DIAGNOSA PENYAKIT *STROKE* MENGGUNAKAN METODE OPTIMASI BAGGING PADA ALGORITMA *NAIVE BAYES CLASSIFIER* BERBASIS *MACHINE LEARNING*

(Studi Kasus Penyakit Stroke)

PROPOSAL SKRIPSI

Oleh

Moh Fiki Budiono

182410103004

PROGRAM STUDI INFORMATIKA

FAKULTAS ILMU KOMPUTER

UNIVERSITAS JEMBER

2022

A. JUDUL

Diagnosa Penyakit *Stroke* Menggunakan Metode Optimasi Bagging Pada Algoritma *Naive Bayes Classifier* Berbasis *Machine Learning* (Studi Kasus Penyakit Stroke)

B. LATAR BELAKANG

Stroke merupakan salah satu penyebab kematian terbesar yang ada di dunia dimana penyakit menempati urutan kedua dari faktor penyebab kematian setelah jantung, di mana tanda tanda penyakit ini sering muncul secara tiba tiba tetapi banyak yang menyerang secara mendadak. Pada tahun 2016 pernah terjadi peningkatan pada pasien *stroke* yang menyebabkan beban besar pada sistem perawatan kesehatan di negara Amerika Serikat (Ali, 2019). Sedangkan di china stroke merupakan penyebab utama kematian studi prediktif telah menemukan bahwa usia di atas 25 tahun beresiko tinggi terkena *stroke* (Wu et al*,* 2020). Tahun 2019 WHO mengatakan tujuh dari sepuluh penyebab kematian yang ada di dunia disebabkan oleh stroke, Kementerian kesehatan mengelompokkan *stroke* sebagai penyakit katastropik karena dampaknya yang luas secara ekonomi dan digital (Faisal et al, 2021). *Stroke* menjadi penyebab pertama kematian yang diprediksi pada tahun 2030 (14,4% dari total kematian)dan penyebab ketiga DALY lost (6% dari total DALY) di negara-negara berpenghasilan menengah (Byna et al, 2020). *Stroke* merupakan masalah yang cukup serius karena serangan *stroke* sebagai keadaan darurat medis yang dapat mengancam kecacatan dan kematian pada pasien jika tidak cepat dan tepat dalam penanganannya (Sakinah et al*,* 2020). Berdasarkan Laporan hasil Riskesdas tahun 2018 *Prevalensi stroke* nasional menyentuh 10,9% hal tersebut membuat penyakit ini menjadi pembunuh nomor 3 di Indonesia (Suryani et al, 2022).

Stroke dianggap sebagai masalah kesehatan yang parah karena tingkat kematiannya yang tinggi, Selain itu juga biaya perawatan *stroke* sangat tinggi sehingga menimbulkan efek negatif bagi perekonomian. Dengan adanya permasalahan diatas maka perlu diciptakannya teknologi canggih yang dapat membantu dalam diagnosis klinis, pengobatan, prediksi kejadian klinis. Deteksi dini *stroke* merupakan langkah penting untuk pengobatan yang efisien dan ML dapat menjadi nilai yang besar dalam proses ini. Untuk dapat melakukan itu, *Machine Learning* (ML) adalah teknologi terbaik yang dapat membantu profesional kesehatan membuat keputusan dan prediksi klinis (Byna et al, 2020). Kompleksitas kondisi seperti *stroke* berpotensi cocok untuk penggunaan metode ML yang mampu menggabungkan berbagai macam variabel dan pengamatan ke dalam satu kerangka prediktif tanpa memerlukan aturan yang telah diprogram sebelumnya (Wang, W et al, 2020). Ada peningkatan minat dalam

penggunaan ML untuk memprediksi hasil stroke, dengan harapan bahwa metode tersebut dapat menggunakan kumpulan data besar yang dikumpulkan secara rutin dan memberikan prognosis pribadi yang akurat.

Seperti penelitian yang dilakukan Ali, A. A. (2019) mengenai "Stroke prediction using distributed machine learning based on Apache spark Stroke" penelitian ini membandingkan berbagai algoritma pembelajaran untuk prediksi penyakit stroke pada Healthcare Dataset Stroke, Algoritma pembelajaran yang diterapkan pada penelitian ini *Decision Tree*, *Support, Vector Machine, Random Forest Classifier*, dan *Logistic Regression*. *Random Forest Classifier* merupakan algoritma pembelajaran yang terbaik dalam penelitian ini dengan akurasi sebesar 90%. Kemudian penelitian yang berjudul "Machine learning–based model for prediction of outcomes in acute stroke" Penelitian ini bertujuan untuk memprediksi pasien yang terkena *stroke* dengan membandingkan berbagai algoritma pembelajaran diantaranya *deep neural network,random forest, and logistic regression*. Dalam penelitian ini data yang diolah dalam penelitian ini sebanyak 2604 dimana dalam penelitian ini model *deep neural network* memiliki tingkat akurasi yang paling baik (Heo, J et al, 2019). Selanjutnya penelitian yang dilakukan oleh Wu, Y., & Fang, Y (2020) yang berjudul "Stroke prediction with machine learning methods among older Chinese International journal of environmental research and public health" Penelitian ini berfokus pada prediksi *stroke* dengan menggunakan data yang tidak seimbang pada populasi lansia di china, data yang digunakan pada penelitian ini sebesar 1131 peserta dengan target 56 pasien *stroke* dan 1075 pasien non stroke yang diambil pada tahun 2012-2014. Metode yang digunakan pada penelitian ini *regularized logistic regression* (RLR), *support vector machine* (SVM), and *random forest* (RF). dari ketiga metode tersebut metode *support vector machine* (SVM) menunjukan hasil yang paling baik.

