PERBANDINGAN METODE ALGORITMA *NAÏVE BAYES* DAN *K-NEAREST NEIGHBORS* UNTUK KLASIFIKASI PENYAKIT *STROKE*

Khairul Akmal [1], Ahmad Faqih [2], Fatihanursari Dikananda [3]
[1,2] Program Studi Teknik Informatika, STMIK IKMI Cirebon
[3] Program Studi Sistem Informasi, STMIK IKMI Cirebon
akmalkhairul447@gmail.com

ABSTRAK

Stroke selaku serupa penyakit yang meninggali strata ketiga di Indonesia sesudah jantung serta kanker. Kerapkali setiap individu bermalas-malasan dalam mendapati terdapatnya penyakit *stroke*. Minimnya kekuatan kedokteran di Indonesia membikin rakyat sukar guna mengetahui dengan cara dini penyakit *stroke. Stroke* yaitu sesuatu sindrom klinis yang diisyarati dengan tandasnya peranan otak dengan cara berat yang bisa mengakibatkan kematian. Tujuan riset ini guna pengelompokan hasil kira-kira penyakit *stroke* dengan pendekatan algoritma *Naïve Bayes* serta *K-Nearest Neighbors* menurut kriteria-kriteria dalam penyakit *stroke* antara lain kategori genitalia, umur pesakit, darah tinggi, riwayat sakit jantung, sempat menikah, kategori profesi, kategori tempat bermukim, kandungan gula darah, BMI, status merokok. Statistik yang dibubuhkan dalam riset ini ialah *Stroke Prediction Dataset* yang diperoleh pada repositori *Kaggle* yang yaitu salah satu yang populer di negeri *Data Science* serta *Machine Learning*. Kali ini perubahan masa revolusi industri 4.0 yang bergerak selaras di segi teknologi serta ilmu kesehatan akibatnya selaku suatu yang bisa berharga dengan mengenakan *Machine Learning*. Banyak sekali faedah yang dibubuhkan dalam menduga sebagian penyakit yang bisa di proyeksi. Eksklusifnya penyakit *stroke* dengan mengenakan tilikan algoritma *Naïve Bayes* serta *K-Nearest Neighbors* guna tiap-tiap elastis nya. Implementasi cara ini mengenakan *Cross Validation* adalah *data training* serta *data testing* dibikin kuota dalam melaksanakan pengetesan. Pemanfaatan algoritma *Naïve Bayes* serta *K-Nearest Neighbors* bisa di terapkan selaku materi evaluasi guna membikin sistem pandai yang dibubuhkan oleh para pakar kesehatan guna pengumpulan ketetapan yang bagus di segi ke penjagaan serta medis dalam memacu hasil pemeriksaan pesakit *stroke*. Hasil ketelitian yang didapat dengan mengenakan algoritma *K-Nearest Neighbors* sebesar 94,36% hasil pengelompokan bisa dibubuhkan guna menolong dokter dalam pemeriksaan penyakit *stroke*.

Kata kunci: *Datamining, Stroke, Naïve Bayes, Klasifikasi, K-Nearest Neighbors*

1. PENDAHULUAN

World Health Organization (WHO) *Stroke* diartikan selaku hilangnya fungsi atau tugas otak secara tiba-tiba yang berlangsung dalam waktu 24 jam atau lebih [1]. Menurut statistik *stroke* global secara garis besar, sekitar 15 juta orang di semua mayapada mengidap *stroke* tiap-tiap tahun, dengan 1 dari 6 orang di dunia menderita *stroke*. Di Indonesia, 43,1% *stroke* di diagnosis oleh tenaga kesehatan pada kelompok berusia 75 tahun ke atas, dan 0,2% pada kelompok usia 15-24 tahun. *Stroke* yang tidak diatasi dengan baik sanggup menimbulkan kematian yang tinggi [2]. *Stroke* merupakan keadaan berbahaya yang butuh ditangani sesegera mungkin karena sel-sel otak bisa mati dalam hitungan menit. Kematian tiba-tiba mampu terjalin bila pengidap dalam situasi yang sangat serius. Perawatan yang tepat dapat mengurangi tingkat kerusakan otak serta kemungkinan komplikasi. Oleh lantaran itu, perlu untuk memprediksi apakah orang tersebut akan mengalami *stroke*. Salah satu metode untuk memprediksi *stroke* adalah dengan klasifikasi. Prediksi akurat penyakit memerlukan klasifikasi *stroke*. Hasil prediksi yang akurat dapat membantu profesional medis membuat keputusan yang tepat [3].

Dalam kedokteran *stroke* sering disebut sebagai *transient ischemic attack*. Kondisi neurologis yang berlangsung kala suplai darah ke bagian otak seketika tersendat. Penatalaksanaan *stroke* harus cepat serta tepat untuk menghindari kecacatan serta komplikasi lebih lanjut. Di masa teknologi tinggi ini, dibuat program pembelajaran mesin untuk mengidentifikasi serta mengenali orang yang terserang *stroke* dan menolong orang menjadi lebih sadar terhadap risiko penyakitnya. Teknik yang dipakai ialah *support vector machine (SVM)* untuk mengklasifikasikan orang yang terserang stroke. Metode *SVM* sangat sesuai sebab akurasi klasifikasi *SVM* cukup tinggi. Dengan program ini, semua orang dapat mengetahui peluang seseorang terserang *stroke* serta peluang tidak terserang *stroke* dengan persentase yang ditampilkan setelah menjalankan program. Tujuan dari dibuatnya program ini yakni guna menyelidiki algoritma SVM dalam pengelompokan data penyakit stroke, program ini menggunakan pengelompokan SVM yang menerima hasil ketelitian setidaknya agung dari data unbalance pada kernel linear adalah 76% dan polynomial sebesar 80%. buat data yang balanced didapat hasil ketelitian pada kernel linear 77%, serta di polynomial 76% [4].

