JATIKOM: Jurnal Aplikasi dan Teori Ilmu Komputer
ISSN: 2615-577X (Online)

Analisis Data Mining Untuk Memprediksi Penyakit Stroke Dengan Algoritma N*aïve Bayes*

Laila Rahmawati[#1], Moch. Hafid[*2], Muhammad Agus Sunandar[#3]

[#]*Teknik Informatika Sekolah Tinggi Teknologi Wastukancana*
Jl. Cikopak No.53 Sadang, Purwakarta, Jawa Barat

[1]lailarahmawati06@wastukancana.ac.id
[2]mhafid@wastukancana.ac.id
[3]agoes.61@stt-wastukancana.ac.id

Abstract— **Stroke is a chronic disease that has a dangerous impact caused by impaired cerebral blood circulation due to blockage of arteries due to blood deposits in blood vessels. The number of stroke cases in the world is very large, one of which is in Indonesia reaching 10.9% or around 2,210,362 people. Stroke occurs due to lack of awareness in carrying out a healthy lifestyle. Risk factors for stroke include diabetes mellitus, heart disease, and hypertension. The number of cases of stroke, the researchers analyzed the prediction of stroke odds, so that people can prevent strokes. This study uses a Data Mining technique using the Naïve Bayes algorithm to predict stroke. By using the Knowledge Discovery In Database (KDD) method, this study can determine the percentage accuracy of stroke prediction. There are five stages of Knowledge Discovery in the Database, namely data selection, data availability, data transformation, data mining processing, and evaluation. This study conducted a test using the RapidMiner application with 90% Training Data and 10% Testing Data.**
Based on the results of measuring the performance of the model using the confusion matrix test method, it is known that the Naïve Bayes algorithm has an accuracy rate of 98%, precision is 80%, and recall is 92%.

Keywords— *Stroke, Naïve Bayes, Confusion Matrix, Data Mining, and RapidMiner*

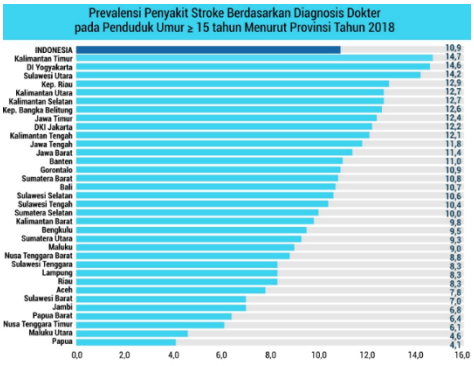
Abstrak— **Stroke merupakan penyakit kronis yang memberikan dampak berbahaya yang diakibatkan oleh gangguan peredaran darah otak karena penyumbatan pembuluh darah arteri akibat endapan darah pada pembuluh darah. Jumlah kasus stroke di dunia sangat banyak, salah satunya pada negara Indonesia mencapai 10,9% atau sekitar 2.210.362 orang. Stroke terjadi akibat kurangnya kesadaran dalam melakukan pola hidup sehat. Faktor risiko penyakit stroke diantaranya diabetes melitus, penyakit jantung, dan hipertensi. Banyaknya kasus penyakit stroke maka peneliti melakukan analisis prediksi peluang stroke, agar masyarakat dapat mencegah terjadinya stroke.**
Penelitian ini menggunakan teknik Data Mining dengan algoritma Naïve Bayes untuk memprediksi penyakit stroke. Dengan menggunakan metode Knowledge Discovery In Databases (KDD), maka penelitian ini dapat mengetahui berapa persen akurasi dari prediksi penyakit stroke. Terdapat lima tahapan dari Knowledge Discovery In Database yaitu seleksi data, pembersihan data, transformasi data, pengolahan data mining, dan evaluasi. Penelitian ini melakukan pengujian menggunakan aplikasi RapidMiner dengan Data Training 90% dan *Data Testing* 10%. Berdasarkan hasil pengukuran performa dari model tersebut dengan menggunakan metode pengujian confusion matrix, diketahui bahwa algoritma *Naïve Bayes* memiliki tingkat akurasi sebesar 98%, *precision* sebesar 80%, dan *recall* sebesar 92%.