Kemudian penelitian yang dilakukan Saputri, N. D (2021) pada penelitian ini mengaplikasikan metode klasifikasi algoritma C4.5 serta metode bagging dan adaboost dari *Ensemble Learning* yang diterapkan untuk prediksi Penyakit Stroke. Hasil percobaan menunjukkan bahwa penambahan bagging pada algoritma C4.5 mengalami peningkatan 3% lebih unggul dari algoritma yang dikombinasikan dengan adaboost dengan dengan peningkatan 2%. Kemudian penelitian yang dilakukan oleh Rahmadani et al, (2022) dalam penelitian ini mendeteksi *stroke* dengan bantuan Algoritma C4.5 dan CART untuk mendapatkan hasil yang terbaik penelitian ini membagi data training dan testing menjadi dua bagian 60:40 dan 70:30, dimana hasil terbaik dalam penelitian ini menggunakan pembagian data 60:40 dengan hasil akurasi sebesar 96% pada algoritma Algoritma C4.5 sedangkan perbandingan data 70:30 menghasilkan tingkat akurasi sebesar 95.76%. Selanjutnya penelitian yang dilakukan

oleh Rachmad, D. U. M., Oktavianto, H., & Rahman, M. (2022) dengan judul "Perbandingan Metode *K-Nearest* yang menggunakan penambahan metode optimasi bagging Dan *Gaussian Naive Baye*s Untuk Klasifikasi Penyakit Stroke" pada penelitian ini metode *Naive Bayes* memiliki tingkat akurasi lebih tinggi yaitu sebesar 74,45% sedangkan *K-Nearest Neighbor* memiliki tingkat akurasi sebesar 68,30% dimana dari perbandingan metode tersebut metode *Naive Bayes* lebih tinggi akurasinya sebesar 6,15%.

Tujuan dari penelitian ini untuk menambah literatur di dunia penelitian, maka dari itu data yang diujikan harus sama dengan beberapa penelitian sebelumnya. Penelitian ini ingin meneliti tingkat keakuratan metode *naive bayes classifier* dengan menambahkan metode bagging untuk menguji tingkat akurasi pada dataset yang sudah ditentukan penggunaannya, dengan perbandingan data testing dan data training yang sama. Dimana penambahan metode optimasi bagging dalam penelitian sebelumnya dapat menambah tingkat keakuratan suatu penelitian.

C. RUMUSAN MASALAH

Ditinjau dari permasalahan yang diuraikan dari latar belakang diatas, maka dapat dirumuskan permasalahan pada penelitian ini sebagai berikut :

1. Bagaimana penerapan metode *Naive Bayes Classifier* pada sistem diagnosa penyakit *stroke* ?
2. Bagaimana tingkat akurasi penggunaan *Naive Bayes Classifier* yang menggunakan penambahan metode optimasi bagging dalam mendiagnosa penyakit *stroke* ?

D. TUJUAN

Tujuan yang ingin dicapai dari penelitian ini adalah sebagai berikut:

1. Untuk membuktikan efisiensi metode *Naive Bayes Classifier* yang diterapkan pada sistem diagnosa penyakit stroke
2. Untuk mengetahui tingkat akurasi metode *Naive Bayes Classifier* yang menggunakan tambahan metode optimasi bagging

E. MANFAAT

Hasil penelitian ini diharapkan memiliki manfaat bagi penulis, keilmuan dan bagi puskeswan. Adapun manfaat yang didapat dari penelitian ini adalah sebagai berikut:

1. Bagi peneliti
    a) Mendapatkan pengalaman dalam mengimplementasikan materi selama kuliah di Universitas Jember
    b) Dapat mengetahui hasil dari penelitian dengan metode *Naive Bayes*

Classifier dalam menentukan penyakit stroke

2. Bagi keilmuan
   a) Dapat digunakan sebagai referensi dalam melakukan pengembangan penelitian yang sejenis dan dapat memberikan kontribusi dalam menerapkan teori yang berkaitan dengan persoalan yang diangkat oleh penelitian

F. BATASAN MASALAH

Adapun batasan-batasan masalah pada penelitian yang dilakukan ini diantaranya adalah sebagai berikut:

1. Data yang diambil dalam penelitian ini diperoleh melalui penyedia basis data Kaggle yang sudah divalidasi oleh Fedesoriano
2. Pengujian data pada penelitian ini menggunakan metode *Naive Bayes Classifier* dengan menggunakan bahasa pemrograman *Python*

G. TINJAUAN PUSTAKA

1. Penelitian Terdahulu

Pelaksanaan penelitian ini tidak lepas dari penelitian terdahulu yang hasilnya dapat digunakan untuk menunjang penelitian saat ini, baik dari segi gagasan ide ataupun metode. Adapun penelitian terdahulu yang digunakan peneliti dalam melakukan penelitian ini adalah sebagai berikut:

Penelitian terdahulu dilakukan oleh Ali, A. A. (2019) dengan judul "Stroke prediction using distributed *machine learning* based on Apache spark" penelitian ini membandingkan berbagai algoritma pembelajaran *machine learning* untuk prediksi penyakit stroke pada Healthcare Dataset Stroke penelitian ini dibantu dengan platform data besar yaitu Apache Spark, Dimana platform ini sudah terintegrasi Spark yang menyediakan berbagai algoritma pembelajaran. Ada 4 jenis algoritma pembelajaran yang diterapkan pada penelitian ini *Decision Tree, Support, Vector Machine, Random Forest Classifier*, dan *Logistic Regression. Random Forest Classifier* merupakan algoritma pembelajaran yang terbaik dalam penelitian ini dengan akurasi sebesar 90%.

