasalahan yang diangkat atau diselesaikan, dan pemilihan metode apa yang digunakan untuk membantu dalam proses penyelesaian.

c. Studi pustaka

   Proses ini, penulis mencari tahu informasi dengan membaca jurnal penelitian dan buku yang diaggap relevan dengan permasalahan yang akan diangkat.

d. Pengambilan dataset penelitian

   Setelah tahap identifikasi masalah dan studi literatur, tahap berikutnya yang akan dilakukan adalah mengambil dataset di kaggle sesuai dengan topik yang dipilih sebelumnya. File yang digunakan pada penelitian ini adalah file yang berekstensi .csv. Kemudian pada tahap studi literatur, dan experiment terhadap beberapa metode untuk menyelasaikan topik yang diangkat, maka

penulis telah menentukan algoritma yang digunakan untuk mencari algoritma mana yang paling efisien terhadap klasifikasi kemungkinan penyakit stroke. Penulis menggunakan algoritma *Extreme Gradient Boosting* dan *SMOTE* pada penelitian ini.

e. Pre-processing data

Sebelum masuk pada pengolahan data maka perlu mekukan pre-processing data untuk memeriksa dan memperbaiki kesalahan yang ditemukan pada data yang di gunakan dalam penelitian ini.

f. Split data

Pada tahap ini membagi dataset menjadi dua bagian yakni bagian yang digunakan untuk training data, dan untuk testing data, pembagian yang dilakukan yaitu berjumlah 70/30, yang dimana 70% untuk data training dan 30% untuk data testing.

g. Balancing data

Data yang di gunakan dalam pengolahan penelitian ini mengalami ketidakseimbangan kelas dimana jumlah kategori stroke dengan total 249 dan tidak stroke total 4861. Pada tahap ini menerapkan metode SMOTE untuk mengatasi ketidakseimbangan kelas.

h. Xgboost Clasification

Tahap ini adalah melakukan klasifikasi terhadap data yang sudah melewati tahapan sebelumnya yaitu Pre-processing data , dan

Balancing data. Data tersebut diklasifikasikan dengan menggunakan metode *Extreme Gradient Boosting*.

i. Tunning Parameter

Tahap ini ialah mengatur berbagai macam aspek dalam machine learning yang sangat berpengaruh pada performa dan model yang dihasilkan yaitu mengatur *hyperparameter*, Pencarian *hyperparameter* dilakukan secara manual, dengan menguji kumpulan *hyperparameter* pada parameter yang ditentukan sebelumnya.

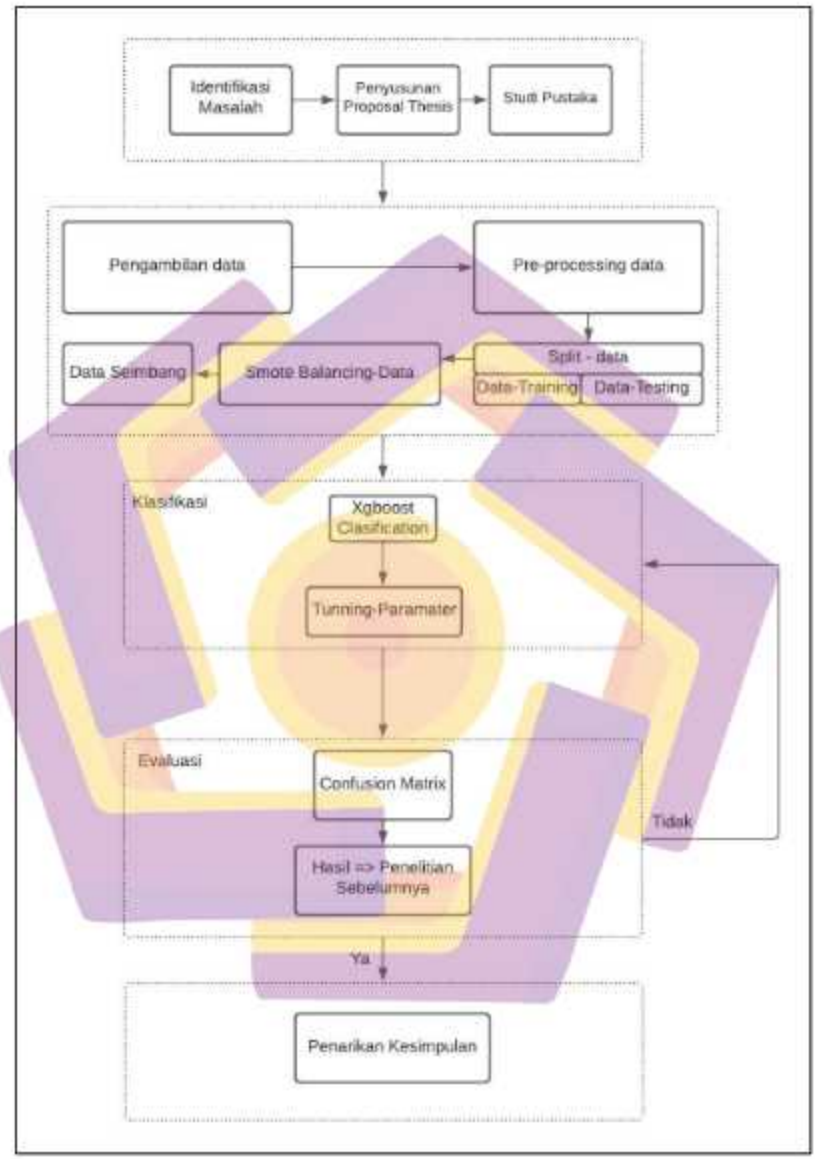
j. Evaluasi Metode

Menilai kinerja model pada proses klasifikasi, tahapan ini menggunakan confusion matrix untuk mengukur kinerja model yang di gunakan dalam proses klasifikasi, jika hasil akurasi dari klasifikasi melebihi hasil akurasi yang dihasilkan pada penelitian sebelumnya maka dilanjutkan pada proses berikutnya, jika tidak maka kembali pada proses sebelumnya yaitu proses klasifikasi dan mengatur parameter hingga mendapatkan hasil yang optimal.

k. Penarikan Kesimpulan

Pada tahap ini merupakan tahap pemberian kesimpulan berdasarkan hasil dari pengujian yang telah dilakukan. Hasil penelitian berupa fakta yang diperoleh metode yang di terapkan untuk pediksi stroke. Hasil pengujian dan evaluasi dijadikan

kesimpulan akhir mengenai metode balancing data, algoritma Xgboost untuk proses klasifikasi, parameter yang diujikan.



Gambar 3. 1 Alur Penelitian

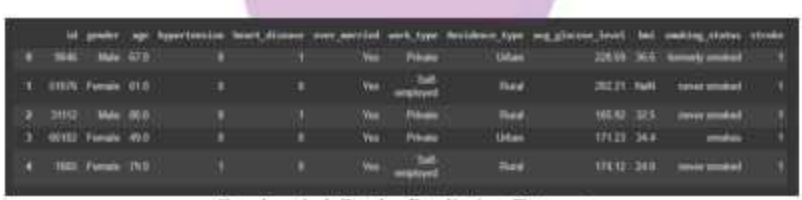
BAB IV

HASIL PENELITIAN DAN PEMBAHASAN

Berdasarkan kajian yang dilakukan sebelumnya yaitu kajian teori dan juga penelitian terdahulu yang dilakukan, maka penjelasan yang terdapat pada bab ini menjelaskan hasil-hasil yang telah di dapatkan dari proses komputasi. pembahasan meliputi tahapan dari penelitian ini diantaranya pengambilan Dataset, preprosessing data, mengatasi ketidakseimbangan kelas pada dataset menggunakan *Smote*, klasifikasi menggunakan metode *Extreme Gradient Boosting*, dan mengevaluasi menggunakan *Confusion Matrix* untuk mengukur performa dari perancangan model yang dilakukan saat proses klasifikasi.

4.1. Analisis Desktriptif

Pada penelitian ini menggunakan data sebagai bahan pengolahan klasifikasi yaitu Stroke Prediction Dataset, data tersebut di ambil dari kaggle dataset https://www.kaggle.com/fedesoriano/stroke-prediction-dataset, deskripsi dataset ini tidak menyampaikan secara spesifik jenis stroke seperti apa yang dimaksud. dataset ini digunakan untuk melatih dan menguji model klasifikasi khususnya pada prediksi penyakit stroke. berikut ini merupakan tampilan dataset yang dapat dilihat pada gambar 4.1.

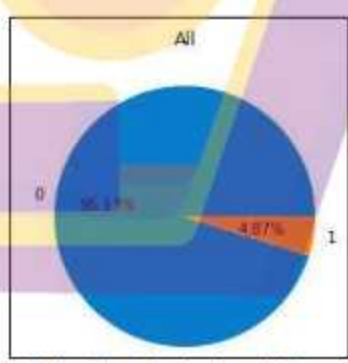


	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
0	9046	Male	67.0	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
1	51676	Female	61.0	0	0	Yes	Self-employed	Rural	202.21	NaN	never smoked	1
2	31112	Male	80.0	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1
3	60182	Female	49.0	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
4	1665	Female	79.0	1	0	Yes	Self-employed	Rural	174.12	24.0	never smoked	1

Gambar 4. 1 Stroke Prediction Dataset.

Pada Gambar 4.1 dataset ini terdiri dari 12 atribut, 10 variabel sebagai fitur tanpa variabel id, dan satu variabel sebagai label atau kelas yang digunakan untuk memprediksi penyakit stroke. 10 variabel yang di maksud adalah jenis Kelamin(Gender), Umur(Age), tekanan darah tinggi (hypertension), Penyakit Jantung(heart disease), pernah menikah(ever_married), tipe pekerjaan (Work_type), Residence_type(Tempat Tinggal), Rata-rata Level Gula(avg_glucose_level), indeks masa tubuh(bmi),dan Status Merokok(smoking_status). Kelas label pada atribut stroke dalam dataset ini memiliki dua nilai yaitu 0 dengan tandanya tidak terindikasi penyakit stroke; sedangkan nilai 1 menandakan terindikasi penyakit stroke.

Salah satu tahapan yang dilakukan sebelum masuk dalam analisis data adalah analisis deskriptif yang digunakan untuk menggambarkan kelas target pada dataset ini sehingga perlu untuk dilakukan keseimbangan kelas pada dataset.



Gambar 4. 2 Pie Chart Stroke dan tidak stroke.