Machine learning yaitu teknologi yang mampu dikenakan buat menduga penyakit *stroke*. Algoritma penerimaan mesin berwatak konstruktif dalam membikin ramalan yang tepat serta memberikan kajian yang tepat. Salah satu algoritma *machine learning classifier* yang mampu dikenakan buat menjalankan

ramalan ialah algoritma tumbuhan ketetapan C4.5 dan algoritma *Naïve Bayes*. Tujuan dari penelitian ini yaitu buat menyamakan kecermatan dan penampilan dari kedua algoritma buat menduga penyakit *stroke.* bersumber pada hasil penelitian ditemukan kalau algoritma C4.5 menjangkau jenjang kecermatan yang lebih mulia adalah 95% sebaliknya algoritma *Naïve Bayes* menjangkau jenjang kecermatan sebesar 91% [5].

Tabel 1. Emblem Data

No	Nama	Keterangan
1	*Id*	Id pengidap
2	*Gender*	Kategori genus
3	*Age*	Umur pengidap
4	*Hypertension*	Darah tinggi
5	*Heart_disease*	Pengidap sakit jantung
6	*Ever_meried*	Sempat menikah
7	*Work_type*	Model karier
8	*Residen_type*	Tempat bermukim
9	*Avg_glukosa_level*	Kandungan glukosa
10	*BMI*	Masa tubuh
11	*Smoking_status*	Sempat merokok
12	*Stroke*	Perkiraan *stroke*

Berdasarkan Tabel 1. Emblem Data yang dibubuhkan dalam penelitian ini yaitu *Stroke Prediction Dataset* yang ada pada repositori *Kaggle* yang adalah salah satu lokasi yang populer di mayapada *Data Science* serta *Machine Learning* yang terdiri dari lebih dari 6000 *dataset.* Keseluruhan *record* yang tampak pada *dataset* ini yakni 5110 data riset dengan 12 emblem. Dengan adanya penelitian ini diharapkan mampu mendiagnosa penyakit *stroke* berdasarkan kriteria-kriteria yang ditentukan.

Studi sebelumnya dengan tajuk "Penggunaan Algoritma *K-Nearest Neighbors (KNN)* guna meraba Penyakit *Stroke*" yang dijalani oleh Muhammad Naja Maskuri pada tahun 2022 memaknakan apabila studi ini mengerjakan bentuk kepada *dataset* penyakit *stroke* dengan memakai algoritma *K-Nearest Neighbors*. Dengan jumlah data yang dibubuhkan dalam riset sebesar 100 data, penghitungan data mengecas serta data tes memakai cara *split validation* dengan skala 80% banding 20%. Poin k yang dibubuhkan guna mencoba prestasi algoritma *K-Nearest Neighbors* dalam melihat penyakit *stroke* adalah k-9, dimana diperoleh ponten ketelitian sebesar 95%. menurut hasil itu algoritma *K-Nearest Neighbors* ada jenjang ketelitian yang bagus dalam mengerjakan prosedur prakiraan penyakit *stroke* [6].

Lalu penelitian lainnya dengan judul "Faktor Risiko Yang Mempengaruhi Kejadian *Stroke*" yang ditulis oleh Putra Agina Widyaswara Suwaryo dalam studinya mengenakan cara pertalian dengan pendekatan *cross-sectional*. Percontoh penelitian terdiri dari 38 penderita yang di preferensi sebagai sederhana *random sampling* yang memenuhi kriteria selamat *stroke*, berusia kurang dari 70 tahun, dan tidak memiliki penyakit jantung penyerta. Hasil penelitian menunjukkan bahwa minum kopi dan merokok tidak berpengaruh terhadap kejadian *stroke*. Perkembangan *stroke* dipengaruhi oleh aktivitas fisik atau raga, pemantauan tekanan darah secara teratur, dan stres [7].

2. TINJAUAN PUSTAKA

2.1. *Datamining*

Datamining ialah sesuatu yang bertujuan menyelesaikan suatu permasalahan dengan menerapkan model matematika agar menemukan pola dari data yang sudah ada sebelumnya secara otomatis. *Datamining* sering dikatakan pula *Knowledge Discovery in Databases* (KDD), yang menggambarkan semua proses mengubah sekumpulan data biasa menjadi pengetahuan yang bermanfaat [8].

2.2. Klassifikasi

Klasifikasi ialah teknik atau metode pengolahan data yang dipakai untuk membagi data ke dalam kelompok yang berbeda berdasarkan kriteria yang berbeda. Dengan mengklasifikasikan data, dapat ditentukan apakah data yang didapatkan akurat atau tidak. Dalam *datamining*, masalah klasifikasi dapat mengidentifikasi masalah akurasi prediksi menggunakan berbagai macam teknik [9].

2.3. *Naïve Bayes*

Naïve Bayes yakni pengategorian dengan prosedur probalitas dan statistik yang dijumpai oleh akademikus kebangsaan Inggris Thomas Bayes, yang memproyeksikan probabilitas di waktu depan bersumber pada pengalaman yang telah berlangsung [10]. Penelitian pada peluang kali ini mengenakan algoritma Naïve Bayes pada penggodokan datanya.