Selanjutnya penelitian yang dilakukan oleh Heo, J., Yoon, J. G., Park, H., Kim, Y. D., Nam, H. S., & Heo, J. H. 2019 dengan judul "Machine learning–based model for prediction of outcomes in acute stroke". Tujuan untuk memprediksi pasien yang terkena stroke dengan membandingkan berbagai algoritma pembelajaran diantaranya *deep neural network, random forest, and logistic regression*. Dalam penelitian ini data yang diolah dalam penelitian ini sebanyak 2604 pasien dan 67% digunakan sebagai data training dan 33% digunakan sebagai data uji, sebanyak 2043 memiliki hasil yang baik dimana dalam penelitian ini model *deep neural network* memiliki tingkat akurasi yang paling baik.

Penelitian terdahulu dilakukan oleh Wu, Y., & Fang, Y. 2020 dengan judul "Stroke

prediction with machine learning methods among older Chinese. International journal of environmental research and public health" Wu, Y., & Fang, Y. (2020). Penelitian ini berfokus pada prediksi stroke dengan menggunakan data yang tidak seimbang pada populasi lansia di china, data yang digunakan pada penelitian ini sebesar 1131 peserta dengan target 56 pasien stroke dan 1075 pasien non stroke yang diambil pada tahun 2012 - 2014. Metode yang digunakan pada penelitian ini *regularized logistic regression* (RLR), *support vector machine(SVM)*, and *random fores*t (RF) yang dimana metode SVM merupakan metode pembelajaran mesin yang kuat dan itu juga menunjukkan kinerja yang sangat baik. terdapat 15 variabel dalam tiga kategori pada tahun 2012 dipilih sebagai prediktor dalam penelitian ini, termasuk variabel demografis, seperti jenis kelamin, usia, dan penyakit penyerta (hipertensi, diabetes, dan penyakit jantung); variabel gaya hidup, seperti merokok dan minum; dan variabel klinis termasuk tekanan darah *sistolik*, tekanan darah *diastolik*, protein C-reaktif sensitivitas tinggi, glukosa darah, kolesterol lipoprotein densitas tinggi, kolesterol lipoprotein densitas rendah, trigliserida , dan asam urat.

Kemudian penelitian terdahulu yang dilakukan oleh Sakinah, N., Badriyah, T., & Syarif, I. 2020 dengan judul "Analisis Kinerja Algoritma Mesin Pembelajaran untuk Klasifikasi Penyakit Stroke Menggunakan Citra CT Scan" Penelitian ini juga bertujuan untuk membandingkan kinerja lima algoritma mesin pembelajaran yaitu *Naïve Bayes, Logistic Regression, Neural Network, Support Vector Machine* dan *Deep Learning* yang diterapkan untuk memprediksi penyakit stroke. Dimana data yang digunakan dalam penelitian ini diambil dari Rumah Sakit Umum Haji Surabaya yang diambil selama periode Januari-Mei 2019 dan berasal dari 102 pasien yang terindikasi stroke. Hasil percobaan menunjukkan bahwa algoritma Deep Learning menghasilkan tingkat performansi paling tinggi yaitu nilai akurasi 96.78%, presisi 97.59% dan recall 95.92%.

Selanjutnya penelitian terdahulu yang dilakukan oleh Saputri, N. D, 2021 dengan judul "Komparasi penerapan metode Bagging dan Adaboost pada Algoritma c4. 5 untuk prediksi Penyakit Stroke". penelitian ini menerapkan metode klasifikasi algoritma C4.5 serta metode bagging dan Adaboost dari Ensemble Learning. k-fold untuk menemukan nilai TP, TN, FP, FN, recall, precision, dan akurasi peneliti menggunakan metode confusion matrix. Hasil dari pengujian menggunakan algoritma C4.5 menghasilkan akurasi sebesar 92.87%. Kemudian hasil akurasi dari algoritma C4.5 dengan metode bagging meningkat menjadi 95.02% dan ketika dikombinasikan dengan metode Adaboost nilai akurasinya juga meningkat menjadi 94.63%.

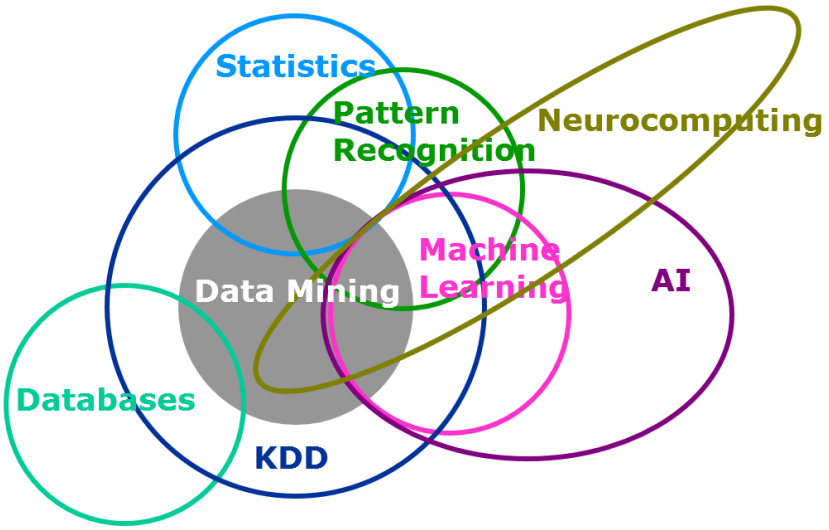
2. Machine Learning

Machine learning merupakan sub bagian dari kecerdasan buatan (Artificial intelligence) yang dapat mengambil keputusan tanpa diprogram *explicit* dengan