Pada gambar 4.2 yang terdapat pie chart yang merupakan persentase terjadinya stroke dan tidak stroke pada dataset stroke prediction. Sebanyak 4.87% pasien terindikasi stroke dan sisanya yaitu sebanyak 95.13% pasien yang tidak

terindikasi penyakit stroke. dengan jumlah tersebut tentunya dataset yang di jadikan pengolahan dalam penelitian ini mengalami ketidakseimbangan kelas karena terdapat salah satu kelas pada kelas target yang memiliki jumlah lebih besar (kelas mayoritas) dibandingkan dengan kelas lain pada kelas target yang memiliki jumlah lebih kecil(kelas minoritas). Maka dengan itu diterapkan-nya metode *Synthetic Minority Over-Sampling Technique* (smote) untuk dapat mengatasi dataset mempunyai masalah imbalance kelas.

4.2. *Pre-Processing Data*

Sebelum masuk dalam proses klasifikasi perlu dilakukan teknik pre-prosessing data, agar menghasilkan data yang menjadi syarat sebagai bahan pengolahan proses klasifikasi dan menghasilkan klasifikasi yang baik. Dengan preprosessing data ini dilakukan agar metode yang digunakan mendapatkan hasil yang baik dalam proses klasifikasi, terdapat beberapa teknik dalam melakukan preprosessing sesuai dengan pola dataset yang terjadi pada penelitian ini, seperti LabelEncoder, dan Replace Missing Values.

a. Ada beberapa categorical data yang terdapat pada atribut dataset ini yaitu pada atribut gender, ever_married, work_type, residence_type, dan smoking_status. Dalam machine learning, data berbentuk seperti ini tidak dapat diproses. Categorical data harus dirubah dulu menjadi bentuk numeric. Maka dengan itu agar dapat diubah menjadi bentuk numeric dengan cara melakukan LabelEncoder yang mengubah setiap nilai dalam kolom menjadi angka yang berurutan yang dapat dilihat pada gambar 4.3.

```
enc = LabelEncoder()
```

```
data['gender'] = enc.fit_transform(data['gender'].values)
data['ever_married'] = enc.fit_transform(data['ever_married'].values)
data['work_type'] = enc.fit_transform(data['work_type'].values)
data['Residence_type'] = enc.fit_transform(data['Residence_type'].values)
data['smoking_status'] = enc.fit_transform(data['smoking_status'].values)
```

Gambar 4. 3 variabel categorical.

Pada gambar 4.3 merupakan tampilan dari atribut yang masih categorical datanya, atribut tersebut akan di ubah menjadi bentuk numeric, dapat dilihat pada gambar 4.4.

	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status
0	9046	1	67.0	0	1	1	2	1	228.69	36.600000	1
1	51676	0	61.0	0	0	1	3	0	202.21	30.471292	2
2	31112	1	80.0	0	1	1	2	0	105.92	32.500000	2
3	60182	0	49.0	0	0	1	2	1	171.23	34.400000	3
4	1665	0	79.0	1	0	1	3	0	174.12	24.000000	2

Gambar 4. 4 variabel categorical.

Dari gambar 4.4 merupakan hasil yang sudah dilakukan pengubahan data categorical ke numerical pada atribut gender, ever_married, work_type, residence_type, dan smoking_status.

b. Ada salah satu atribut yang mempunyai nilai kosong didalamnya, atribut tersebut ialah bmi (Body Mass Index) yang dapat dilihat pada gambar 4.5.

```
data.isnull().sum()

id                   0
gender               0
age                  0
hypertension         0
heart_disease        0
ever_married         0
work_type            0
Residence_type       0
avg_glucose_level    0
bmi                202
smoking_status       0
stroke               0
```

Gambar 4. 5 Missing Value.

Pada gambar 4.5 disampaikan bahwa pada atribut memiliki nilai yang kosong (Missing Value) sebanyak 202 data pada kolom bmi. Dengan adanya missing value ini maka menggantikan atribut yang bernilai kosong dengan rata-rata bmi mengalami stroke dan bmi tidak mengalami stroke menjadi pilihan, dibandingkan dengan menghapus baris data yang terdapat nilai kosong, karena tidak ada data yang terbuang.

4.3. Smote Balancing Data

Berdasarkan hasil ilustrasi label dalam dataset stroke prediction ini terjadi imbalance kelas, hal ini juga merupakan kondisi dari dataset yang dapat menghasilkan klasifikasi yang tidak optimal. dengan mengatasi imbalance yang terjadi digunakannya metode *Synthetic Minority Over-Sampling Technique* (SMOTE) untuk dapat mengatasi hal tersebut dengan cara menambah kelas minoritas agar sama dengan kelas mayoritas dengan menambahkan data buatan, data buatan atau sintesis tersebut di buat berdasarkan *k-tetangga* terdekat. berikut contoh perhitungan smote pada dataset sebelum dilakukan proses SMOTE (Sulistiyono et al., 2021).

Tabel 4. 1 Contoh Dataset Tidak Seimbang

Sumber : (Sulistiyono et al., 2021)

No	Atribut 1	Atribut 2	Kelas
1	1	2	1
2	2	3	1
3	4	3	1
4	6	2	2
5	6	4	2
6	5	4	2
7	4	4	2
8	5	6	2
9	6	3	2
10	4	5	2
11	6	7	2
12	5	3	2

Pada Tabel 4.1 diketahui kelas minoritas atau pada kelas 1 berjumlah 3, dan kelas mayoritas atau pada kelas 2 berjumlah 9, tahapan yang dilakukan untuk memperbanyak kelas minoritas menghitung jarak menggunakan euclidean distance dapat dihitung dengan persamaan (4.1).

$$dist = \sqrt{(X_1 - Y_1)^2 + (X_2 - Y_2)^2} + \cdots + (X_n - Y_n)^2 \quad (4.1)$$

agar mendapatkan instance tetangga terdekat Xknn dengan setiap kelas minoritas lainnya, berikut ini replikasi setiap kelas minoritas sebagai berikut :

Data ke-1 dari setiap kelas minoritas :

$$d\left(\begin{bmatrix}1\\2\end{bmatrix},\begin{bmatrix}1\\2\end{bmatrix}\right) = \sqrt{(1-1)^2 + (2-2)^2} = \sqrt{0}$$

$$d\left(\begin{bmatrix}1\\2\end{bmatrix},\begin{bmatrix}2\\3\end{bmatrix}\right) = \sqrt{(1-2)^2 + (2-3)^2} = \sqrt{2}$$

$$d\left(\begin{bmatrix}1\\2\end{bmatrix},\begin{bmatrix}4\\3\end{bmatrix}\right) = \sqrt{(1-4)^2 + (2-3)^2} = \sqrt{10}$$

Jarak dari setiap data ke-1 dengan data minoritas diurutkan dari yang terkecil = $\sqrt{0}, \sqrt{2}, \sqrt{10}$. berikut Data ke-2 dari setiap kelas minoritas :

$$d\left(\begin{bmatrix}2\\3\end{bmatrix},\begin{bmatrix}1\\2\end{bmatrix}\right) = \sqrt{(2-1)^2 + (3-2)^2} = \sqrt{2}$$

$$d\left(\begin{bmatrix}2\\3\end{bmatrix},\begin{bmatrix}2\\3\end{bmatrix}\right) = \sqrt{(2-2)^2 + (3-3)^2} = \sqrt{0}$$

$$d\left(\begin{bmatrix}2\\3\end{bmatrix},\begin{bmatrix}4\\3\end{bmatrix}\right) = \sqrt{(2-4)^2 + (3-3)^2} = \sqrt{4}$$

Jarak dari setiap data ke-2 dengan data minoritas diurutkan dari yang terkecil = $\sqrt{0}, \sqrt{2}, \sqrt{4}$. berikut Data ke-3 dari setiap kelas minoritas :

$$d\left(\begin{bmatrix}4\\3\end{bmatrix},\begin{bmatrix}1\\2\end{bmatrix}\right) = \sqrt{(4-1)^2 + (3-2)^2} = \sqrt{10}$$

$$d\left(\begin{bmatrix}4\\3\end{bmatrix},\begin{bmatrix}2\\3\end{bmatrix}\right) = \sqrt{(4-2)^2 + (3-3)^2} = \sqrt{4}$$

$$d\left(\begin{bmatrix}4\\3\end{bmatrix},\begin{bmatrix}4\\3\end{bmatrix}\right) = \sqrt{(4-4)^2 + (3-3)^2} = \sqrt{0}$$

Hasil dari proses smote dapat dilihat pada gambar 4.3 dibawah ini. Jarak dari setiap data ke-3 dengan data minoritas diurutkan dari yang terkecil = $\sqrt{0}, \sqrt{4}, \sqrt{10}$. Karena kelas mayoritas berjumlah 9 maka replikasi dari setiap data minoritas harus di replikasi sebanyak dua kali N dari SMOTE = 200.

Langkah berikutnya membangkitkan data sintesis, menggunakan persamaan (2.1). berikut ini perhitungan data synthesis pada kelas minoritas,

dengan jumlah N(replikasi smote) = 200 Xknn yang digunakan adalah acak data dari banyak N.

Data ke-1 :

$$X_{syn} = [1,2] + ([1,2] - [1,2]) \times 0,12 = [1,2]$$
$$X_{syn} = [1,2] + ([2,3] - [1,2]) \times 0,26 = [2,26 , 3,26]$$

Data ke-2 :

$$X_{syn} = [2,3] + ([4,3] - [2,3]) \times 0,31 = [2,62 , 3]$$
$$X_{syn} = [2,3] + ([1,2] - [2,3]) \times 0,21 = [1,79 , 2,79]$$

Data ke-3 :

$$X_{syn} = [4,3] + ([1,2] - [4,3]) \times 0,34 = [2,98 , 2,66]$$
$$X_{syn} = [4,3] + ([2,3] - [4,3]) \times 0,17 = [3,66 , 3]$$

berikut ini hasil setelah dilakukan SMOTE, yang dapat dilihat pada tabel Tabel 4.2.

Tabel 4. 2 Dataset Setelah Dilakukan Smote

Sumber : (Sulistiyono et al., 2021)

No	Atribut 1	Atribut 2	Kelas
1	1	2	1
2	2	3	1
3	4	3	1
4	6	2	2
5	6	4	2
6	5	4	2
7	4	4	2
8	5	6	2
9	6	3	2
10	4	5	2
11	6	7	2

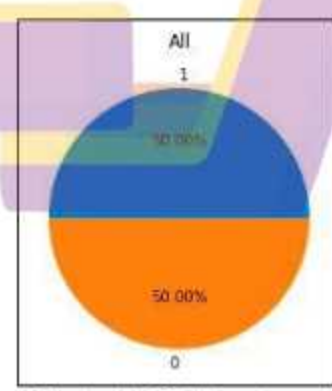
Tabel 4. 2 Dataset Setelah Dilakukan Smote (Lanjutan)

Sumber : (Sulistiyono et al., 2021)

No	Atribut 1	Atribut 2	Kelas
12	5	3	2
13*	1	2	1
14*	2.26	3.26	1
15*	2.62	3	1
16*	1.79	2,79	1
17*	2.98	2.66	1

*) data syntesis baru

Hasil dari proses balancing kelas pada dataset yang digunakan penelitian ini menggunakan smote menghasilkan pasien yang terkena stroke dengan persentase 50% dengan total jumlah 4868 , dan 0 untuk pasien yang tidak terindikasi stroke dengan persentase 50% dengan jumlah 4868. yang dapat dilihat pada gambar 4.3



Gambar 4. 6 Hasil smote.