2.4. *Rapidminer*

RapidMiner yaitu aplikasi yang mempunyai fungsi sebagai sarana pembelajaran pada ilmu *data mining*. Basis ini dibesarkan oleh perseroan yang ber pengkhususan dalam seluruh tingkat data besar bidang usaha profitabel, penelitian, pembelajaran, dan penelaahan. *RapidMiner* mempunyai ratusan solusi pembelajaran untuk pengelompokan, klasifikasi, asosiasi, serta analisis regresi [11].

RapidMiner membawa kecendekiaan bikinan buat perseroan dengan basis ilmu data yang terbuka dan sanggup di rasio kan. Melampaui dari satu juta pengguna global menggunakan produk *RapidMiner* untuk mencari solusi permasalahan dengan *data mining* [12].

2.5. *K-Nearest Neighbors*

Algoritma *K-Nearest Neighbors* maupun sanggup jua disingkat KNN yaitu salah satu sistem penggolongan *data mining* yang bisa buat rekognisi bermacam perkara, kayak studi yang sudah terlihat adalah buat memproyeksikan penyakit *stroke*. Pemanfaatan algoritma *K-Nearest Neighbors* diseleksi gara-gara kemampuannya dalam memutuskan grup data dengan mengestimasi jarak orang sebelah terdekat. Algoritma *K-Nearest Neighbors* pada

kebanyakan dalam ramalan subjek menurut pada data penataran yang mempunyai ponten perbedaan kecil serta jarak orang sebelah terdekat dengan subjek [6].

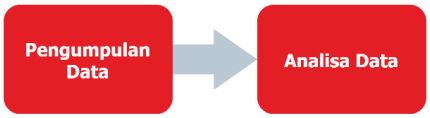
2.6. Uji *T-Test*

T-Test yakni cara percobaan dugaan dengan menggunakan satu pribadi (entitas studi) dengan mengenakan 2 perlakuan yang bertentangan. Walaupun dengan mengenakan entitas yang selevel Namun percontoh senantiasa terurai selaku 2 ialah data dengan perlakuan kesatu serta data dengan perlakuan ke2. Performance sanggup diketahui dengan teknik menyamakan situasi entitas studi kesatu serta situasi entitas pada studi ke2. uji coba t *(t-test)* ialah menyamakan jalinan antara 2 faktor ialah faktor respon serta faktor predictor. uji coba t *sample* ganda *(paired-sample t-test)* dipakai buat menakar pedoman jarak dua rata-rata dari dua sample yang ganda dengan sangkaan kalau data terdistribusi wajar [13].

3. METODE PENELITIAN

Pada studi ini pengarang memakai teknik studi kuantitatif. Data Kuantitatif adalah data yang didapat dari data *Stroke Prediction Dataset | Kaggle* seterusnya data itu hendak di analisa mengenakan pendekatan teknik algoritma *Naïve Bayes* dan *K-Nearest Neighbors.*

Ada pula fase riset yang dipakai pada riset ini selaku seterusnya:



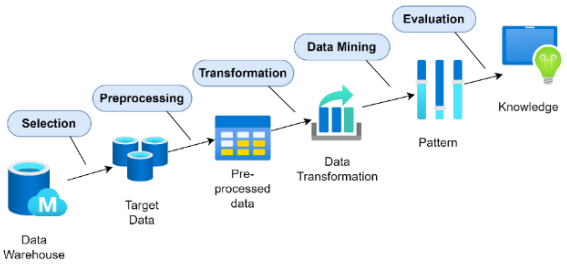
Gambar 1. Fase riset

3.1 Pengumpulan Data

1. Observasi
   Pada riset ini pengarang mengenakan prosedur riset kuantitatif. Fakta Kuantitatif yakni data yang dihasilkan dari data *Stroke Prediction Dataset | Kaggle* setelah itu data itu hendak di kajian mengenakan pendekatan prosedur algoritma *Naïve Bayes* serta *K-Nearest Neighbors.*
2. Riset Pustaka
   pemanfaatan referensi lama dipakai dalam studi selaku pangkal data sebagaimana literatur sebagai sumber data digunakan untuk mengkaji, menafsirkan, bahkan memprediksi.

3.2 Analisa Data

Prosedur analisa data yang dikenakan pada riset ini mengenakan teknik *Knowledge Discovery In Database (KDD)*, dengan beberapa tahap mulai dari *Data Selection*, *Data Preprocessing*, *Data Transformation*, *datamining*, hingga *evaluation*. Penjelasan rinci tentang tahapan tersebut sebagai berikut:



Gambar 2. Tahapan KDD
Sumber: www.researchgate.net

1. *Data Selection*
   Statistik *selection* yakni teknik meminimalkan jumlah data yang buat teknik *mining* dengan senantiasa menunjukkan data aslinya. Statistik hasil pilihan buat dalam teknik *data mining* ditaruh dalam *file* yang terpisah dari *database* produksi. Pada tingkatan ini, data yang didapat dari *Stroke Prediction Dataset | Kaggle.*
2. *Data Preprocessing*
   Keterangan *preprocesing* adalah teknik buat purifikasi penggeledahan data maupun teknik mencium serta memulihkan (maupun meniadakan) peringatan yang cacat maupun tidak cermat dari kerumunan peringatan, grafik, maupun *database* serta mengarahkan pada mengenali bagian data yang tidak komplit, salah, tidak cermat maupun tidak relevan serta setelah itu mengubah, memodifikasi, maupun meniadakan data kotor maupun agresif.
3. *Data Transformation*
   Data Transformation ialah metode transmutasi data dari satu dimensi ke format yang ada. Alih bentuk data yang setidaknya biasa merupakan merombak data belantah jadi yang bersih serta sanggup, merombak kategori data, membersihkan data tembusan, serta memperkaya data guna berguna jaringan. sepanjang metode alterasi data, seorang analis akan menentukan struktur, melakukan pemetaan data, mengekstrak data dari sumber asli, melakukan transformasi, dan akhirnya menyimpan data dalam *database* yang sesuai.
4. *Data mining*
   Data mining yaitu teknik mencari pola maupun data menarik dalam data tertentu dengan memanfaatkan cara maupun cara khusus. metode, cara, maupun algoritma dalam data mining amat berbagai macam. Penyortiran cara maupun algoritma yang pas amat tergantung pada tujuan serta teknik KDD dengan cara totalitas. Tujuan pentingnya merupakan buat mensaring rombongan data besar buat mengenali gaya, pola, serta ikatan buat menunjang pemungutan ketentuan serta perancangan yang pas. prosedur penggalian data pada riset ini memanfaatkan prosedur algoritma *Naïve Bayes* serta *K-Nearest Neighbors.*
5. *Evaluation*
   Evaluation merupakan interprestasi atau kesimpulan dari hasil datamining. Kesimpulan akhir

ditarik dengan menyatukan berbagai hipotesis yang dihasilkan dari datamining.

4. HASIL DAN PEMBAHASAN

4.1 Hasil

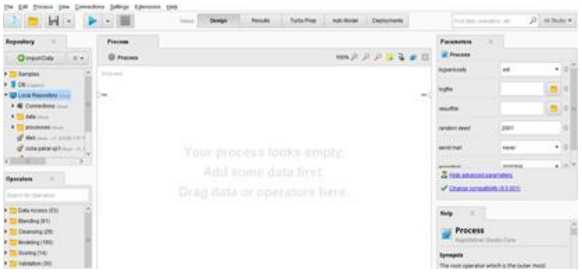
Studi ini digeluti berdasarkan *dataset* yang diperoleh dari data public *Stroke Prediction Dataset | Kaggle.* Berikut ini adalah penjabaran dari hasil proses penelitian klasifikasi dugaan penyakit *stroke* mengenakan algoritma *Naïve Bayes* serta *K-Nearest Neighbors.*

Pada tingkatan ini data yang dipakai hendak di koleksi dengan teknik menatap, kecocokan data dengan pokok maupun tajuk studi yang hendak di cermat, dalam perihal ini data yang didapat dari data massa *Stroke Prediction Dataset | Kaggle* telah sesuai dengan struktur data yang terdiri dari 5110 *record* yang terdiri dari 10 lambang *predictor* adalah jenis_kelamin, heart_disease, umur, darah_tinggi, sempat_menikah, jenis_pekerjaan, tempat_bermukim, kandungan_gula_darah, bmi, sempat_merokok, serta 1 lambang tujuan adalah *stroke.*

Adapun tabel *dataset* diagnosa penyakit *stroke* adalah sebagai berikut:

Tabel 2. Tabel *dataset* diagnosa penyakit *stroke*

No	Id	Gender	Heart disease	age	Hyper Tension	Ever married	Work type	Residence Type	Avg Glucose level	bmi	Smoking status	stroke
1	9046	Male	1.0	67.0	0.0	Yes	Private	Urban	228.69	36.6	formerly smoked	Yes
2	51676	Female	0.0	61.0	0.0	Yes	Self-employed	Rural	202.21	36.6	never smoked	Yes
3	31112	Male	1.0	80.0	0.0	Yes	Private	Rural	105.92	32.5	never smoked	Yes
5108	19723	Female	0.0	35.0	0.0	Yes	Self-employed	Rural	82.99	30.6	never smoked	No
5109	37544	Male	0.0	51.0	0.0	Yes	Private	Rural	166.29	25.6	formerly smoked	No
5110	44679	Female	0.0	44.0	0.0	Yes	Govt_job	Urban	85.28	26.2	never smoked	No

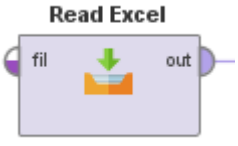


Gambar 3. Halaman utama rapidminer

Setelah pengumpulan data dilakukan, selanjutnya dilakukan analisa data menggunakan metode *Knowledge Discovery In Database* dengan lima tahapan didalamnya yang diantaranya yakni, *Data Selection, Data Preprocessing, Data Transformation, data mining,* hingga *evaluation*. Selanjutnya hasil yang didapatkan dari proses analisa data:

4.2 Data Selection

Dataset dibaca menggunakan operator *read excel* pada aplikasi rapidminer dan menghasilkan informasi statistik seperti berikut:

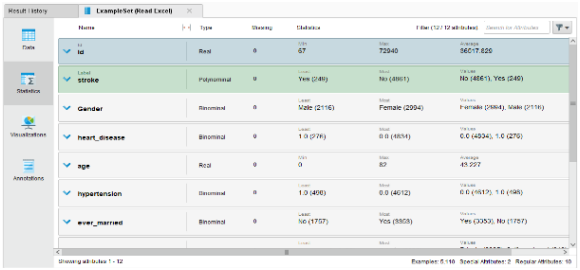


Gambar 3. Statistik dataset

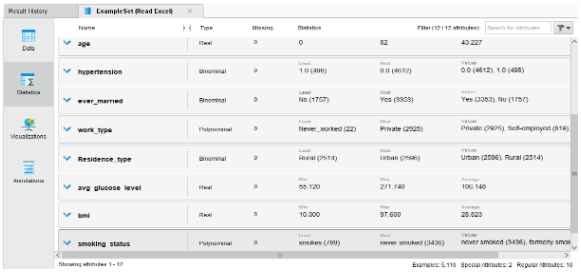
Untuk membaca *file excel* di Rapidminer dibutuhkan operator read excel, seperti Gambar 3. Operator *read excel*.