memanfaatkan data yang telah dibangun (Ibnu Daqiqil, 2021). *Machine learning* merupakan bidang ilmu komputer yang mana terjadi proses pembelajaran untuk memahami input yang tidak diketahui sebelumnya, sehingga memiliki hasil sebuah pola yang dapat dimengerti. Dimana algoritma dalam mesin pembelajaran ini bersifat *generic.* dimana dalam proses penyelesaian masalah tersebut algoritma akan menghasilkan sebuah data yang menarik tanpa menulis kode yang spesifik.

Machine learning beroperasi untuk membangun model dari banyak inputan untuk membuat prediksi atau keputusan. Algoritma *machine learning* ini akan mempertimbangkan semua titik data dalam kumpulan data tanpa bias manusia karena pengetahuan sebelumnya. Dikarenakan dalam penyelesaian masalah *machine learning* memiliki peramalan yang metode pengukuran berbeda tergantung pada tugas yang diberikan kepada sistem. Dimana tujuan *machine learning* diciptakan untuk membangun sebuah sistem yang dapat diterapkan ke dalam komputer untuk mempelajari data yang nantinya siap digunakan untuk memecahkan kasus tertentu berdasarkan jenis data (Id, I. D, 2021).

Klasifikasi *machine learning* berdasarkan cara belajar dalam mengolah suatu data dikelompokan menjadi tiga meliputi *Supervised Learning, Unsupervised Learning, dan Reinforcement Learning* (Id, I. D, 2021). Berikut ini gambar hubungan AI dengan *machine learning.*



Gambar 1 Posisi *Machine Learning* (Sumber : Mitchell-Guthrie, 2014)

Supervised Learning

Pembelajaran terarah atau terawasi dimana dalam proses prosesnya sistem ini akan mempelajari data training yang berisi label. Salah satu rumusan standar tugas *supervised*

learning yaitu masalah klasifikasi dan regresi. Adapun beberapa contoh algoritma yang termasuk dalam kategori ini diantaranya : *Random Forests*, *Linear Regression* (Regresi Linear), *Naive Bayes Classifier*, *k-Nearest Neighbors*, *Support Vector Machines* (SVMs), *Neural networks*, dan *Linear Regression* (Regresi Linear).

- *Unsupervised Learning*

Algoritma yang prosesnya dilakukan tanpa adanya petunjuk dari awal dimana algoritma dalam komputer yang belajar untuk menemukan pola data tersebut. dalam penerapanya algoritma *unsupervised learning* dibagi menjadi dua jenis yaitu asosiasi dan *clustering*. Masalah pengelompokan (*clustering*) adalah tempat untuk menemukan pengelompokan yang melekat dalam data, seperti mengelompokkan benda berdasarkan warna dan bentuk. Sedangkan masalah asosiasi adalah aturan yang menggambarkan sebagian besar data yang ada, seperti orang yang membeli A juga cenderung membeli B. Adapun contoh algoritma *unsupervised learning* sederhana adalah *K-means.* (Roihan, 2020).

- · *Reinforcement Learning*

Reinforcement learning merupakan *machine learning* dimana sebuah komputer dapat mengambil keputusan untuk memaksimalkan beberapa gagasan tentang imbalan kumulatif di lingkungan yang sangat dinamis. sebagai contoh seperti pada permainan catur atau *self-driving*.

Klasifikasi *machine learning* berdasarkan cara kerjanya dalam mengolah suatu data dikelompokan dua diantaranya: *Instance-based learning* dan *Model based learning* (Id, I. D, 2021).

- *Instance-based learning*

Instance-based learning yang memiliki cara kerja dengan cara membandingkan data testing yang telah dipelajari pada saat proses training sehingga algoritma ini sering disebut juga *memory-based learning* yaitu membandingkan data yang telah disimpan di memori. Contoh algoritma ini *k-nearest neighbors, kernel machines*, dan RBF network.

- *Model based learning*

Model based learning merupakan *machine learning* yang kerjanya kebalikan dari *Instance-based learning.* dimana algoritma ini menggunakan memori dalam memecahkan masalah, algoritma ini membuat model yang bersifat generik.