Pada gambar 4.6 di atas merupakan hasil smote yang menambahkan hasil pada kelas minoritas pada kelas 1 (pasien yang terindikasi stroke) agar seimbah dengan kelas mayoritas pada kelas 0 (pasien yang tidak terindikasi stroke), jumlah yang ditambahkan yaitu berjumlah 4619 data, berikut hasil data balancing dapat dilihat pada gambar 4.7.

5118	0.740463	64.29768	0.259537	0	1	2.259537	0.740463	207.9856	29.66185	2	1
5119	1	74.73956	1	0	1	3	0.420147	194.6305	29.43685	1.420147	1
5120	0	80.61187	0	0	0.129376	0.258751	1	69.96058	25.64475	2.870624	1
5121	1	54.87991	0	0	1	0.439957	0.219979	86.62912	30.9617	3	1
5122	0	78.50515	0	0	1	2.167715	0	79.33663	19.14717	0.33543	1
5123	0.808169	81.61634	0.191831	0	1	2.191831	0	87.46406	30.13228	1	1
5124	0	65.4815	0	0	1	2	0.839501	206.619	44.92465	0.839501	1
5125	0.65717	40.62668	0	0	0.65717	2.34283	0.34283	83.0226	24.97004	0.34283	1
5126	1	57.18539	0	0	1	2	0.185388	84.46667	35.54527	0.556165	1
5127	0.749918	74.99967	0	0	1	1.499836	1	137.9288	29.52488	1.250082	1
5128	1	57.37251	0	0	1	2	0	87.04604	38.07827	0	1
5129	0.693155	47.45476	0.306845	0	1	2	0.306845	61.65212	29.67616	1.386311	1
5130	0	13.14034	0	0	0	4	0.067797	58.77319	30.87094	0	1
5131	0.375942	80.62406	0	0.375942	0.375942	0.751885	1	85.63458	25.19849	3	1
5132	1	76.08561	0.042804	0	1	2.042804	0.957196	139.8731	30.09817	0.957196	1
5133	0.262547	78.52509	0.262547	0	1	2	0.262547	236.801	30.9085	2	1
5134	0.007142	72.00714	0.007142	0	1	3	0.007142	90.02235	31.98908	1.992858	1
5135	0.408115	69.22434	0	0.408115	0.408115	0.816229	1	81.96124	26.62587	1.224344	1
5136	0.444042	58.44042	0	0.444042	1	2.555958	0.444042	191.7288	29.11228	3	1
5137	0	71.45365	0.728624	0	1	2.271376	0	75.09713	23.65513	1	1
5138	0.384766	79.23047	1	0	1	3	0.384766	91.83185	31.96552	1	1
5139	0	73.50624	0	0.218751	1	0.437502	0.781249	85.79531	30.80112	0.218751	1
5140	0.794202	60.5684	0	0.794202	1	2	0	112.7628	28.57971	3	1
5141	0.827734	57.34453	0.827734	0	1	0.344533	0.827734	78.84282	28.1774	0.827734	1
5142	1	72.88552	0.057241	1	0.942759	0	0	219.5703	28.73169	2	1

Gambar 4. 7 Hasil Balancing Kelas Dataset

Pada Gambar 4.7 diatas merupakan beberapa data yang sudah diseimbangkan menggunakan smote pada dataset stroke prediction hasil tersebut merupakan nilai dari fitur dataset berupa data sytesis yang dibangun oleh smote, gambar 4.1 menampilkan hasil balancing data sebanyak 25 data mulai dari data hasil smote 5118 sampai dengan data ke-5142.

4.4. *Normalisasi Data*

Hasil balancing pada dataset ini masih mempunyai atribut dengan nilai skala yang berbeda jauh, hal ini dapat membuat model Machine Learning tidak optimal

dalam melakukan klasifikasi, Karena atribut pada dataset ini memiliki skala yang berbeda (Wahanani et al., 2020). skala yang berbeda contohnya pada dataset ini yaitu value pada kolom hypertensi dan pada kolom avg_glukosa_level yang dapat dilihat pada gambar 4.1 dataset, maka perlu dilakukan standarisasi untuk memiliki skala yang sama saat membangun model pembelajaran mesin. Teknik normalisasi data yang digunakan adalah Min Max normalization yang merupakan metode normalisasi dengan melakukan transformasi linier terhadap data asli sehingga menghasilkan keseimbangan nilai perbandingan antar atribut, atribut-atribut ini ketika dikonversi dan menghasilkan perhitungan yang similaritas, maka nantinya dapat berada di rentang 0 hingga 1.

keseimbangan tersebut adalah nilai skala antar atribut yang terdapat dalam dataset. Metode ini dapat menggunakan rumus sebagai berikut (Wahanani et al., 2020) :

$$N = \frac{MinRange + (X - MinValue)(MaxRange - MinRange)}{MaxValue - MinValue} \quad (4.1)$$

Dimana :

N = Normalisasi Min_Max

MinRange = Nilai Konversi Kecil Yang ditentukan

MaxRange = Nilai Konversi Terbesar yang ditentukan

MaxValue = Nilai Terbesar pada atribut yang dibandingkan

MinValue = Nilai Terkecil pada atribut yang dibandingkan

Berikut ini contoh normalisasi pada data ke 5118 yang dapat dilihat pada gambar 4.1 diatas. contoh perhitungan normalisasi Min-Max pada atribut Gender dengan atribut Smoking_status.

Diketahui :

MinRange = 0

MaxRange = 1

X = 64.29768464

MinValue = 0

MaxValue = 207.9855855

Normalisasi = $\frac{0+(64.29768464-0)(1-0)}{207.9855855-0} = 0.30914482996$

Berikut ini hasil standarisasi menggunakan teknik normalisasi min_max pada dataset stroke prediction yang dapat dilihat pada gambar 4.8.

```
1  Scaler.fit_transform(XTrainSmote,YTrainSmote)

array([[0.        , 0.62158285, 0.        , ..., 0.41558502, 0.2416953 ,
        0.        ],
       [0.5       , 0.91455078, 0.        , ..., 0.70391242, 0.23138603,
        0.33333333],
       [0.        , 0.63378906, 0.        , ..., 0.5941152 , 0.18213058,
        0.66666667],
       ...,
       [0.5       , 0.90386152, 0.        , ..., 0.22649866, 0.22224005,
        0.07843337],
       [0.5       , 0.70872411, 0.        , ..., 0.12746552, 0.23030852,
        0.90754716],
       [0.40663504, 0.72845361, 0.        , ..., 0.25381433, 0.23105713,
        0.27042336]])
```

Gambar 4. 8 Hasil Normalisasi.

4.5. Train/Test Split

Membagikan dataset menjadi data training dan data testing, pembagian pada penelitian ini dibagikan data training dan data testing menjadi 70/30. Dengan jumlah pembagian tersebut mempunyai tujuan melihat model dalam memprediksi ketika mempunyai data test dengan total 2922, Secara umum model machine learning mendapatkan hasil akurasi yang baik jika memiliki jumlah data testing yang sedikit. Maka dalam penelitian ini meningkatkan data test dan

menguji apakah model mendapatkan hasil yang baik atau tidak. Pada Tabel 4.1 menggambarkan pembagian data yang di lakukan.

Tabel 4. 3 Train/Test Split

Keterangan	Data Training	Data Testing	Total
Proporsi	70%	30%	100%
Jumlah	6815	2922	9737

Pada Tabel 4.3 di jelaskan bahwa pembagian data training dan data testing yang dilakukan yaitu membagikan/split data menjadi 70/30, 70% untuk data training yang berjumlah 6815 data dan 30% untuk data testing yang berjumlah 2922 data, dengan itu jumlah keseluruhan data dari dataset berjumlah 9737 total data.

4.6. Klasifikasi Xgboost

Penggunaan metode dalam penelitian ini adalah metode extreme gradient boosting, berikut ini langkah-langkah penyusunan Algoritma Extreme Gradient Boosting dengan diberikan suatu dataset dengan dua variabel yaitu variabel dosis obat, dan efektifitas (Ichi.Pro, 2020).

Tabel 4. 4 Contoh Dataset

Sumber : (Ichi.Pro, 2020).

X	Y
2	0
8	1
12	1
18	0

Dengan dosis obat = X, efektifitas = Y

Parameter-parameter yang dipakai dalam pembangunan metode yaitu n_estimators atau model yang dibangun sebanyak 2, max_depth = 2, learning_rate = 1, gamma = 2, min_child_weight = 0, reg_lambda = 0, dan base_score = 0.5.

a. Prediksi awal

Biarkan prediksi awal atau base_score menjadi 0,5 untuk semua titik data dalam dataset yaitu Fo (x) = ho (x) = 0,5.

Tabel 4. 5 Perhitungan Residual

Sumber : (Ichi.Pro, 2020).

X	Y	Fo(x)	Ỹ = Y - Fo(x)
2	0	0.5	-0.5
8	1	0.5	0.5
12	1	0.5	0.5
18	0	0.5	-0.5

b. Model latih

Melatih model Pertama (Model pelatihan disini mengacu pada bangunan pohon) yaitu M1 dengan menggunakan [X, Ỹ] sebagai datanya dan model tersebut adalah pohon Xgboost khusus yang dibangun secara berbeda jika dibandingkan dengan pohon keputusan normal. dalam membangun pohon *XGBoost* menggunakan rumus untuk menyelesaikan masalah optimasi pada XGBoost sebagai berikut.

$$\text{Gain} = \text{Left}_{\text{Similarity}} + \text{Right}_{\text{Similarity}} - \text{Root}_{\text{Similarity}}$$

$$\text{Similarity Score} = \frac{(\sum \text{Residual}_i)^2}{\sum [\text{Previous Probability}_i \times (1 - \text{Previous Probability}_i)] + \lambda}$$

$$\text{Output Value} = \frac{(\sum \text{Residual}_i)}{\sum [\text{Previous Probability}_i \times (1 - \text{Previous Probability}_i)] + \lambda}$$

Dimana :

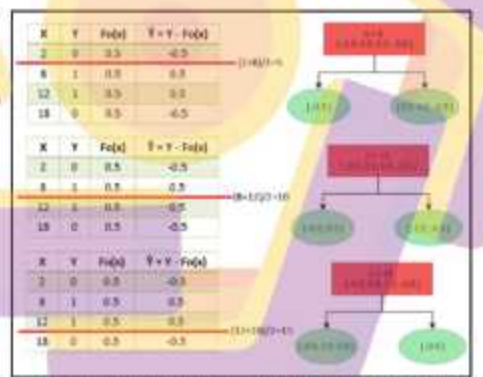
$\hat{y}$ = Residual ke-i

λ = reg_lambda

Previous f1(x) = probabilitasi sebelumnya

- Perhitungan nilai gain hanya untuk menghitung akar pohon
- Perhitungan similarity untuk semua node
- Perhitungan output value hanya untuk leaf node
- Lambda merupakan parameter, apabila nilai lambda meningkat akan menghasilkan pruning lebih banyak node pada pohon yang dibangun.

i. Pohon dibangun dengan membagi data menjadi dua partisi dari berbagai kemungkikan pemisahan atau split.