4.3 Data Preprocessing

Dalam sesuatu data yang besar selalu kali ditemui data yang cacat serupa *missing value* alias data yang sirna.



Gambar 4. Data statistik dataset (1)

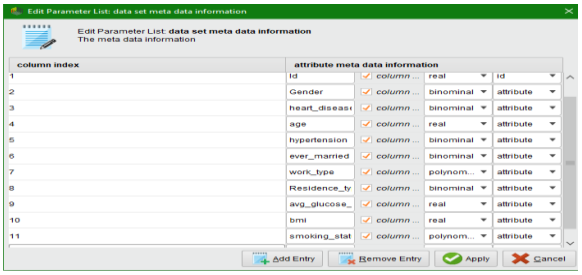


Gambar 5. Data statistik dataset (2)

Missing Value merupakan *record* yang salah satu alias sampai-sampai lebih dari atributnya tidak diketahui poinnya, pada skandal *missing value* kerap kali riset dijalani dengan metode memasukkan alias memuat dengan nilai rata–rata yang kerap timbul alias pula membersihkan atributnya. Dalam riset ini tidak dipakai *Operator Replace Missing Values* pada *tool RapidMiner* akibat sehabis di analisa lebih lanjut pada Gambar 4. Data statistik *dataset* (1) dan Gambar 5. Data statistik *dataset* (2) tidak ditemukan adanya *missing* atau suatu data yang hilang atau kosong.

4.4 Data Transformation

Pada tahapan ini data akan diubah menjadi bentuk yang sesuai untuk proses datamining. Berikut gambar tahapan *data transformation*:

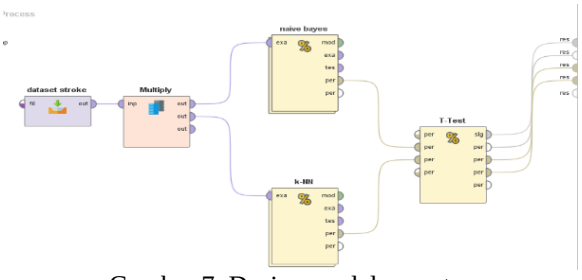


Gambar 6. Data Transformation

Pada aplikasi *RapidMiner* Studio, data yang sudah *via* sistem sebelumnya hendak di transfigurasi supaya sanggup cocok dengan algoritma yang dikenakan adalah algoritma *Naïve Bayes* serta *K-Nearest Neighbors.* Pada strata ini tanda yang dibubuhkan hendak diberi merek mencontoh hal data - data pada tanda itu.

4.5 Data mining

Proses *data mining* menggunakan dua operator pada rapidminer yaitu *FP-Growth* dan *Create Association Rules* dengan penggunaan nilai minimum *support*, mininum *confidence* dan minimum *lift* sebagai berikut:



Gambar 7. Design model operator

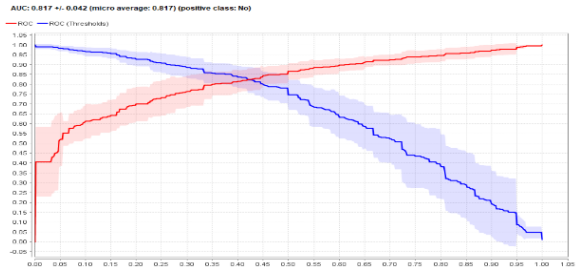
Pada gambar diatas model proses algoritma *Naïve Bayes* serta *K-Nearest Neighbors* dihubungkan *T-Test* menggambarkan proses setelah dilakukannya tahapan *data selection, preprocessing, transformation*, pada *software RapidMiner Studio* selanjutnya adalah memasukkan *operator multiply*, algoritma *Naïve Bayes, K-Nearest Neighbors* dan operator *T-Test*. Seluruh operator di hubung-hubungkan akibatnya bakal menciptakan tiruan kayak pada ilustrasi diatas. Kedua algoritma disatukan dengan *operator multiply* serta operator *T-Test*.

Metode pemeriksaan prosedur diawali dari penghitungan *dataset* dengan prosedur *10-fold cross validation*, sesudah itu dipraktikkan strata penilaian memakai *Area Under Curve (AUC)* guna mengukur hasil kecermatan dari penampilan bentuk pengelompokan. Hasil kecermatan ditinjau memakai kurva *Receiver Operating Characteristic (ROC)*. ROC menciptakan 2 garis dengan tatanan *true positive* selaku garis tegak serta *false positive* selaku garis melintang.

Kurva ROC ialah tool 2 sukatan yang guna memperkirakan kapasitas pengategorian yang memanfaatkan 2 *class* ketetapan, masing-masing subjek dipetakan ke salah satu komponen dari set pendamping, positif ataupun minus. Pada kurva ROC, TP *rate* di jalan cerita pada sumbu Y serta FP rate di jalan cerita pada sumbu X. Guna pengategorian *datamining*, angka AUC mampu dibelah sebagai sebagian regu, yakni *Excellent classification, good classification, fair classification, poor classification serta failure* [14].