3. Metode *Naive Bayes Classifier*

Metode *Naive Bayes Classifier merupakan* metode pengklasifikasian dengan cara menghitung probabilitas dalam menentukan suatu kelas, teorema ini dikemukakan oleh ilmuwan Inggris, Thomas Bayes. Dalam menentukan sebuah kategori kelas yang optimal, teorema Bayes ini menghitung nilai dari atribut setiap kelas kemudian diklasifikasikan, setelah itu nilai tersebut masuk dapat di kategori kelas yang paling optimal yang dimana setiap kelas sudah memiliki nilainya masing masing. Dalam metode ini terdapat dua proses *learning/training dan testing/classify.* Pada proses *learning* sebagian data yang sudah diketahui kategorinya diumpankan dalam bentuk model perkiraan. Kemudian dalam proses *testing* model yang sudah dibentuk diuji dengan sebagian data lainya untuk mengetahui tingkat akurasi dari model tersebut (Byna et all, 2021), Berikut ini perhitungan metode *Naive Bayes Classifier* persamaan (1)

$$P(X|H) = \frac{P(X|H)}{P(X)} . P(H)$$

Keterangan

X : Vektor Inputan (data kelas yang belum diketahui) H :
kelas spesifik (hipotesis data)
P(H|X) : Probabilitas hipotesis kelas H pada kondisi X (posteriori probability) P(H) :
Probabilitas hipotesis H (prior probabilitas)
P(X|H) : Probabilitas hipótesis P(X) :
Probabilitas X

Pada saat proses pengklasifikasian data, metode *Naive Bayes Classifier* ini membutuhkan beberapa tahapan untuk menentukan kelas yang sesuai untuk proses menganalisis sampel data yang digunakan dalam penelitian. Sehingga persamaan (1) disesuaikan dengan persamaan (2).

$$P (C|F1….Fn) = \frac{P(C) . P(F1….Fn|C)}{P (F1….Fn)} . P (H)$$

Pada persamaan diatas (2), Variabel C merupakan sebuah kelas dan variabel F1...Fn adalah sebuah atribut karakteristik dari petunjuk yang digunakan pada saat proses klasifikasi sebuah data. Probabilitas munculnya kelas C (sebelum masuknya sampel disebut prior) dipengaruhi dari probabilitas masuknya sampel karakteristik tertentu pada kelas C (posterior), kemudian kelas tersebut dikalikan dengan probabilitas munculnya karakteristik dari sebuah sampel pada kelas C (likelihood), setelah itu dibagi

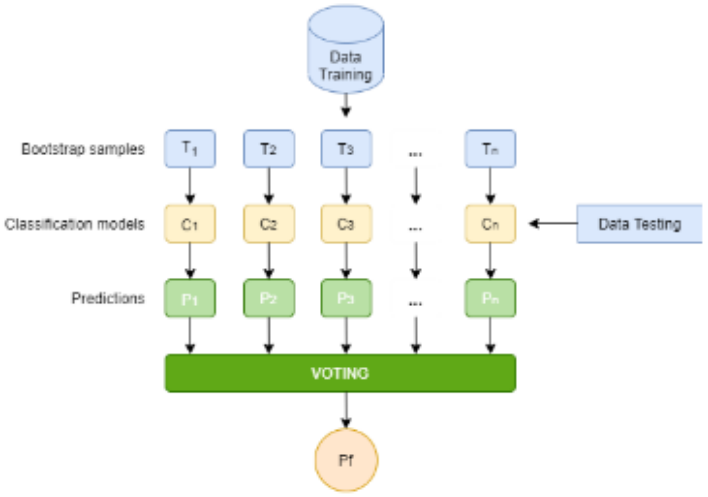
dengan probabilitas munculnya karakteristik pada sampel secara global (evidence). Pada persamaan (3) penggunaan rumus sederhana pada metode *Naive Bayes Classifier*

$$Posterior = \frac{Prior\ x\ Likelihood}{Evidence}$$

Untuk menentukan termasuk kelas apa sebuah sampel diklasifikasikan yaitu dengan cara membandingkan nilai *posterior* yang ada dalam setiap kelas, Dimana dalam satu sampel terdapat sebuah kelas dimana kelas tersebut sudah memiliki nilai *evidence* yang selalu tetap dan tidak bisa diubah.

4. Bagging

Bagging adalah singkatan dari bootstrap aggregating, menggunakan sub-dataset (bootstrap) untuk menghasilkan set pelatihan L (learning), L melatih dasar belajar menggunakan prosedur pembelajaran yang tidak stabil, dan kemudian selama pengujian, mengambil rata-rata. Bagging adalah salah satu teknik dari ensemble method dengan cara memanipulasi data training, data training di duplikasi sebanyak d kali dengan pengembalian (sampling with replacement), yang akan menghasilkan sebanyak data training yang baru, kemudian dari data training tersebut akan dibangun classifiers classifier yang disebut sebagai bagged classifier.). Hal inilah yang menjadi kunci kenapa bagging bisa meningkatkan akurasi karena dengan sampling with replacement dapat memperkecil variance dari dataset. Kurniawan, A., & Prihandono, A. (2021). Dalam beberapa penelitian penggunaan metode ansambel sebagai algoritma pembelajaran memberikan hasil yang sangat baik dalam melakukan dalam melakukan prediksi (Id, I. D, 2021). Tingkat akurasi pada metode bagging ini merupakan nilai tertinggi dari *voting* dari setiap data yang di training Nugroho, A., & Religia, Y. (2021).