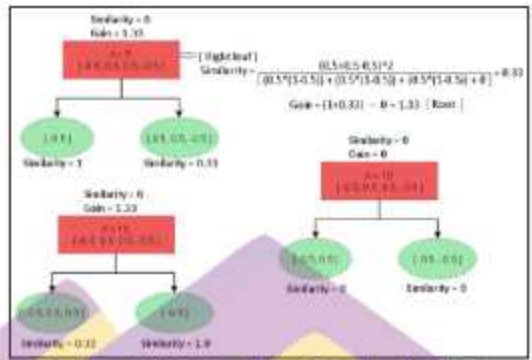


Gambar 4. 9 Contoh Pembangunan Pohon Extreme Gradient Boosting.

Sumber : Sumber : (Ichi.Pro, 2020).

Ambang batas akar dihitung dengan mengambil rata-rata dua titik dekat di antara perpecahan dan sisa menuju ke daun masing-masing.

ii. Setelah pohon dibangun dilanjutkan perhitungan Gain dan similarity yang dapat dilihat pada Gambar 4. 2.

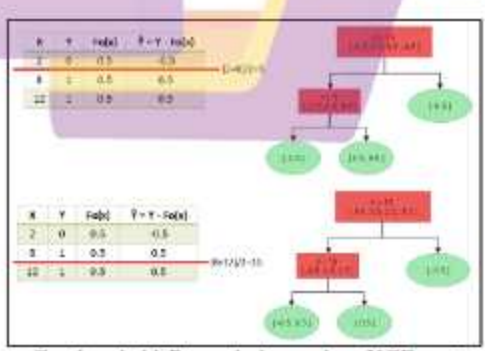


Gambar 4. 10 Contoh Nilai Gain dan Similarity.

Sumber : Sumber : (Ichi.Pro, 2020).

Setelah hasil Gain dan similarity dihasilkan yang dapat dilihat pada gambar 4.10.

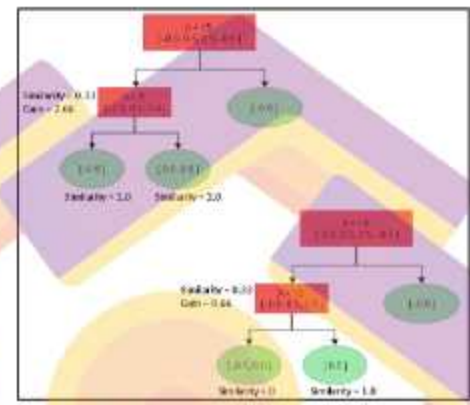
iii. langkah selanjutnya pilih nilai *Gain* yang *maksimum*, dengan itu terdapat pada pohon X<15 yang terpilih. karena nilai parameter *max_dept* adalah 2 maka penumbuhan pohon lebih jauh atau sebanyak nilai parameter *max_dept* dengan memisahkan data yang ada di daun kiri pohon sebelumnya. yang dapat diliha pada gambar 4.11.



Gambar 4. 11 Penumbuhan pohon XGBoost

Sumber : Sumber : (Ichi.Pro, 2020).

Pemisahan data yang baru saja dilakukan untuk memilih root internal yang memiliki Penguatan maksimal. Hitung kembali Persamaan(similarity) dan Keuntungan(Gain) untuk memilih root internal yang memiliki Penguatan maksimal, yang dapat dilihat pada gambar 4.12.

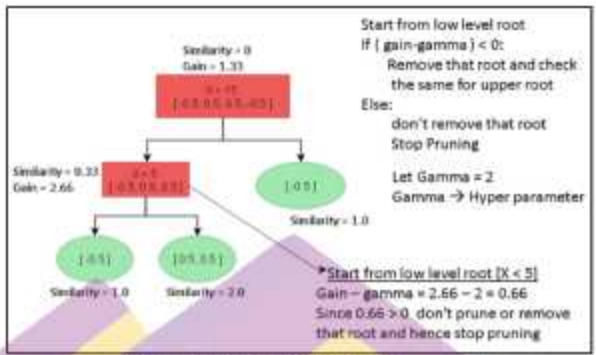


Gambar 4. 12 Gain dan Similarity pohon lanjutan

Sumber : Sumber : (Ichi.Pro, 2020).

Hasil yang terdapat pada gambar 4.12. pohon X<5 dipilih sebagai akar internal untuk konstruksi yang merupakan pohon pertama dalam gambar karena akar internal tersebut memiliki Gain maksimum 2.66.

iv. setelah itu dilakukan pruning yang bertujuan untuk memperkecil ukuran pohon keputusan dengan menghilangkan bagian pohon yang memiliki kekuatan yang kurang untuk mengklasifikasikan kejadian. langkah pruning dapat dilihat pada gambar 4.13.

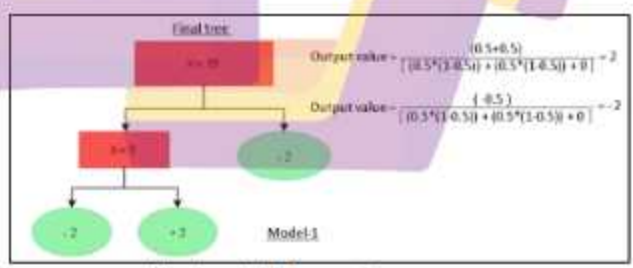


Gambar 4. 13 Pemangkasan

Sumber : Sumber : (Ichi.Pro, 2020).

Pada gambar 4.13. merupakan langkah pruning(pemangkasan) hal ini jika nilai *gain* kuranf *gamma* lebih kecil dari 0 maka tidak dilakukan penghapusan.

v. langkah berikutnya yaitu Menghitung nilai output untuk semua leaf untuk mendapatkan pohon terakhir pada model 1, karena beberapa leaf memiliki lebih dari satu residual, perhitungan tersebut dapat dilihat pada gambar 4.14.



Gambar 4. 14 Pemangkasan

Sumber : Sumber : (Ichi.Pro, 2020).

Setelah mendapatkan hasil dari perhitungan nilai output.

vi. dilanjutkan Menghitung prediksi dari model 1, untuk mendapatkan *h1(x)* dan hitung prediksi *f1(x)* dan *residu*.

Diberikan nilai learning_rate = 1.0

Maka

$$F1(x)=\sigma([ho(x)/(1-ho(x))]+[\eta * h1(x)])$$

Memecahkan *f1(x)* pada klasifikasi didapatkan:

$$F1(x) = \sigma\,(0+1*(h1(x)))$$

Fungsi sigmoid :

$$\sigma\,(x)= 1/(1 + \exp(-x))$$

Tabel 4. 6 Perhitungan Nilai Prediksi pada Model-1

Sumber : Sumber : (Ichi.Pro, 2020).

X	Y	h1(x)	$F1(x) = \sigma(0+1*(h1(x)))$	Y = Y – F1(x)
2	0	-2	0.11	-0.11
8	1	2	0.88	0.12
12	1	2	0.88	0.12
18	0	2	0.11	-0.11

Pada tabel 4.6 merupakan hasil dari prediksi model 1, barikutnya akan dilanjutkan perhitungan prediksi model berikutnya yaitu pada model 2 yang sama seperti perhitungan yang dilakukan sebelumnya, Ulangi Langkah 3 untuk membangun model berikutnya.

Dalam melakukan klasifikasi kemungkinan pasien terindikasi penyakit stroke atau tidak sebagaimana yang dilakukan dalam penelitian ini. Untuk proses

klasifikasi langkah utama yang dilakukan ialah tuning parameter. Hasil dari tuning parameter dapat dilihat pada tabel 4.3 dibawah ini.

Tabel 4. 7 Hasil *Tuning* Parameter Metode *Extreme Gradient Boosting*.

No.	*max_ depth*	*Eta (learning_rate)*	*min_ child_ weight*	*n_ estimators*	*Random_ State*	*Sub-sample*	*Accuracy*	
Percobaan pada data yang tidak seimbang.								
1.	4	0.2	1	40	-	-	96%	
2.	5	0.1	1	50	0	0.1	91%	
3.	5	0.2	1	60	0	0.1	94%	
4.	7	0.1	1	70	5	0.1	95%	
5.	8	0.1	1	70	5	0.2	95%	
6.	7	0.1	1	90	5	0.3	95%	
7.	10	0.2	1	100	0	0.2	95%	
8.	15	0.2	1	100	-	0.5	94%	
9	15	0.09	1	200	-	1	95%	

Tabel 4. 7 Hasil *Tuning* Parameter Metode *Extreme Gradient Boosting*. (Lanjutan)

No.	*max_ depth*	*Eta (learning_rate)*	*min_ child_ weight*	*n_ estimators*	*Random_ State*	*Sub-sample*	*Accuracy Normalisasi*	*Accurasi Tanpa Normalisasi*
Percobaan pada data seimbang.								
10.	4	0.2	1	40	-	-	89%	88%
11.	5	0.1	1	50	0	0.1	93%	92%
12.	5	0.2	1	60	0	0.1	92%	91%
13.	7	0.1	1	70	5	0.1	94%	93%
14.	8	0.1	1	70	5	0.2	95%	95%
15.	7	0.1	1	90	5	0.3	97%	96%
16.	10	0.2	1	100	0	0.2	97%	97%
17	15	0.2	1	100	-	0.5	98%	98%
18.	15	0.09	1	200	-	1	99%	99%

Pada tabel 4.7 diatas dapat dilihat bahwa parameter yang digunakan pada proses klasifikasi dari percobaan-percobaan yang dilakukan adalah *max_depth, learning_rate, min_child_weight, n_estimator, random_state, lambda, sub-sample,* dan *gamma*. split data yang digunakan sebelum masuk pada proses klasifikasi adalah 70/30, 70 untuk data training dan 30 untuk data testing. Penelitian ini terdapat 18 kali percobaan klasifikasi, jumlah percobaan ini dibagi dua yaitu 9 untuk percobaan klasifikasi menggunakan data yang tidak seimbang, dan 9 lainnya menggunakan data seimbang. Sembilan percobaan pertama dan kedua memiliki nilai parameter yang sama. Penelitian ini terdiri dari enam

tahapan diantaranya pre-prosessing data, balancing kelas pada dataset, split data, klasifikasi dengan XGBClassifier, pengaturan paramater, dan evaluasi.