Kata Kunci: ***Stroke, Naïve Bayes, Confusion Matrix, Data Mining*, RapidMiner**

I. PENDAHULUAN

Stroke adalah terjadinya penyumbatan atau pecahnya pembuluh darah otak sehingga mengganggu atau mengurangi suplai darah ke otak secara tiba-tiba. Otak kemudian tidak bisa menerima oksigen yang cukup, dan sel-sel pada otak mulai mengalami kerusakan bahkan mati[1]. Menurut data *World Stroke Organization* (WSO), jumlah kasus *stroke* mencapai 13,7 juta kasus baru *stroke*, dan sekitar 5,5 juta kematian terjadi akibat penyakit *stroke*. Hal tersebut menyebabkan *stroke* menjadi penyebab kematian kedua di dunia [2].

Terjadinya penyakit *stroke* disebabkan oleh kurangnya kesadaran masyarakat dalam melaksanakan pola hidup sehat. Dikutip data dari Kementerian Kesehatan Republik Indonesia jumlah kasus *stroke* di Indonesia pada tahun 2018 berdasarkan diagnosis dokter pada penduduk dengan umur $\geqslant 15$ tahun sebanyak 10,9% atau diperkirakan 2.210.362 orang. Provinsi dengan prevalensi tertinggi *stroke* adalah Kalimantan Timur (14,7%), seperti terlihat pada (Gambar 1) [2].



Gambar 1. Prevalensi Stroke Di Indonesia Tahun 2018 [2].

Inovasi dalam pencegahan dan pengendalian serta pengobatan *stroke* sangat penting untuk dilakukan. Namun, selain membentuk berbagai kebijakan untuk pengendalian dan pencegahan *stroke*, sudah terdapat berbagai penelitian yang dilakukan di Indonesia mengenai penyakit *stroke*. Penelitian-penelitian tersebut di antaranya melakukan penelitian untuk memprediksi *stroke* berdasarkan data riwayat kondisi pasien menggunakan data mining. Data mining adalah suatu proses yang menggunakan teknik statistik untuk mengekstraksi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai basis data. Ada suatu metode dalam data mining yaitu disebut *classification*, yaitu metode untuk memprediksi atau klasifikasi kategori suatu data berdasarkan sekumpulan variable atau atribut dari data tersebut [3]. Salah satu algoritma yang digunakan dalam *classification* adalah *Naïve Bayes*. *Naïve Bayes* adalah pengklasifikasi probabilistik sederhana yang menghitung probabilitas dengan menjumlahkan frekuensi dan kombinasi nilai dari kumpulan data yang diberikan [4].

Berdasarkan pemaparan di atas, maka dalam penelitian ini penulis akan mengimpementasikan algoritma *Naïve Bayes* untuk memprediksi diagnosis pasien. Dengan harapan penelitian ini dapat menghasilkan akurasi yang lebih tinggi dalam memprediksi diagnosis *stroke*.

II. KAJIAN PUSTAKA

A. *Penelitian Terdahulu*

Annisa Puspitawuri melakukan penelitian untuk prediksi stroke dengan *Deep Neural Network* (DNN). Dalam penelitian ini data yang digunakan juga sama seperti yang digunakan penulis, yaitu set data *stroke* dari *Fedesoriano* yang dikelola oleh *Kaggle*. Penelitian ini membandingkan tiga teknik oversampling untuk mendapatkan model prediksi yang lebih baik. Hasil yang paling baik didapatkan pada teknik sampling *SMOTE Tomek* dan arsitektur DNN dengan lima hidden layer, *optimal* Adam, *learning rate* 0.001, dan jumlah *epoch* 500. Skor akurasi, presisi, *recall* dan fl-score masing-masing mendapatkan 0.96, 0.9614, dan 0.9611 [5].

Anas Faisal juga melakukan penelitian mengenai stroke, penelitian ini menggunakan data kondisi kesehatan masyarakat kota Malang tahun 2015, menggunakan sebanyak 150 data responden. Hasil penelitian diuji dan dianalisis dengan pengujian pengaruh sebaran data seimbang dan tidak seimbang. Nilai akurasi tertinggi diperoleh pada data kelas seimbang yaitu 96.67% dengan data latih sebanyak 45 dan nilai K=15-22. Sedangkan pada data kelas tidak seimbang akurasi tertinggi 100% dengan jumlah data latih sebanyak 60 dan nilai K=20-30 [6].