Gambar 1 cara kerja bagging

5. Stroke

Stroke merupakan penyakit yang tidak menular dimana terjadi karena gangguan fungsional otak yang disebabkan tersumbatnya aliran darah ke otak dimana suplai darah yang terhenti secara mendadak dapat mengakibatkan stroke, tanpa suplai darah sel sel yang ada di otak akan perlahan mati (Dritsas, E., & Trigka, M. 2022). Penyakit ini tergolong ke dalam cerebrovascular disease karena membutuhkan penanganan selama 24 jam, jika tidak ditangani secara cepat dapat menyebabkan kematian (Saputri, 2021). Pasien yang terkena stroke harus segera ditangani secepatnya karena sel otak dapat mati dalam hitungan menit, Tindakan penanganan stroke secara cepat dan tepat dapat mengurangi resiko kerusakan otak dan mencegah terjadinya komplikasi. Stroke terdiri dari sekelompok gangguan heterogen yang ditandai dengan gangguan mendadak dan fokal dari suplai vaskular otak, menyebabkan gejala neurologis yang menetap lebih dari 24 jam. Secara umum stroke dapat diklasifikasikan menjadi iskemik dan hemoragik. Stroke iskemik terjadi ketika pembuluh darah tersumbat oleh trombus atau embolus, yang mengakibatkan iskemia otak. Stroke hemoragik disebabkan oleh pecahnya dan pendarahan pembuluh darah yang melemah ke jaringan otak di sekitarnya, yang biasanya menyebabkan tekanan intrakranial. pengidap stroke hemoragik cenderung mengkonsumsi alkohol dan rokok. Sedangkan iskemik, lebih cenderung karena diabetes. Ada banyak faktor risiko yang dapat memicu stroke, Set pertama faktor risiko termasuk usia, jenis kelamin, ras dan etnis. Sebaliknya, faktor risiko yang dapat dimodifikasi terkait dengan kondisi klinis, seperti penyakit jantung (misalnya, hipertensi, fibrilasi atrium, hiperkolesterolemia dan diabetes mellitus, serta faktor gaya hidup yaitu sedentarisme, obesitas, gizi buruk, penggunaan tembakau dan konsumsi alkohol (Sirsat et al, 2020).

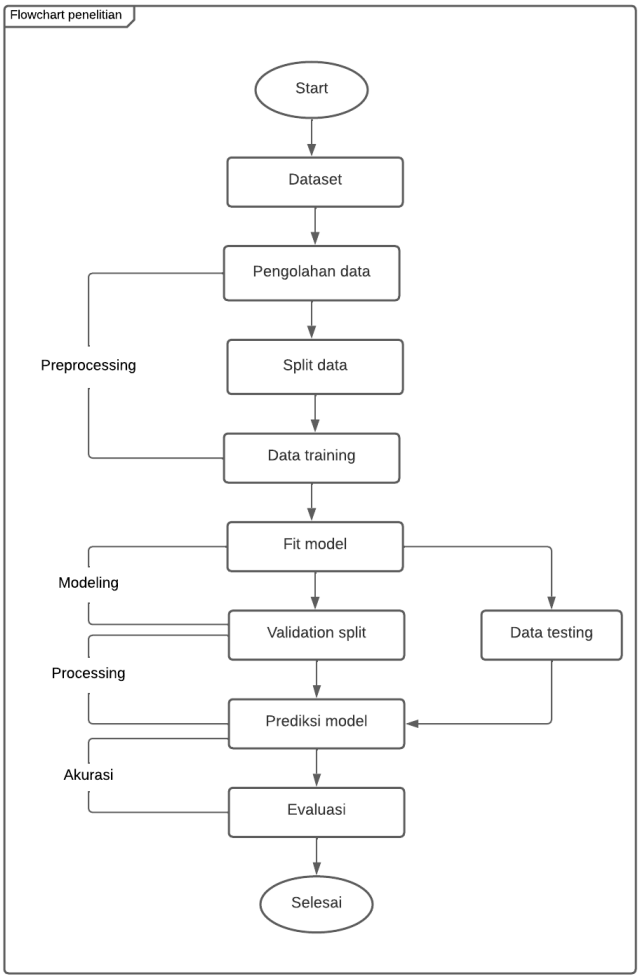
H. METODOLOGI PENELITIAN

H 1. Jenis Penelitian

Jenis Penelitian ini termasuk dalam jenis penelitian terapan *(applied research)* penelitian terapan merupakan jenis penelitian yang dikembangkan untuk memecahkan permasalah praktis bukan ke arah teoritis di kehidupan nyata dimana penelitian itu dilakukan, Penelitian terapan memiliki tujuan untuk menyelesaikan suatu permasalahan serta menghasilkan sesuatu yang berguna bagi kehidupan manusia pada bidang tertentu.

H 2. Tahapan Penelitian

Penelitian ini akan dilakukan dalam 5 tahapan, dimulai dari pengumpulan data, Preprocessing, Modeling, Prosesing, dan Akurasi metode yang digunakan dan pengembangan sistem.



Gambar 2 Flowchart tahapan penelitian

H 4.1 Dataset

Penelitian menggunakan dataset stroke prediction yang diunduh dari Kaggle.com. jumlah keseluruhan data sebanyak 5110 terdiri dari 12 atribut dan 1 target diantaranya Id, Jenis Kelamin, Umur, Hipertensi, Penyakit Jantung, Pernah Menikah, Tipe Pekerjaan, Tipe Tempat Tinggal, Kadar Glukosa, BMI, Status Merokok, Stroke. Dimana data tersebut masih belum seimbang dikarenakan jumlah data pasien stroke tidak sama sebanyak 249 stroke dan 4861 data non stroke, dimana data tersebut akan dibagi menjadi dua yaitu data training dan data testing pada tahap preprocessing.