Sembilan Percobaan pertama mempunyai hasil akurasi terbaik yaitu 96% yang mempunyai empat puluh pohon yang digunakan, kemudian setiap pohon memiliki empat percabangan, eta menggunakan nilai 0.2 yang digunakan sebagai tingkat pembelajaran yang mempengaruhi algoritma xgboost dalam membuat klasifikasi berbentuk pohon, jumlah minimum berat yang digunakan bernilai 1 dimana jika partisi pohon menghasilkan simpul daun dengan jumlah bobot instance kurang dari min_child_weight, maka proses pembangunan akan menghentikan partisi lebih lanjut, dan subsample merupakan rasio sampel dari instance pelatihan, yang digunakan dalam percobaan ini adalah 1, berarti XGBoost akan mengambil sampel secara acak keseluruhan dari data pelatihan sebelum menanam pohon.

Berikutnya hasil percobaan ke-18 yang mendapatkan hasil terbaik diantara percobaan ke-10 hingga ke-17 menggunakan data seimbang. percobaan ini memiliki hasil akurasi 99%, hasil tunning parameter yang diatur mempunyai perbedaan dengan hasil terbaik pada percobaan 1 sebelumnya karena jumlah dataset ini menjadi lebih banyak jika sudah dilakukan balancing kelas pada dataset maka dengan itu jumlah pohon yang diatur mengalami peningkatan sebanyak dua ratus pohon yang digunakan, agar mampu memperbaiki kesalahan yang dibuat oleh model di iterasi sebelumnya, kemudian setiap pohon memiliki lima belas percabangan, learning_rate diatur menurun yaitu 0.09 agar eror dari setiap pembaruan klasifikasi menjadi turun, subsample merupakan rasio sampel dari

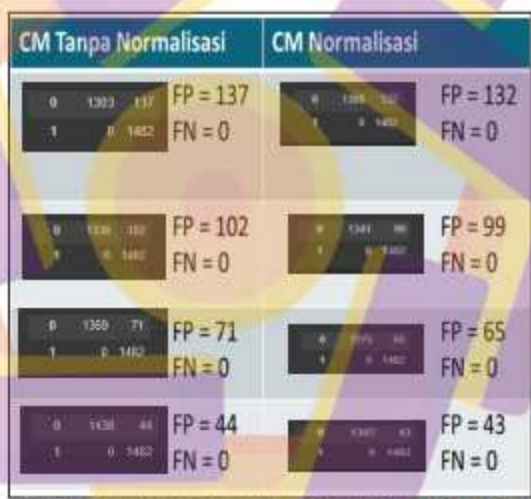
instance pelatihan, yang digunakan dalam percobaan ini adalah 1, dan Jumlah bobot minimum pada child node adalah 1.

Pada tabel diatas, sembilan percobaan pertama dengan menggunakan data tidak seimbang menghasilkan akurasi tertinggi sebesar 96% yaitu pada percobaan ke 1. hasil terbaik pada sembilan percobaan pertama dengan dengan menggunakan data yang tidak seimbang ini ketika menggunakan nilai parameter seperti n_estimator yang tidak kurang dari 40 dan tidak lebih besar dari 40, lalu max_dept tidak lebih dari 4, dan tidak membatasi subsample maka akurasi yang dihasilkan akan lebih baik jika dibandingkan dengan meningkatkan-nya kedua parameter n_estimator, max_dept dan membatasi sub_sample. berbeda dengan klasifikasi menggunakan data seimbang yang dimana sebelumnya sudah di terapkan metode smote untuk mengatasi ketidakseimbangan kelas, ketika parameter n_estimator, max_dept ditingkatkan lalu learning_rate nya dikurangin, dan sub_sample nya tidak dibatasi maka dapat meningkatkan akurasi sebanyak 3% yaitu 99% dari hasil penelitan sebelumnya yang hanya menggunakan algoritma Xgboost saja tanpa menggunakan metode imbalance yaitu metode smote mendapatkan akurasi 96%, dengan hasil 99% ini merupakan hasil dari percobaan ke-18.

Pada Tabel 4.7 diatas terdapat perbedaan antara klasifikasi tanpa normalisasi data dan klasifikasi menggunakan normalisasi data, Terdapat lima percobaan dengan perlakuan yang sama, klasifikasi menggunakan teknik normalisasi data mampu menaikan akurasi 1% dari hasil akurasi klasifikasi tanpa menggunakan teknik normalisasi. Terdapat 4 percobaan dengan teknik normalisasi data yang mendapatkan hasil akurasi yang sama dengan hasil tanpa dilakukan normalisasi,

Patokan nilai akurasi yaitu pada classification report jadi jika hasil akurasi yang dihasilkan mengalami kenaikan tidak lebih dari 0,5 maka akurasi dianggap sama contoh akurasi yang dihasilkan adalah 95,45 maka hasil yang di tampilkan adalah 95% akurasinya, begitupun sebaliknya jika lebih maka mengalami kenaikan 1% contohnya 95,67 maka yang di tampilkan akurasinya sebesar 96%.

Tetapi dengan teknik normalisasi data dapat mengurangi kesalahan dalam mengklasifikasi stroke atau tidak yaitu pada nilai FP dan FN, berikut perbedaannya dapat dilihat pada gambar 4.1.

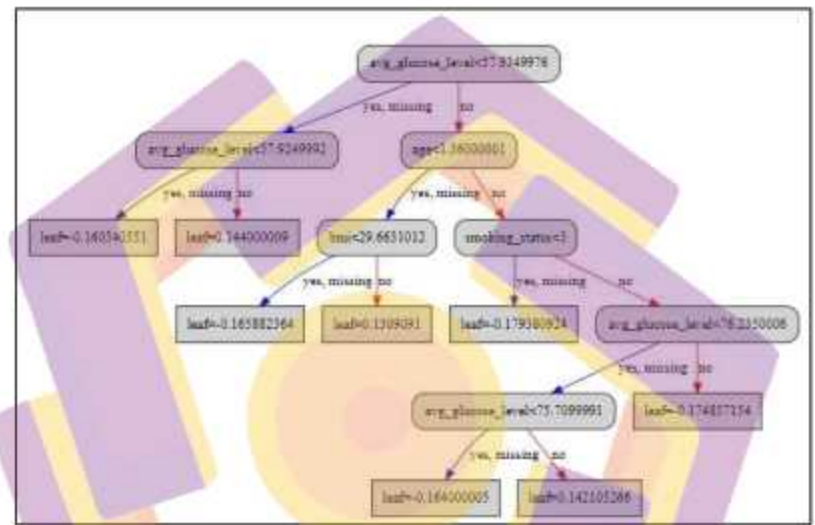


Gambar 4. 15 Perbedaan Hasil Confusion Matrix Normalisasi dan tanpa Normalisasi

pada gambar 4.15 adalah hasil Confusion Matrix dari proses klasifikasi menggunakan teknik normalisasi dan tanpa normalisasi, ada beberapa hal yang terjadi ketika melakukan normalisasi data yaitu Berkurangnya kesalahan yaitu pada nilai FP pada percobaan ke 5,7,8 dan 9. Percobaan ke 5 FP = 137 berkurang

menjadi 132. Percobaan ke 7 FP = 102 berkurang menjadi 99 Percobaan ke 8 FP = 71 berkurang menjadi 65 Percobaan ke 5 FP = 44 berkurang menjadi 43.

berikutnya contoh pohon dari Extreme Gradient Boosting (XGBoost) pada percobaan yang memiliki hasil terbaik dari penelitian ini.



Gambar 4. 16 Contoh Pohon Extreme Gradient Boosting.

Dari Gambar 4.16 hasil dari Contoh Pohon Extreme Gradient Boosting. Hasil yang didapatkan dari dari pohon diatas sebagai berikut :

a. Jika avg_glucose_level kurang dari 57.9349976, avg_glucose_level kurang dari 57.9249992 maka akan mengembalikan -0.160540551 untuk pohon selanjutnya dan yang lain akan mengembalikan 0.144000009.

b. Jika avg_glucose_level kurang dari 57.9349976, age kurang dari 1.36000001, bmi kurang dari 29.6631012 maka akan

mengembalikan -0.165882364 untuk pohon selanjutnya, dan yang lain akan mengembalikan 0.1309091.

c. Jika avg_glucose_level kurang dari 57.9349976, age kurang dari 1.36000001, dan smoking_status kurang dari 3 maka akan mengembalikan -0.179380924.

d. Jika avg_glucose_level kurang dari 57.9349976, age kurang dari 1.36000001, dan smoking_status lebih dari 3, avg_glucose_level lebih dari 76.2350006 maka akan mengembalikan -0.174857154.

e. Jika avg_glucose_level kurang dari 57.9349976, age kurang dari 1.36000001, dan smoking_status lebih dari 3, avg_glucose_level kurang dari 76.2350006, dan avg_glucose_level kurang dari 75.7099991 maka akan mengembalikan -0.164000005 untuk pohon selanjutnya, dan yang lain akan mengembalikan 0.142105266.

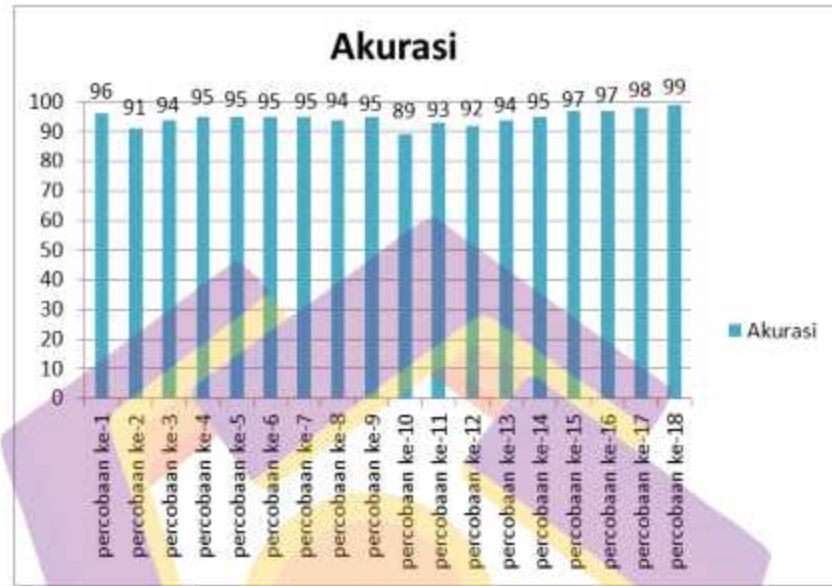
Akurasi dari delapan belas percobaan yang dilakukan, dari hasil percobaan-percobaan tersebut adalah menunjukan bahwa Percobaan ke-18 dengan menggunakan split data 70/30 yang sudah di seimbangkan kelasnya menggunakan smote mendapatkan nilai akurasi terbaik yaitu 99%, berikut 10 hasil klasifikasi dari percobaan ke- 18 yang dapat dilihat pada gambar 4.17.

```
HasilPrediksi = pd.DataFrame(y_predict)
print("HASIL KLASIFIKASI STROKE : \n - 1 = pasien stroke \n - 0 = pasien tidak stroke",HasilPrediksi.head(10))
```

```
HASIL KLASIFIKASI STROKE :
 - 1 = pasien stroke
 - 0 = pasien tidak stroke
   0
0  0
1  1
2  0
3  0
4  1
5  1
6  0
7  1
8  0
9  1
```

Gambar 4. 17 Hasil Klasifikasi

Pada gambar 4.17 ini merupakan hasil klasifikasi yang ditampilkan yang berjumlah 10 data dari hasil data test, yang pertama model memprediksi sebagai kelas 0 atau kelas pasien tidak mengalami stroke, dan pada data test yang kedua model memprediksi kelas 1 atau kelas pasien yang mengalami stroke, hasil klasifikasi data test berikutnya dapat dilihat pada gambar 4.17 diatas. hasil akurasi dari keseluruhan percobaan-percobaan yang dilakukan. Dapat dilihat pada gambar 4.18.