B. *Stroke*

Stroke adalah suatu kondisi terjadi ketika asupan darah menuju ke otak terganggu atau berkurang akibat penyumbatan *(stroke iskemik)* atau pecahnya pembuluh darah *(stroke hemoragik)*, sehingga jaringan otak kekurangan oksigen dan nutrisi yang mengakibatkan sel-sel dalam otak akan mulai mati. Kondisi ini menyebabkan bagian tubuh yang dikendalikan oleh otak tidak dapat berfungsi dengan baik [7].

Faktor risiko yang menyebabkan penyakit *stroke* ada 3 yaitu:

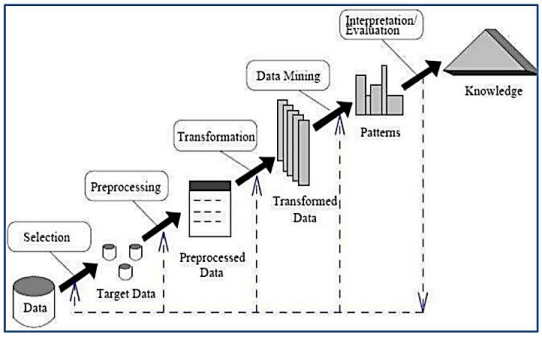
1. Hipertensi
   Hipertensi adalah kondisi ketika tekanan darah *sistolik* berada di angka ≥ 140 mmHg dan tekanan darah diastolic ≥ 90 mm/Hg.
2. Penyakit Jantung
   Penyakit jantung adalah kondisi di mana jantung mengalami gangguan seperti gangguan pembuluh darah atau otot jantung. Penyakit jantung juga dapat disebabkan oleh infeksi atau kelainan lahir.
3. Diabetes Melitus
   Diabetes melitus adalah di mana kondisi gula darah seseorang sangat tinggi yaitu ≥ 200 mg/dL [2].

C. *Data mining*

Data Mining adalah proses mengekstraksi dan mengidentifikasi informasi yang berguna menggunakan statistik, matematika, kecerdasan buatan, dan pembelajaran mesin. Penambangan data didefinisikan sebagai proses menemukan pola dalam data. Berdasarkan tugas tersebut, penambangan data dikelompokkan menjadi deskripsi, ramalan, prediksi, klasifikasi, pengelompokan, dan asosiasi [8].

D. *Knowledge Discovery in Databases (KDD)*

Knowledge Discovery in Databases (KDD) adalah sekumpulan proses untuk menemukan pengetahuan yang bermanfaat dari data. KDD terdiri dari langkah-langkah perubahan, termasuk data *pre-processing* dan *post-processing* [9]. Proses KDD secara garis besar dapat dilihat pada Gambar 2.



Gambar 2. Tahapan Knowledge Discovery in Databases

E. Klasifikasi

Klasifikasi merupakan bagian dari data mining. Klasifikasi adalah metode yang menggunakan data dengan target *(class/label)* yang berupa nilai kategorikal/nominal. Algoritma yang digunakan pada klasifikasi adalah *Naïve Bayes*, KNN, C4.5, ID3, dan lain sebagainya [10].

F. Prediksi

Prediksi adalah proses memperkirakan sesuatu yang paling mungkin akan terbukti dengan membandingkan informasi yang dimiliki masa lau dengan informasi yang sudah dimiliki sekarang dengan tujuan agar kesalahan selisih antara sesuatu yang terjadi dengan hasil perkiraan dapat diperkecil [11].

Algoritma yang digunakan dalam prediksi adalah algoritma C4.5, algoritma K-Means, algoritma Apriori, dan *Naïve Bayes* [12].

G. Algoritma Naïve Bayes

Naïve Bayes adalah klasifikasi menggunakan metode probabilistik dan statistik yang diusulkan oleh ilmuwan Inggris Thomas Bayes. Klasifikasi *Naïve Bayes* mengasumsikan bahwa ada atau tidak adanya ciri tertentu dari satu kelas tidak ada hubungannya dengan fitur kelas lain [13]. Notasi algoritma *Naïve Bayes* sebagai berikut:

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

Rumus 1. Notasi Algoritma Naïve Bayes [14]

H. Confusion Matrix

Confusion matrix merupakan salah satu metode yang dapat digunakan untuk mengukur performansi suatu metode klasifikasi. Pada dasarnya, *Confusion Matrix* berisi informasi yang membandingkan hasil klasifikasi yang dilakukan oleh sistem dengan hasil klasifikasi yang diharapkan [15]. *Confusion matrix* meliputi 3 hasil yaitu akurasi, presisi, dan *recall*. Rumus *Confusion Matrix* [16] sebagai berikut :

$$Accuracy = \frac{\mathrm{TP + TN}}{\mathrm{TP + TN + FP + FN}} \mathrm{x100\%}$$