Tabel 1.1 Atribut Deskripsi dan Tipe Data Stroke

Atribut	**Deskripsi**	**Tipe Data**
Id	Kode pengenal	Numerik
Jenis Kelamin	Pria,Wanita	Kategorikal
Umur	Usia Pasien (Tahun)	Numerik

Atribut	Deskripsi	Tipe Data
Id	Kode pengenal	Numerik
Hipertensi	Ya,Tidak	Kategorikal
Penyakit Jantung	Ya,Tidak	Kategorikal
Pernah Menikah	Ya,Tidak	Kategorikal
Tipe Pekerjaan	Pekerja pemerintah, Tidak Bekerja, Pribadi, Wiraswasta, Anak-anak	Kategorikal
Tipe Tempat Tinggal	Perkotaan, Pedesaan	Kategorikal
Kadar Glukosa	Kadar Gula darah (ml/g)	Numerik
BMI	Indeks Massa Tubuh	Numerik
Status Merokok	Tidak Diketahui, Sebelumnya Perokok, Tidak Pernah Merokok, Merokok	Kategorikal
Stroke	Tidak Stroke, Stroke	Kategorikal

H 4.2 Tahap preprocessing

Tahap preprocessing dimana tahap ini bertujuan untuk meningkatkan hasil kinerja metode yang digunakan. dalam penelitian ini tahapan preprocessing dibagi menjadi dua tahap yaitu tahap pengolahan data dan split data, kedua pembagian tersebut memiliki fungsi tersendiri , diantaranya :

1. Tahapan pengolahan data merupakan tahap mengubah data mentah khususnya data yang bersifat kategorik menjadi data yang dapat dengan mudah diolah dengan proses data mining. Banyak data noise seperti tidak ada keterangan “N/A” pada variabel dan BMI dan data kosong tanpa keterangan pada variabel merokok serta data. Untuk mendapatkan nilai dari data kosong pada label IBM dan Status merokok perlu diterapkan maen untuk mengisi data yang hilang. kemudian bagian data yang masih bersifat kategorikal diubah menjadi data numerik, dimana dalam merubah tipe data tersebut menggunakan *OneHotEncoder* pada library *Scikit-learn* yang ada dalam bahasa pemrograman *python.*
2. *Split* data dimana dalam proses ini data akan dibagi menjadi dua menggunakan teknik *split validation*, dimana dataset yang disajikan diatas akan dibagi kedalam dua bagian yakni 70% dari database akan dijadikan sebagai data training dan 30% sisanya akan dijadikan sebagai data testing atau uji.

H 4.3 Modeling dan Processing

Proses *machine learning* yang dilakukan yaitu menggunakan algoritma *Naive Bayes Classifier* yang dibuat menggunakan bahasa pemrograman *python*, algoritma ini mampu mengatasi masalah dengan cara membandingkan probabilitas data satu dengan data lainya. tahap ini setelah data pada proses preprocessing sudah dijalankan, kemudian akan berlanjut pada proses pengklasifikasian menggunakan algoritma *naive bayes* dengan menggunakan library *Scikit-learn.* Pada tahap ini pengujian dilakukan dua kali yang pertama murni menggunakan menggunakan algoritma *Naive Bayes Classifier* pengujian yang kedua menggunakan algoritma *Naive Bayes Classifier* yang dioptimalkan dengan metode bagging dengan harapan memperoleh akurasi yang lebih akurat. berikut ini skenario pembagian klasifikasi model:

Tabel 1.4 Proses Skenario Klasifikasi

Naive Bayes Classifier	*Naive Bayes Classifier* + Bagging
Pada skenario pertama akan dilakukan pengujian prediksi penyakit stroke menggunakan algoritma *Naive Bayes Classifier* saja tanpa menerapkan metode bagging	Pada skenario kedua akan dilakukan pengujian menggunakan algoritma *Naive Bayes Classifier* dengan tambahan metode bagging untuk menghasilkan klasifikasi yang lebih akurat

H 4.4 Pengujian

Pada tahapan pengujian penguji melakukan percobaan terhadap program yang telah dibuat, percobaan (testing) akan dilakukan dengan percobaan menginputkan gejala-gejala dan melihat nilai kebenaran dari hasil analisa sistem. Pada tahapan ini akan dilakukan pengujian akurasi.

Pada tahapan tahap uji akurasi dilakukan dengan cara mengkonversi data dari hasil klasifikasi menggunakan algoritma tersebut dengan menggunakan tabel *confusion matrix* agar mendapatkan nilai accuracy (akurasi), precision dan recall (Sakinah, 2020), Rumus tabel *confusion matrix* sebagai berikut :

Tabel 1.4 *Ilustrasi tabel confusion matrix*

	Nilai aktual positif	Nilai aktual negatif
Nilai prediksi positif	TP	FP
Nilai prediksi negatif	FN	TN

Keterangan :

TP (True positif) : Jumlah data positif yang benar diprediksi positif FP

(False positif) : Jumlah data negatif yang bernilai positif

TN (True negatif) : Jumlah data positif yang benar diprediksi negatif TP (False negatif) : Jumlah data negatif yang bernilai positif