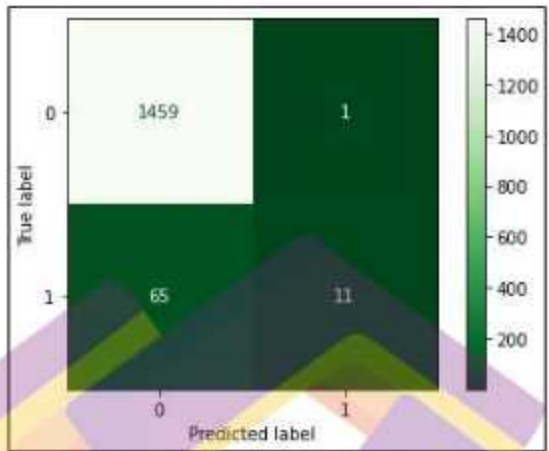


Gambar 4. 18 Hasil Akurasi.

4.7. Evaluasi Metode dengan Confusion matrix dan Classification Report

4.6.1 Hasil Confusion Matrix dari Uji Coba Klasifikasi.

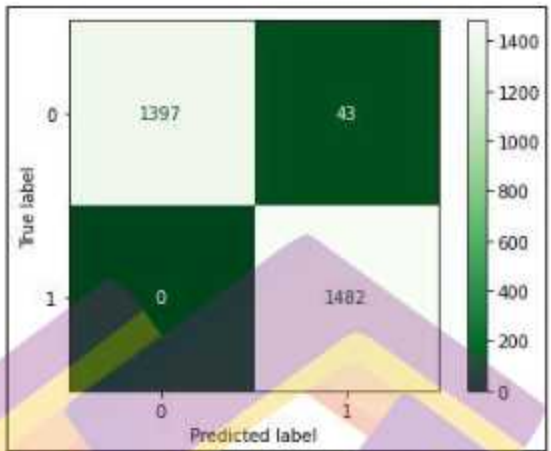
Dari percobaan ke-1 sampai pada percobaan ke-9 dengan split data 70/30 dengan menggunakan data yang tidak seimbang, dari sembilan percobaan tersebut hasil Confusion Matrix terbaik yaitu pada percobaan ke-1 dapat dilihat pada gambar 4.19.



Gambar 4. 19 Hasil Confusion Matrix percobaan ke-1.

Pada Gambar 4.19. merupakan hasil Confusion Matrix dari eksekusi python, hasil dari Confusion Matrix ini merupakan hasil dari percobaan yang memiliki hasil terbaik dengan menggunakan data yang tidak seimbang yaitu pada percobaan ke-1. Berikut Hasil confusion matrix pada Percobaan ke-1 yaitu True Positive (TP) berjumlah 1459, untuk True Negatif (TN) berjumlah 11, untuk False Positif (FP) adalah 1, dan untuk False Negative (FN) adalah 65.

Berikutnya pada percobaan ke-10 sampai dengan ke-18, dengan sembilan percobaan tersebut menggunakan split data yang sama dengan data seimbang, dari sembilan percobaan tersebut hasil Confusion Matrix terbaik yaitu pada percobaan ke-18, yang dapat dilihat pada gambar 4.20.



Gambar 4. 20 Hasil Confusion Matrix percobaan ke-18.

Pada Gambar 4.20. merupakan hasil Confusion Matrix dari eksekusi python, hasil tersebut merupakan hasil percobaan ke-18, dimana hasil tersebut merupakan hasil terbaik dari hasil confusion matrix lainnya yaitu pada percobaan ke-10 sampai dengan ke-17, hasil yang didapatkan pada percobaan ke-18 yaitu True Positive (TP) adalah 1397, untuk True Negatif (TN) adalah 1482, untuk False Positif (FP) adalah 43, dan untuk False Negative (FN) adalah 0.

Hasil pada Confusion matrix diatas diantaranya adalah Nilai True Positive (TP), True Negatif (TN), False Positif (FP), False Negative (FN) maksud antara nilai-nilai hasil confusion matrix tersebut ialah sebagai berikut.

a. True Positive (TP) dimana pasien yang diprediksi stroke, memang benar secara actual bahwa pasien tersebut stroke.
b. True Negatif (TN) dimana pasien yang di prediksi tidak stroke, dan memang benar bahwa pasien tersebut tidak stroke.

c. False Positif (FP) dimana pasien secara actual nya tidak terindikasi stroke, tetapi diprediksi stroke.

d. False Negative (FN) dimana pasien tersebut terindikasi stroke, tetapi diprediksi tidak stroke.

4.6.2 Classification Report dari Uji Coba Klasifikasi.

Hasil pengujian sebelumnya, lalu dilakukannya evaluasi menggunakan confusion matrix untuk menilai kinerja dari model yang dibangun yang menggunakan algoritma XGBoost (Xtreme Gradient Boosting) pada data yang mengalami ketidakseimbangan kelas dan pada data yang sudah seimbang kelas. terlihat hasil terbaik dari percobaan yang dilakukan dengan beberapa parameter yang di ujikan maka percobaan ke-1 memiliki hasil akurasi terbaik pada data yang tidak seimbang kelasnya yaitu dengan nilai akurasinya adalah 96%. lalu pada percobaan yang menggunakan data seimbang, percobaan ke-18 memiliki hasil akurasi terbaik yaitu dengan nilai akurasinya adalah 99%. kedua hasil terbaik pada klasifikasi menggunakan kelas seimbang dan tidak menggunakan data yang sudah seimbang kelasnya dapat diilihat pada gambar 4.21.

Percobaan ke-1 pada data tidak seimbang

	precision	recall	f1-score	support
0	0.96	1.00	0.98	1460
1	0.92	0.14	0.25	76
accuracy			0.96	1536
macro avg	0.94	0.57	0.61	1536
weighted avg	0.96	0.96	0.94	1536

Percobaan ke-18 pada data seimbang

	precision	recall	f1-score	support
0	1.00	0.97	0.98	1440
1	0.97	1.00	0.99	1482
accuracy			0.99	2922
macro avg	0.99	0.99	0.99	2922
weighted avg	0.99	0.99	0.99	2922

Gambar 4. 21 Hasil Classification Report percobaan ke-1 dan Percobaan ke-18.

Dari gambar 4.21 merupakan Hasil Classification Report pada percobaan ke-1 dan percobaan ke-18 yang dimana percobaan ke-1 menggunakan data yang tidak seimbang kelasnya, dan untuk percobaan ke-18 menggunakan data yang seimbang kelasnya. hasil pada percobaan ke-1 diatas memiliki nilai accuracy 96% yang dimana dengan accuracy ini merupakan rasio pasien yang benar diprediksi Stroke dan Tidak Stroke dari kesuluruhan Pasien yang ada pada dataset, lalu recall mengetahui bahwa Pasien yang diprediksi Stroke dibandingkan dengan keseluruhan Pasien yang sebenarnya Stroke, dengan nilai 100%, Presisi menunjukan rasio Pasien yang benar Stroke yang ada pada dataset dari keseluruhan Pasien yang diprediksi Stroke yaitu 96%, perbandingan rata-rata presisi dan recall yang dibobotkan (F1 Score) yang bernilai 98%. Untuk percobaan ke-18 memiliki nilai accuracy sebesar 99%, lalu recall 97%, Presisi 100%, F1 Score yang bernilai 98%. perhitungan dari Presisi, Recal, dan F1-Score diatas merupakan hasil perhitungan pada kelas 0 (pasien tidak terindikasi stroke). Untuk hasil Presisi, Recal, dan F1-Score pada kelas 1 (pasien terindikasi stroke) mendapatkan hasil yang baik juga yang dimana hasil presisi yang diperoleh 97%, Recal 100%, F1-Score 99%.

Hasil accuracy, presisi, recal, dan f1-score. dapat dihitung dengan menggunakannya hasil dari confusion matrix yaitu nilai True Positive (TP), True Negatif (TN), False Positif (FP), False Negative (TN).

Hasil akurasi percobaan ke-1 ini dapat diketahui dengan menggunakan rumus perhitungan sebagai berikut:

$$Accuracy\ (\%) = \frac{(TP+TN)}{(TP+TN+FP+FN)}\ x\ 100 = \frac{(1459+11)}{(1459+11+1+65)} = 95{,}7\%.$$

selain nilai akurasi juga dapat mengukur ukuran lainnya seperti Presisi, Recal, dan F1-Score, dengan menggunakan rumus perhitungan sebagai berikut:

$$Recall\ (\%) = \frac{(TP)}{(TP+FP)}\ x\ 100 = \frac{(1459)}{(1459+1)} = 99{,}9\%.$$

$$Precision\ (\%) = \frac{(TP)}{(TP+FN)}\ x\ 100 = \frac{(1459)}{(1459+65)}\ x\ 100 = 95{,}7\%.$$

$$F1\text{-}Score\ (\%) = 2\ x\ \frac{(Precision\ x\ Recall)}{(Precision\ +\ Recall)} = 98\%.$$

hasil dari Presisi, Recal, dan F1-Score diatas merupakan hasil yang paling baik pada kelas 0 (pasien tidak terindikasi stroke) dibandingkan dengan hasil dari kelas 1 (pasien terindikasi stroke).