Rumus 2. Notasi menghitung *Accuracy*

$$Precision(P) = \frac{\mathrm{TP}}{\mathrm{TP + FP}}$$

Rumus 3. Notasi menghitung *Precision*

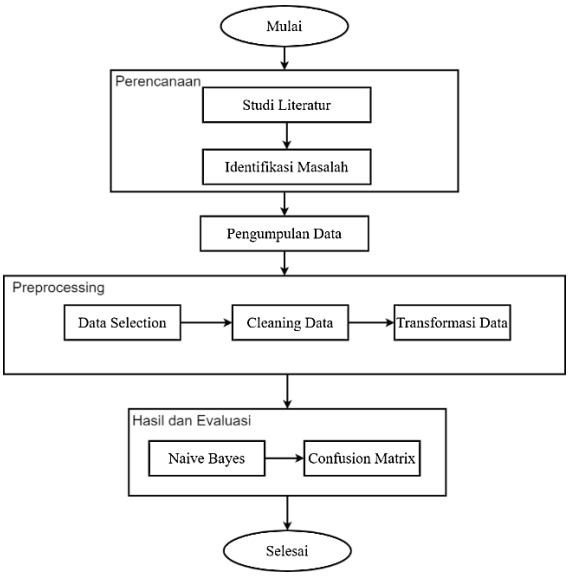
$$Recall(R) = \frac{\mathrm{TP}}{\mathrm{TP + FN}}$$

Rumus 4. Notasi Rumus Menghitung *Recall*

III. METODOLOGI PENELITIAN.

A. Kerangka Berpikir

Kerangka berpikir adalah model konseptual tentang bagaimana teori berhubungan dengan berbagai faktor yang diidentifikasi sebagai masalah yang penting [17]. Kerangka berpikir dapat dilihat pada Gambar 3.



Gambar 3. Kerangka Berpikir

B. Metode Penelitian

Adapun metode yang dilakukan dalam penelitian ini dibagi ke dalam beberapa bagian, yaitu metode pengumpulan data, data *pre-processing*, proses data mining, dan evaluasi.

1) Pengumpulan data: Dalam proses pengumpulan data yang digunakan pada penelitian ini didapatkan dari *Stroke Predict* data set. Data tersebut disediakan oleh *flatform* data set yaitu *Kaggle*.
2) Data *Pre-processing*: Tahap data *pre-processing* meliputi beberapa proses di antaranya yaitu:
   a. Data *selection*, bertujuan untuk memilih dan memilah data apa saja yang akan digunakan.
   b. Data *Cleaning*, dilakukan untuk membersihkan data dengan cara melengkapi data, menghapus data duplikat dan data kosong serta data yang tidak digunakan. Data yang sudah dipilih pada tahap seleksi kemudian dibersihkan untuk menghilangkan *missing value* dan *redundant data*.

3) Transformasi data bertujuan untuk menyesuaikan data dengan proses *data mining*.
4) Proses *Data Mining*: Tahap ini dilakukan dengan algoritma *Naïve Bayes*. Algoritma *Naïve Bayes* digunakan untuk menentukan diagnosis dari dokumen data pasien.
5) Evaluasi: Tahap ini dilakukan untuk menguji performa algoritma yang digunakan dalam penelitian.

IV. HASIL DAN PEMBAHASAN

A. *Pengumpulan Data*

Data yang digunakan dalam penelitian ini merupakan data pasien dengan gejala menunjukan Stroke yang disediakan oleh *flatform* dataset yaitu *Kaggle*, yang dapat diunduh secara bebas melalui link: *https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset*

Data set ini berisi 5110 baris data dengan 12 kolom yang terdiri atas: *independent features*, yaitu *gender*, *age*, *hypertension*, *heart disease*, *ever married*, *work type*, *residence type*, *avg glucose level*, *bmi*, *smoking status*, seperti pada Tabel 1. Kemudian 1 kolom *dependent features* bernama *stroke* yang merupakan label hasil prediksi.