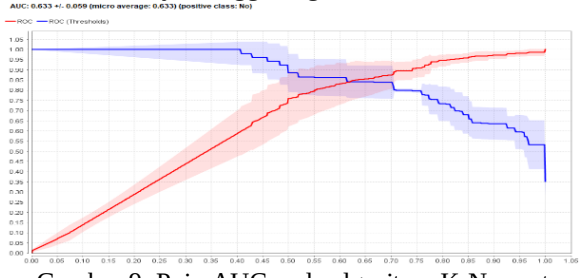
Tabel. 3 Performance ketepatan AUC

Performance	**Penggolongan**
0.90-1.00	Penggolongan Sempurna
0.80-0.90	Penggolongan Baik
0.70-0.80	Penggolongan Adil
0.60-0.70	Penggolongan Cukup
0.50-0.60	Penggolongan Tidak Baik



Gambar 8. Poin AUC pada algoritma Naïve Bayes

Dari percobaan dengan replika algoritma *Naïve Bayes*, ditemukan tabel ROC serupa Gambar 8 dengan angka AUC *(zona Under Curve)* sebesar 0,817 dengan dianosis hasilnya Penggolongan Baik.



Gambar 9. Poin AUC pada algoritma K-Nearest Neighbors

Dari percobaan dengan miniatur algoritma *K-Nearest Neighbors*, diperoleh diagram ROC serupa sketsa 9 dengan angka AUC *(zona Under Curve)* sebesar 0,633 dengan pemeriksaan hasilnya Penggolongan Cukup

Dalam eksperimen studi ini memakai dataset yang terdiri dari 10 tanda, prosedur yang dicoba yaitu algoritma kategorisasi Naïve Bayes, serta K-Nearest Neighbors. Hasil riset yang disuguhkan yaitu accuracy, precision, recall, serta AUC.

4.6 Evaluation

Tahapan evaluasi untuk melihat nilai akurasi dari algoritma *Naïve Bayes* serta *K-Nearest Neighbors.*

Tabel 4. Tabel hasil eksperimen algoritma *Naïve Bayes*, dan *K-Nearest Neighbors*

No	Algoritma	*PerformanceVector*
1	*Naïve Bayes*	90.10%
2	*K-Nearest Neighbors*	94.36%

Hasil *PerformanceVector* algoritma *Naïve Bayes*, serta *K-Nearest Neighbors* selaku selanjutnya:

Tabel 5. Pertimbangan *performanceVector* algoritma *Naïve Bayes*, dan *K-Nearest Neighbors*

No	*Criteria*	*Naïve Bayes*	*K-Nearest Neighbor*
1	*Accuracy*	90.10%	94.36%
2	*Precision*	96.28%	95.18%
3	*Recall*	93.19%	99.09%
4	*AUC*	0.817	0.633

Berlandaskan hasil *performancevector* algoritma *Naïve Bayes* serta *K-Nearest Neighbors* dalam pengelompokkan penyakit *stroke* mampu dipandang pada Tabel 5. Pertimbangan *performancevector* algoritma *Naïve Bayes* serta *K-Nearest Neighbors* bahwasannya algoritma *K-Nearest Neighbors* menciptakan angka ketepatan paling tinggi dengan angka ketepatan 94,36%.

Selanjutnya adalah tahapan evaluasi untuk melihat nilai akurasi dari algoritma *Naïve Bayes* serta *K-Nearest Neighbors.*

accuracy: 90.10% +/- 1.06% (micro average: 90.10%)

	true Yes	true No	class precision
pred. Yes	74	331	18.27%
pred. No	175	4530	96.28%
class recall	29.72%	93.19%	

Gambar 10. Angka akurasi algoritma Naïve Bayes

Menurut Gambar 10 Angka kecermatan algoritma *Naïve Bayes* yakni 90,10% tampak jika jumlah *true* positif (tp) yakni 74 *record* di kategorisasi selaku *true* positif serta *false* positif (fn) sejumlah 175 *record* di kategorisasi selaku positif tapi tidak positif. selanjutnya 331 *record false* positif (fp) di kategorisasi selaku tidak positif tapi positif serta 4530 *record* guna *true* positif (tn) di kategorisasi selaku tidak positif.

accuracy: 94.36% +/- 0.81% (micro average: 94.36%)

	true Yes	true No	class precision
pred. Yes	5	44	10.20%
pred. No	244	4817	95.18%
class recall	2.01%	99.09%	

Gambar 11. Angka akurasi algoritma K-Nearest Neighbors

Bersumber pada Gambar 11 Angka ketepatan algoritma *K-Nearest Neighbors* angka ketepatan nya 94,36% terpandang apabila jumlah *true* positif (tp) merupakan 5 *record* di penggolongan selaku *true* positif serta *false* positif (fn) sebesar 244 *record* di penggolongan selaku positif tapi tidak positif. Selanjutnya 44 *record false* positif (fp) di penggolongan selaku tidak positif tapi positif serta 4817 *record* guna *true* positif (tn) di penggolongan selaku tidak positif.

Sesudah itu pada riset ini dilakoni percobaan dengan memanfaatkan tes *T-Test*. Dalam signifikasi tes harga memanfaatkan jenjang signifikasi statistik sebagai 0.05. Berarti selaku statistik kurang dari 0.05 memperlihatkan terdapatnya antagonisme penting antara angka rata-rata, dengan seperti itu wajib menyanggah asumsi kosong (H_0). Sesudah itu dilakoni tes *T-Test* guna mencium penggolongan berselisih selaku penting.