Berikutnya Hasil akurasi percobaan ke-18 ini dapat diketahui dengan menggunakan rumus perhitungan sebagai berikut:

$$Accuracy\ (\%) = \frac{(TP+TN)}{(TP+TN+FP+FN)}\ x\ 100 = 99\%.$$

selain nilai akurasi juga dapat mengukur ukuran lainnya seperti presisi, recal, dan f1-score, dengan menggunakan rumus perhitungan sebagai berikut:

$$Recall\ (\%) = \frac{(TP)}{(TP+FP)}\ x\ 100 = 97\%.$$

$$Precision\ (\%) = \frac{(TP)}{(TP+FN)}\ x\ 100 = 100\%.$$

$$F1\text{-}Score\ (\%) = 2\ x\ \frac{(Precision\ x\ Recall)}{(Precision\ +\ Recall)} = 98\%.$$

Terdapat beberapa teknik, penggunaan metode klasifikasi dan penerapan metode untuk mengatasi imbalance kelas yang diimplementasikan kedalam experiment ini sehingga terjadi peningkatan akurasi yang lebih baik dari akurasi yang dihasilkan sebelumnya, teknik tersebut adalah teknik yang diterapkan dalam proses preprosessing data yaitu menggantikan nilai kosong yang terdapat pada

dataset, lalu untuk metode klasifikasi yang digunakan yaitu metode *Xgboost* dan mengatur parameter yang sesuai dengan pola yang terjadi dalam dataset agar menghasilkan model yang optimal, dan metode untuk dapat mengatasi imbalance kelas yaitu diterapkannya metode *smote*.

Preprosessing data dengan menerapkan teknik yaitu menggantikan nilai kosong pada atribut dalam dataset, dengan ini maka tidak lagi terdapat missing value dalam variabel dataset, jika terdapat missing value dalam dataset maka model klasifikasi akan menghasilkan output yang bias, tentunya dengan hal ini akan mengurangi hasil akurasi model dalam mengklasifikasikan penyakit stroke ini. adapun teknik lain untuk mengatasi missing value yaitu menghapuskan missing value pada atribut dataset, dengan teknik ini maka banyak data dalam dataset akan terbuang, dibandingkan dengan replace missing value tidak menghapus/membuang banyak data dalam dataset.

Metode *smote* untuk mengatasi imbalance kelas pada dataset, jika model melakukan klasifikasi pada data dengan jumlah kelas yang tidak seimbang maka model hanya dapat mengklasifikasi dengan benar pada kelas mayoritas dan ketika model mempredikasi kelas minoritas maka akan diprediksi sebagai kelas mayoritas, terdapat beberapa metode selain metode *smote* yang dapat mengatasi imbalance kelas yaitu under-sampling teknik dan over sampling teknik. under sampling teknik menghapus kelas mayoritas secara acak agar menyamakan kelas minoritas sedangkan over sampling teknik menduplikat kelas minoritas secara random agar menyamakan kelas mayoritas. dampak yang terjadi ketika digunakan

kedua metode tersebut adalah kehilangan banyak data, dan terjadinya overfitting dalam melakukan klasifikasi.

Terdapat beberapa kondisi yang terjadi pada dataset stroke prediction ini yang dapat menghasilkan klasifikasi menjadi tidak optimal seperti terdapat missing value pada atribut bmi yang berjumlah 202, Skala atribut dalam dataset yang memiliki perbedaan value yang jauh, dan imbalance kelas pada dataset, dengan kondisi yang terjadi tersebut maka perlu dilakukan teknik preprosessing data, normalisasi data, dan mengatasi imbalance kelas, agar data yang mengalami kesalahan, skala yang berbeda, dan imbalance kelas dapat teratasi. hasil dari klasifikasi ini menghasilkan klasifikasi yang optimal, ketika melakukan klasifikasi terhadap data yang sudah melewati teknik-teknik dan penerapan metode untuk melakukan perbaikan dataset dengan menghasilkan akurasi 99%.

Pengaruh fitur dalam melakukan prediksi apakah seseorang terindikasi stroke atau tidak dalam dataset, yang dapat dilihat pada gambar 4.22.


Data Visualization Result

	No Stroke	Yes Stroke	Note
gender (Most)	Female	Female	No Clear Difference
age (Median)	43	71	the median age of stroke patients is higher than patient with no stroke
hypertension (Most)	0	0	the patient who has hypertension from stroke patient is 13 % higher than the patient with no stroke
heart_disease (Most)	0	0	the patient who has heart disease from stroke patient is 14 % higher than the patient with no stroke
ever_married (Most)	Yes	Yes	the patient who ever married from stroke patient is 24 % higher than the patient with no stroke
work_type (Most)	Private	Private	the patient who work as self employed from stroke patient is 11.4% higher than the patient with no stroke
Residence_type (Most)	Urban	Urban	No Clear Difference
avg_glucose_level (Median)	91.5	105.2	the median of avg_glucose_level from Stroke Patient is higher than the Patient with no Stroke
bmi (Median)	28.3	30.5	the median of bmi from Stroke Patient is little higher than the Patient with no Stroke
smoking_status (Most)	never smoked	never smoked	The patient who smokes or formerly smoked from is 13% higher than the patient with no stroke

Gambar 4. 22 Pengaruh Fitur dalam prediksi penyakit.

Pada gambar 4.22 menjelaskan bahwa pengaruh fitur dalam dataset ketika melakukan prediksi penyakit stroke, dalam memprediksi seseorang berisiko penyakit stroke tidak dapat berpatokan terhadap fitur Umur(Gender) saja, untuk

fitur umur(Age) dapat dijadikan patokan dalam melakukan prediksi stroke, jika umur seseorang jauh lebih tua maka beresiko lebih tinggi terkena stroke, pada fitur tekanan darah tinggi(hypertension), jika seseorang mempunyai riwayat darah tinggi maka sekitar 18% berisiko terkena stroke, untuk penyakit jantung(heart_disease) jika seseorang mempunyai riwayat penyakit jantung maka 14% berisiko terkena stroke, lalu fitur pernah menikah(ever_married) meskipun didominasi pasien yang belum menikah namun jika seseorang pernah menikah maka sekitar 24% berisiko stroke, untuk tipe kerja(work_type) jika seseorang yang mempunyai pekerjaan wiraswasta maka sekitar 11,4% berisiko terkena stroke, lalu pada fitur tipe tempat tinggal(residence_type) meskipun didominasi type tempat tinggal urban tetapi tidak mempunyai pengaruh dalam mengetahui terkena stroke atau tidak, berikutnya fitur level penyakit gula(avg_glukosa_level) jika seseorang mempunyai tinggal gulah darah lebih tinggi maka dapat berisiko terkena stroke, selanjutnya fitur indeks masa tubuh(bmi) jika seseorang mempunyai nilai bmi nya lebih tinggi sedikit yaitu sekitar 30.5 maka berisiko terkena stroke, dan yang terakhir yaitu status merokok(smoking_status) jika pasien yang merokok maka sekitar 13% akan berisiko terkena stroke.

Dalam melakukan klasifikasi kemungkinan pasien terindikasi penyakit stroke atau tidak, metode yang digunakan yaitu metode *Xgboost,* sebelum masuk pada proses klasifikasi langkah utama yang dilakukan ialah tuning parameter yang dimana parameter metode diatur agar model menghasilkan hasil yang optimal. Paramater yang diatur adalah *max_depth, learning_rate, min_child_weight, n_estimator, random_state,* dan *subsample.*

dengan parameter-parameter tersebut diatur berdasarkan berjalannya pola pada dataset.

Tabel 4. 8 Perbandingan Penelitian.

Data Tidak Seimbang.			
Author	Dataset	Metode	Hasil Akurasi
Colak C, Dkk	Stroke Prediction, (Fedesoriano, n.d.).	Ann, Svm	93%
Ahmed H, dkk	Stroke Prediction, (Fedesoriano, n.d.).	Decision Tree, Svm, Random forest, Logistic regression.	90%
Nugroho .A.S	Stroke Prediction, (Fedesoriano, n.d.).	Knn, Fuzzy K-Nearest Neighbor	83%
Liu T, Dkk	Stroke Prediction, (Fedesoriano, n.d.).	DNN Berbasis Autohpo	72%.
Rohman R S, Dkk	Stroke Prediction, (Fedesoriano, n.d.).	C4.5, C4.5 Berbasis Pso, C4.5 Ga	92%
Huang M S N, Dkk	Stroke Prediction, (Fedesoriano, n.d.).	Rf,Svm	94%
Penelitian yang dilajukan	**Stroke Prediction, (Fedesoriano, n.d.).**	**Xgboost**	**96%**

Tabel 4. 8 Perbandingan Penelitian.(Lanjutan)

Data Seimbang			
Author	Dataset	Metode	Hasil Akurasi
Anas Faisal & Agus Subekti	Stroke Prediction data seimbang, (Fedesoriano, n.d.).	Deep Neural Network untuk Prediksi Stroke & Smote	96%
Nur Diana Saputri	Stroke Prediction data seimbang, (Fedesoriano, n.d.).	algoritma C4.5 dan C4.5 setelah menerapkan metode bagging & Smote	95%
Mutmainah. S	Stroke Prediction data seimbang, (Fedesoriano, n.d.).	Random Forest, Naïve Bayes,Knn, C4.5, SVM, Random Forest,CNN & Smote	95%
Penelitian yang diajukan	**Stroke Prediction data seimbang, (Fedesoriano, n.d.).**	**XGBoost & Smote**	**99%**

Pada tabel 4.8 merupakan tabel Perbandingan penelitian sebelumnya dengan penelitian ini. dataset yang diolah dalam penelitian sebelumnya dengan penelitian ini sama yaitu dataset stroke prediction yang belum dilakukan balancing dataset dan sudah dilakukan balancing dataset, terdapat perbedaan antara penelitian sebelumnya dengan penelitian yang dilakukan yaitu penggunaan metode klasifikasi. Metode klasifikasi yang digunakan antara lain yaitu Decision

Tree, Svm, Random forest, Logistic regression dll, yang dapat dilihat pada tabel 4.8.

Dari perbandingan tersebut ketika menggunakan metode Xgboost dalam melakukan klasifikasi terhadap data seimbang dan data tidak seimbang, memiliki hasil akurasi yang lebih baik dibandingkan dengan hasil akurasi yang dihasilkan oleh penelitian sebelumnya. hasil dari klasifikasi menggunakan metode Xgboost terhadap data tidak seimbang memiliki hasil akurasi yang lebih baik dari hasil akurasi penelitian sebelumnya yaitu sebesar 96%, dan terhadap data seimbang memiliki hasil akurasi yang lebih baik juga dari hasil akurasi penelitian sebelumnya yaitu sebesar 99%.