1. Menghitung tingkat akurasi

   Nilai precision merupakan perbandingan probabilitas nilai positif dengan banyaknya data yang diprediksi positif :

$$\text{Accuracy} = \frac{TP + TN}{TP+FP+FN+TN}$$

2. Menghitung nilai precision

   Nilai precision merupakan perbandingan probabilitas nilai positif dengan banyaknya data yang diprediksi positif :

$$\text{Precision} = \frac{TP}{TP+FP}$$

3. Menghitung nilai recall

   Nilai precision merupakan perbandingan probabilitas nilai positif dengan banyaknya data yang sebenarnya positif :

$$\text{Recall} = \frac{TP}{TP + FP}$$

4. Menghitung nilai F1-Score

   F1-Score merupakan harmonic maen dari nilai precision dan recall dimana nilai terbaik dalam perhitungan F1-Score adalah 1.0 dan nilai terburuknya 0 berikut rumus persamaan dalam menentukan nilai F1-Score

$$\frac{1}{F1} = \frac{1}{2}\left(\frac{1}{Precision} + \frac{1}{Recall}\right)$$

I. Jadwal Penelitian

Luaran dari penelitian ini antara lain skripsi, artikel yang akan di submit pada jurnal dan aplikasi *machine learning* untuk menambah literatur di dunia penelitian.

Tabel 5 Data Jadwal Penelitian

No	Jadwal Kegiatan	2022					
		Agu	Sep	Okt	Nop	Des	Jan
1	Identifikasi Masalah	■					
2	Studi Pustaka		■				
3	Pengumpulan Data		■				
4	Pengolahan Data			■			
5	Perancangan Sistem			■			
6	Implementasi				■		
7	Pengujian					■	
8	Penyusunan Skripsi						■

Daftar Pustaka

Ali, A. A. (2019). Stroke prediction using distributed machine learning based on Apache spark. Stroke, 28(15), 89-97. Q2

Wu, Y., & Fang, Y. (2020). Stroke prediction with machine learning methods among older Chinese. International journal of environmental research and public health, 17(6), 1828. Q1

Faisal, A., & Subekti, A. Deep Neural Network untuk Prediksi Stroke. JEPIN (Jurnal Edukasi dan Penelitian Informatika), 7(3), 443-449. S2

Byna, A., & Basit, M. (2020). Penerapan Metode Adaboost Untuk Mengoptimasi Prediksi Penyakit Stroke Dengan Algoritma *Naïve Bayes*. Jurnal Sisfokom (Sistem Informasi dan Komputer), 9(3), 407-411. S3

Sakinah, N., Badriyah, T., & Syarif, I. (2020). Analisis Kinerja Algoritma Mesin Pembelajaran untuk Klasifikasi Penyakit Stroke Menggunakan Citra CT Scan. Jurnal Teknologi Informasi dan Ilmu Komputer, 7(4), 833-844. S2

Heo, J., Yoon, J. G., Park, H., Kim, Y. D., Nam, H. S., & Heo, J. H. (2019). Machine learning–based model for prediction of outcomes in acute stroke. Stroke, 50(5), 1263-1265. Q1

Dritsas, E., & Trigka, M. (2022). Stroke risk prediction with machine learning techniques. Sensors, 22(13), 4670 Q1

Saputri, N. D. (2021). Komparasi penerapan metode Bagging dan Adaboost pada Algoritma c4. 5 untuk prediksi Penyakit Stroke (Doctoral dissertation, UIN Sunan Ampel Surabaya).

Id, I. D. (2021). Machine Learning: Teori, Studi Kasus dan Implementasi Menggunakan Python (Vol. 1). Unri Press. buku

Sirsat, M. S., Fermé, E., & Câmara, J. (2020). Machine learning for brain stroke: a review. Journal of Stroke and Cerebrovascular Diseases, 29(10), 105162.

Roihan, A., Sunarya, P. A., & Rafika, A. S. (2020). Pemanfaatan Machine Learning dalam Berbagai Bidang. Jurnal Khatulistiwa Informatika, 5(1), 490845. Shinta 4

Wang, W., Kiik, M., Peek, N., Curcin, V., Marshall, I. J., Rudd, A. G., ... & Bray, B. (2020). A systematic review of machine learning models for predicting outcomes of stroke with structured data. PloS one, 15(6), e0234722. Q1

Saputri, N. D. (2021). Komparasi penerapan metode Bagging dan Adaboost pada Algoritma c4. 5 untuk prediksi Penyakit Stroke (Doctoral dissertation, UIN Sunan Ampel Surabaya).

Supriyatna, A., & Mustika, W. P. (2018). Komparasi Algoritma *Naive baye*s dan SVM Untuk Memprediksi Keberhasilan Imunoterapi Pada Penyakit Kutil. J-SAKTI (Jurnal Sains Komputer dan Informatika), 2(2), 152-161. S4

Rahmadani, D., Muzafar, A. A., Hamid, A., & Annisa, R. (2022, September). Analisis Perbandingan Algoritma C4. 5 dan CART Untuk Klasifikasi Penyakit Stroke: Comparative Analysis of C4. 5 and CART Algorithms for Classification of Stroke. In SENTIMAS: Seminar Nasional Penelitian dan Pengabdian Masyarakat (Vol. 1, No. 1, pp. 197-206). IRPI

Rachmad, D. U. M., Oktavianto, H., & Rahman, M. (2022). Perbandingan Metode *K-Nearest Neighbor* Dan *Gaussian Naive Bayes* Untuk Klasifikasi Penyakit Stroke. Jurnal Smart Teknologi, 3(4), 405-412.

Nugroho, A., & Religia, Y. (2021). Analisis Optimasi Algoritma Klasifikasi Naive Bayes menggunakan Genetic Algorithm dan Bagging. Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi), 5(3), 504-510. S2