A	B	C
	0.901 +/- 0.011	0.944 +/- 0.008
0.901 +/- 0.011		0.000
0.944 +/- 0.008		

Gambar 12. T-Test Signification (T-Test)

Dari hasil *T-Test Signification (T-Test)* menampakkan harga yang lebih kecil dari *alpha* = 0.050 yang menampakkan terdapatnya parak yang bermakna antara harga rata-rata. Tentang hal catatan angka prestasi algoritma *Naïve Bayes* = 0.901, serta *K-Nearest Neighbors* menampakkan hasil yang nilainya = 0.944.

Selanjutnya dijalani tes T guna menyamakan jalinan antara 2 fleksibel ialah fleksibel respon serta fleksibel *predictor*. Dalam tes t, *sample* berangkap *(paired-sample t-test)* dipakai guna menakar analogi perbedaan 2 rata-rata dari 2 *sample* yang berangkap dengan anggapan kalau data terdistribusi wajar.

Dari hasil *T-Test Signification (T-Test)* membuktikan angka yang lebih kecil dari *alpha* 0,05 menampakkan terdapatnya antagonisme yang bermakna antara angka rata-rata kejituan nya. Adapun Daftar nilai kinerja algoritma *Naïve Bayes* adalah 90.10%, dan *K-Nearest Neighbors* menunjukan hasil yang nilainya 94.36%.

Selanjutnya dihitung perbedaan selisih akurasi antara algoritma *Naïve Bayes* dan *K-Nearest Neighbors* sebesar 4.26%. Berasas uji coba T yang

menampakkan terdapatnya kontras serta memandang angka kecermatan yang terdapatnya kontras bisa di artikan kalau algoritma *K-Nearest Neighbors* lebih cakap dari algoritma *Naïve Bayes* guna riset perkara ini.

4.7 Penerapan Klasifikasi

Penerapan model klasifikasi algoritma *Naïve Bayes* serta *K-Nearest Neighbors* guna memprediksi hasil diagnosis penyakit *stroke* dilakukan melalui beberapa alur atau tahapan. Penelitian ini menggunakan motede pendekatan *Knowledge Discovery in Database (KDD).*

Tahapan pertama yaitu pada peringkat ini data yang ditemukan dari data *public Stroke Prediction Dataset | Kaggle* telah menggenapi patokan yang cocok. dengan struktur data yang terdiri dari 5110 *record* yang terdiri dari 10 lambang *predictor* adalah jenis_kelamin, heart_disease, umur, darah_tinggi, sempat_menikah, jenis_pekerjaan, tempat_bermukim, kandungan_gula_darah, bmi, sempat_merokok, serta 1 lambang tujuan adalah *stroke.*

Selanjutnya dilakukan *preprocessing*, pada tahapan ini biasanya terdapat *Missing Value* atau *record* yang salah satu maupun sampai-sampai lebih dari atributnya tidak diketahui angkanya, pada persoalan *missing value* kerap kali riset digeluti dengan teknik memasukkan maupun memuat dengan nilai rata–rata yang kerap timbul serta serta meniadakan atributnya. Bakal tapi dalam riset ini tidak dibubuhkan *Operator Replace Missing Values* pada *tool RapidMiner* gara-gara sesudah di analisa lebih lanjut pada data statistik *dataset* tidak dijumpai terdapatnya *Missing* maupun sesuatu data yang sirna maupun kosong.

Selanjutnya tahapan *data transformation*, mengunakan *software datamining* yakni *RapidMiner* studio, sehingga data yang dengan sistem sebelumnya hendak di transmutasi supaya mampu pantas dengan algoritma yang digunakan yakni algoritma *Naïve Bayes* serta *K-Nearest Neighbors*. Setiap atribut yang ada dirubah tipe datanya agar sesuai berdasarkan isi dari setiap *record*nya.

Berikutnya adalah tahapan *datamining*, pada riset ini data diolah dengan menggunakan *Software* Rapidminer *Studio* yang menerapkan metode algoritma *Naïve Bayes* serta *K-Nearest Neighbors*.

Selanjutnya yakni tahapan evaluasi, poin ketepatan *performancevec*tor algoritma *Naïve Bayes* sebesar 90.10% sedangkan nilai akurasi *performancevector* algoritma K-*Nearest Neighbors* sebesar 94.36%. Dari hasil tes T membuktikan poin yang lebih kecil dari *alpha* = 0,050 membuktikan terdapatnya parak yang bermakna antara poin rata-rata. Ada pula himpunan poin kapasitas algoritma *Naïve Bayes* = 0.901, serta *K-Nearest Neighbors* membuktikan hasil poin yang poinnya = 0.944. Pada hasil Gambar 12 diketahui kalau poin alpha dari segenap algoritma ada nilai poin lebih kecil dari 0.050 yang bisa diartikan kalau kedapatan parak bermakna antara algoritma *Naïve Bayes* dengan *K-Nearest Neighbors.*

4.8 Perbandingan Akurasi Klasifikasi

Dari hasil *T-Test Signification (T-Test)* membuktikan angka yang lebih kecil dari *alpha* 0,05 memperlihatkan terdapatnya variasi yang berarti antara angka rata-rata kejituannya. Ada pula catatan angka kapasitas algoritma *Naïve Bayes* yaitu 90.10%, serta *K-Nearest Neighbors* membuktikan hasil yang nilainya 94.36%. kemudian dihitung variasi kacek akurasi antara algoritma *Naïve Bayes* serta *K-Nearest Neighbors* sebesar 4.26%.