BAB V

PENUTUP

5.1. Kesimpulan

Berdasarkan hasil yang didapatkan pada proses analisa yang dilakukan sebelumnya maka dapat di tarik kesimpulan sebagai berikut :

a. Penelitian ini mendapatkan hasil akurasi yang lebih baik dari penelitian sebelumnya, ketika menggunakan metode xgboost saja, mampu menaikan akurasi 1% diatas hasil akurasi penelitian sebelumnya yaitu menjadi 96%.
b. Penelitian ini ketika menggunakan metode untuk mengatasi ketidakseimbangan kelas yaitu metode smote, dan untuk proses klasifikasi menggunakan metode Xgboost mampu menaikan akurasi 3% dari percobaan yang hanya menggunakan Xgboost tanpa smote, hasil akurasi yang dihasilkan pada penelitian ini adalah 99%.
c. Penerapan teknik normalisasi data dengan min max, maka setiap hasil percobaan yang dilakukan terjadi pengurangan hasil eror berupa fp, dengan itu mendapatkan peningkatan pada hasil akurasi sebesar 1%.

5.2. Saran

Pada penelitian ini melakukan pemilihan hyperparameter masih secara manual pada metode Xtreme Gradient Boosting (XGBoost) untuk proses klasifikasi, yang dimana akan terdapat banyak uji coba yang harus dilakukan agar mendapatkan hasil yang optimal untuk proses klasifikasi. Dengan ini saran yang diberikan adalah penerapan Grid Seacrh untuk menemukan parameter terbaik dalam suatu model dalam melakukan klasifikasi.

DAFTAR PUSTAKA

Rahim,A.M.A., Sunyoto,A., & Arif,R.M.,Abd Mizwar A. Rahim, Andi Sunyoto, M. R. A. (2022). Stroke Prediction Using Machine Learning Method with Extreme Gradient Boosting Algorithm. Matrik : Jurnal Manajemen, Teknik Informatika, Dan Rekayasa Komputer.

Agarwal, A. K., Wadhwa, S., & Chandra, S. (1994). Diagnosis of tuberculosis--newer tests. The Journal of the Association of Physicians of India, 42(8), 665.

Ahmed, H, Dkk,. (2019). Stroke prediction using distributed machine learning based on apache spark. International Journal of Advanced Science and Technology, 28(15), 89–97. https://doi.org/10.13140/RG.2.2.13478.68162

Aji Seto Arifianto, Moechammad Sarosa, O. S. (2014). Klasifikasi Stroke Berdasarkan Kelainan Patologis dengan Learning Vector Quantiation. Eeccis, 8(2), 117–122.

Mabrur, A, G, & L. R. (2012). Penerapan Data Mining Untuk Memprediksi Kriteria Nasabah Kredit. Jurnal Komputer Dan Informatika (KOMPUTA), 1(1), 53–57.

Azizah, N. (2021). Komparasi Metode Klasifikasi Decision Tree Algoritma C4.5 Dan Random Forest Untuk Prediksi Penyakit Stroke. Undergraduate Thesis, Sriwijaya University. https://repository.unsri.ac.id/58331/

Bunkhumpornpat, C., Dkk, (2012). DBSMOTE: Density-based synthetic minority over-sampling technique. Applied Intelligence, 36(3), 664–684. https://doi.org/10.1007/s10489-011-0287-y

I, Cholissodin, Dkk. (2017). Klasifikasi Tingkat Resiko Stroke Menggunakan Improved Particle Swarm Optimization dan Support Vector Machine. Konferensi Nasional Sistem & Informasi 2016, At STT Ibnu Sina, February, 11–13.

Claesen, M., & De Moor, B. (2015). Hyperparameter Search in Machine Learning. 10–14. http://arxiv.org/abs/1502.02127

Doreswamy, & Hemanth, K. S. (2011). Performance Evaluation of Predictive Engineering Materials Data Sets. Artificial Intelligent Systems Ans Machine Learning, 3(3), 1–8.

Agustin, S. (2021). No Title. Alodokter. https://www.alodokter.com/hati-hati-awalnya-stroke-ringan-selanjutnya-stroke

Fedesoriano. (n.d.). Stroke Prediction Dataset. https://www.kaggle.com/fedesoriano/stroke-prediction-dataset

Gorunescu, F. (2011). Data Mining: Concepts, models and techniques. https://link.springer.com/book/10.1007%2F978-3-642-19721-5

Handayani, A., Jamal, A., & Septiandri, A. A. (2017). 350-565-1-Sm. 6(4), 394–403.

ICHI.PRO. (2020). No Title. https://ichi.pro/id/apa-itu-xgboost-dan-bagaimana-cara-mengoptimalkannya-232831486673332

Junaidi, I. (2011). Stroke Waspadai Ancamannya, Yogyakarta: ANDI (M. S. Dra. Dorce Tandung (ed.)).

Kasanah, A. N, Dkk. (2019). Penerapan Teknik SMOTE untuk Mengatasi Imbalance Class dalam Klasifikasi Objektivitas Berita Online Menggunakan

Algoritma KNN. Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi), 3(2), 196–201. https://doi.org/10.29207/resti.v3i2.945

Kaur, Dkk. (2015). Review of decision tree data mining algorithms: ID3 and C4. 5. Proceedings of International Conference on Information Technology and Computer Science.

Kemenkes RI. (2018a). Hasil Riset Kesehatan Dasar Tahun 2018. Kementrian Kesehatan RI, 53(9), 1689–1699.

Kemenkes RI. (2018b). Stroke Dont Be The One (p. 10).

Kementerian Kesehatan Republik Indonesia. (2018). Hasil Utama Riskesdas 2018 Kesehatan. Riskesdas, 52. http://www.depkes.go.id/resources/download/info

Kovács, B, Dkk, (2020). Unfolding the effects of different forestry treatments on microclimate in oak forests: results of a 4-yr experiment. Ecological Applications, 30(2), 321–357. https://doi.org/10.1002/eap.2043

Mardi, Y. (2017). Data Mining : Klasifikasi Menggunakan Algoritma C4.5. Edik Informatika, 2(2), 213–219. https://doi.org/10.22202/ei.2016.v2i2.1465

Mo, H, Dkk, (2019). Developing window behavior models for residential buildings using XGBoost algorithm. Energy and Buildings, 205, 109564. https://doi.org/10.1016/j.enbuild.2019.109564

Mohan, S, Dkk, (2019). Effective heart disease prediction using hybrid machine learning techniques. IEEE Access, 7, 81542–81554. https://doi.org/10.1109/ACCESS.2019.2923707

Mutmainah, S. (2021). Penanganan Imbalance Data Pada Klasifikasi. 1, 10–16.

Naqa, I. El, & Murphy, M. J. (2015). Machine Learning in Radiation Oncology. Machine Learning in Radiation Oncology, 3–11. https://doi.org/10.1007/978-3-319-18305-3

Nugroho, S. A. J. I. (2020). Naskah publikasi perbandingan metode fuzzy k-nearest neighbor dan neighbor weighted k-nearest neighbor untuk deteksi penyakit stroke.

Permatasari, N. (2020). Perbandingan Stroke Non Hemoragik dengan Gangguan Motorik Pasien Memiliki Faktor Resiko Diabetes Melitus dan Hipertensi. Jurnal Ilmiah Kesehatan Sandi Husada, 11(1), 298–304. https://doi.org/10.35816/jiskh.v11i1.273

Phipps, M. S., & Cronin, C. A. (2020). Management of acute ischemic stroke. The BMJ, 368. https://doi.org/10.1136/bmj.l6983

Pierot, L, Dkk, (2018). Standards of practice in acute ischemic stroke intervention: International recommendations. American Journal of Neuroradiology, 39(11), E112–E117. https://doi.org/10.3174/ajnr.A5853

Putatunda, S., & Rama, K. (2018). A comparative analysis of hyperopt as against other approaches for hyper-parameter optimization of XGBoost. ACM International Conference Proceeding Series, 6–10. https://doi.org/10.1145/3297067.3297080

Riyantoko, P. A. (2021). Exploratory Data Analysis and Machine Learning Algorithms to Classifying Stroke Disease. March, 77–81.

Royal Society of Great Britain. (2017). Machine learning : the power and promise of computers that learn by example. In Report by the Royal Society (Vol. 66, Issue January).

S Mujiasih. (2011). Pemanfaatan Data Mining Untuk Prakiraan Cuaca. Jurnal Meteorologi Dan Geofisika, Vol. 12, N, 207–217.

Sabiq Sofyan, A. P. (2013). Penerapan Synthetic Minority Oversampling Technique (Smote) Terhadap Data Tidak Seimbang Pada Pembuatan Model Komposisi Jamu. Xplore: Journal of Statistics, 1(1), 868–877. https://doi.org/10.29244/xplore.v1i1.12424

Sailasya, G., & Kumari, G. L. A. (2021). Analyzing the Performance of Stroke Prediction using ML Classification Algorithms. International Journal of Advanced Computer Science and Applications, 12(6), 539–545. https://doi.org/10.14569/IJACSA.2021.0120662

Saleh, A. (2015). Implementasi Metode Klasifikasi Naïve Bayes Dalam Memprediksi Besarnya Penggunaan Listrik Rumah Tangga. Creative Information Technology Journal, 2(3), 207–217.

Septian, N. Y. (2009). Data Mining Menggunakan Algoritma Naïve Bayes Untuk Klasifikasi Kelulusan Mahasiswa Universitas Dian Nuswantoro. Jurnal Semantik 2013, 1–11.

Shameer, K., Johnson, K. W., Glicksberg, B. S., Dudley, J. T., & Sengupta, P. P. (2018). Machine learning in cardiovascular medicine: Are we there yet? Heart, 104(14), 1156–1164. https://doi.org/10.1136/heartjnl-2017-311198

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. Information Processing and Management, 45(4), 427–437. https://doi.org/10.1016/j.ipm.2009.03.002

Statistika, P. S., Matematika, F., Ilmu, D. A. N., Alam, P., & Indonesia, U. I. (2020). Xgboost Pada Klasifikasi Customer Churn.

Syahrani, I. M., Kusuma, W. A., & Wahjuni, S. (2019). Analisis Pembandingan Teknik Ensemble Secara Boosting(XGBoost) Dan Bagging (Random Forest) Pada Klasifikasi Kategori Sambatan Sekuens DNA. Jurnal Penelitian Pos Dan Informatika. https://doi.org/10.17933/jppi.2019.090103

Syarif, I., Zaluska, E., Prugel-Bennett, A., & Wills, G. (2012). Application of bagging, boosting and stacking to intrusion detection. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 7376 LNAI, 593–602. https://doi.org/10.1007/978-3-642-31537-4_46

Utomo, D. P., & Mesran, M. (2020). Analisis Komparasi Metode Klasifikasi Data Mining dan Reduksi Atribut Pada Data Set Penyakit Jantung. Jurnal Media Informatika Budidarma, 4(2), 437. https://doi.org/10.30865/mib.v4i2.2080

Yaman, E., & Subasi, A. (2019). Comparison of Bagging and Boosting Ensemble Machine Learning Methods for Automated EMG Signal Classification. BioMed Research International, 2019. https://doi.org/10.1155/2019/9152506