5. KESIMPULAN DAN SARAN

Hasil perbandingan terhadap diagnosa penyakit stroke menggunakan klasifikasi metode *Naïve Bayes* serta *K-Nearest Neighbors* menghasilkan nilai ketepatan *PerformanceVector* terbaik adalah algoritma *K-Nearest Neighbors* sebesar 94.36% dan *accuracy* terendah dimiliki oleh algoritma *Naïve Bayes* sebesar 90.10%.

Hasil perbandingan berlandaskan hasil pengetesan ditemui diagram ROC semacam pada Tabel 4 dengan poin AUC *(zona Under Curve)* algoritma *Naïve Bayes* sebesar 0.817 dengan pemeriksaan hasilnya Pengelompokkan Baik, dan algoritma *K-Nearest Neighbors* membuktikan hasil AUC sebesar 0.633 dengan pemeriksaan hasilnya Pengelompokkan Cukup.

Untuk menghasilkan hasil klasifikasi yang lebih baik dan lebih berkualitas perlu dilakukan penelitian lebih lanjut pada metode serta algoritma lainnya.

DAFTAR PUSTAKA

[1] Permatasari, N., 2020. Perbandingan *stroke* non hemoragik dengan gangguan motorik pasien memiliki faktor resiko diabetes melitus dan hipertensi. *Jurnal Ilmiah Kesehatan Sandi Husada*, *9*(1), pp.298-304.

[2] Amelia, U., Indra, J. and Masruriyah, A., 2022. Implementasi Algoritma *Support Vector Machine (SVM)* untuk Prediksi Penyakit *Stroke* dengan Atribut Berpengaruh. *Scientific Student Journal for Information, Technology and Science*, *3*(2), pp.254-259.

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

[4] Sulaeman, K.R., Setianingsih, C. and Saputra, R.E., 2022. Analisis Algoritma *Support Vector Machine* Dalam Klasifikasi Penyakit *Stroke*. *eProceedings of Engineering*, 9(3).

[5] Kohsasih, K.L. and Situmorang, Z., 2022. Analisis Perbandingan Algoritma C4. 5 dan *Naïve Bayes* Dalam Memprediksi Penyakit *Cerebrovascular*. Jurnal Informatika, 9(1), pp.13-17.

[5] Maskuri, M.N., Harliana, H., Sukerti, K. and Bhakti, R.M.H., 2022. Penerapan Algoritma *K-Nearest Neighbor* (KNN) untuk Prediksi Penyakit *Stroke*. Jurnal Ilmiah *Intech: Information Technology Journal of* UMUS, 4(01), pp.130-140.

[6] Suwaryo, P.A.W., Widodo, W.T. *and* Setianingsih, E., 2019. Faktor risiko yang mempengaruhi kejadian *stroke*. Jurnal Keperawatan, 11(4), pp.251-260.

[7] Mirbod, M. and Dehghani, H., 2023. *Smart Trip Prediction Model for Metro Traffic Control Using Data Mining Techniques. Procedia Computer Science*, 217, pp.72-81.

[8] Bahtiar, A., Dwilestari, G., Basysyar, F.M. and Suarna, N., 2021, February. *Data mining techniques with machine learning algorithm to predict patients of heart disease. In IOP Conference Series: Materials Science and Engineering* (Vol. 1088, No. 1, p. 012035). IOP Publishing.

[9] Sarwiyah, Y., Rahaningsih, N. and Basysyar, F.M., 2020. Klasifikasi Data Nasabah Produk Asuransi Kendaraan Menggunakan Algoritma *Naive Bayes* Pada PT. Jasaraharja Putera. KOPERTIP: Jurnal Ilmiah Manajemen Informatika dan Komputer, 4(3), pp.124-130.

[10] V. R. Prasetyo, H. Lazuardi, A. A. Mulyono, and C. Lauw, "Penerapan Aplikasi *RapidMiner* Untuk Prediksi Nilai Tukar Rupiah Terhadap *US Dollar* Dengan Metode *Linear Regression*," *Jurnal Nasional Teknologi dan Sistem Informasi*, vol. 7, no. 1, pp. 8–17, May 2021, doi: 10.25077/teknosi.v7i1.2021.8-17.

[11] "*Why Choose Us*? | *RapidMiner*." https://rapidminer.com/why-rapidminer/ (accessed Jan. 23, 2023).

[12] Annisa, R., 2019. Analisis Komparasi Algoritma Klasifikasi Data *Mining* Untuk Prediksi Penderita Penyakit Jantung. JTIK (Jurnal Teknik Informatika Kaputama), 3(1), pp.22-28.

[13] Ilham, A., 2017. Komparasi Algoritma Kasifikasi dengan Pendekatan Level Data Untuk Menangani Data Kelas Tidak Seimbang. Jurnal Ilmiah Ilmu Komputer Fakultas Ilmu Komputer Universitas Al Asyariah Mandar, 3(1), pp.1-6.

[14] Amelia Nastuti and Syaiful Zuhri Harahap, "Teknik Data Minig untuk Penentuan Paket Hemat Sembako dan Kebutuhan Harian dengan Menggunakan Algoritma *FP-Growth* (Studi Kasus di Ulfamart Lubuk Alung)," *Informatika: Jurnal IlmiahFakultas Sains dan Teknologi*, 2019.