TABEL I. DATA PASIEN STROKE

gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
Male	67	0	1	Yes	Private	Urban	228,69	36,6	formerly smoked	Yes
Male	80	0	1	Yes	Private	Rural	105,92	32,5	never smoked	Yes
Female	49	0	0	Yes	Private	Urban	171,23	34,4	smokes	Yes
Female	79	1	0	Yes	Self-employed	Rural	174,12	24	never smoked	Yes
Male	81	0	0	Yes	Private	Urban	186,21	29	formerly smoked	Yes
Male	74	1	1	Yes	Private	Rural	70,09	27,4	never smoked	Yes
Female	69	0	0	No	Private	Urban	94,39	22,8	never smoked	Yes
Male	68	0	0	Yes	Self-employed	Urban	91,68	40,8	Unknown	No
Male	9	0	0	No	children	Urban	71,88	17,5	Unknown	No
Male	82	1	0	Yes	Self-employed	Rural	71,97	28,3	never smoked	No
Female	45	0	0	Yes	Private	Urban	97,95	24,5	Unknown	No
Female	57	0	0	Yes	Private	Rural	77,93	21,7	never smoked	No
Female	18	0	0	No	Private	Urban	82,85	46,9	Unknown	No
Female	13	0	0	No	children	Rural	103,08	18,6	Unknown	No
Female	81	0	0	Yes	Self-employed	Urban	125,2	40	never smoked	No
Female	35	0	0	Yes	Self-employed	Rural	82,99	30,6	never smoked	No
Male	51	0	0	Yes	Private	Rural	166,29	25,6	formerly smoked	No
Female	44	0	0	Yes	Govt_job	Urban	85,28	26,2	Unknown	No

B. *Data Pre-processing*

Tahap data *pre-processing* terdapat beberapa bagian yaitu *selection*, *cleaning*, dan *transformation*.

Tahapan pertama yaitu seleksi data. Tahapan ini bertujuan untuk memilih data apa saja yang akan digunakan. Dalam penelitian ini data prediksi *stroke* yang digunakan adalah data prediksi stroke dari umur 1-83 tahun. Setelah melakukan tahapan ini data berkurang dari 5110 baris data menjadi 4995 baris data.

Kemudian tahapan selanjutnya adalah *cleaning* data. Tahapan ini bertujuan untuk membersihkan data dari data yang memiliki nilai *null* dan *missing value*, serta *redundant data*. Setelah melalui tahapan *cleaning* data yang ada berkurang dari 4995 baris data 4799 baris data.

Selanjutnya yaitu tahap transformasi data. Tahap ini dilakukan menyesuaikan data dengan proses *data mining* yang akan dilakukan. (Tabel 2).

TABEL II. PRATINJAU DATA SETELAH DITRANSFORMASI

JK	Umur	kanan Darah Tin	enyakit Jantun	Status Menikah	Tipe Pekerjaan	Tipe Tempat Tingga	Rata-Rata Glukosa	BMI	Status Merokok	stroke
Male	67	0	1	1	Private	Urban	Diabetes	Overweig	formerly smoke	Yes
Male	80	0	1	1	Private	Rural	Normal	Overweig	never smoked	Yes
Female	49	0	0	1	Private	Urban	Normal	Overweig	smokes	Yes
Female	79	1	0	1	Self-employed	Rural	Normal	Normal w	never smoked	Yes
Male	81	0	0	1	Private	Urban	Normal	Overweig	formerly smoke	Yes
Male	74	1	1	1	Private	Rural	Normal	Overweig	never smoked	Yes
Female	69	0	0	0	Private	Urban	Normal	Normal w	never smoked	Yes
Female	78	0	0	1	Private	Urban	Normal	Normal w	Unknown	Yes
Female	81	1	0	1	Private	Rural	Normal	Overweig	never smoked	Yes
Female	61	0	1	1	Govt_job	Rural	Normal	Overweig	smokes	Yes
Male	82	1	0	1	Self-employed	Rural	Normal	Overweig	never smoked	No
Female	45	0	0	1	Private	Urban	Normal	Normal w	Unknown	No
Female	57	0	0	1	Private	Rural	Normal	Normal w	never smoked	No
Female	18	0	0	0	Private	Urban	Normal	Overweig	Unknown	No
Female	13	0	0	0	children	Rural	Normal	Normal w	Unknown	No
Female	81	0	0	1	Self-employed	Urban	Normal	Overweig	never smoked	No
Female	35	0	0	1	Self-employed	Rural	Normal	Overweig	never smoked	No
Male	51	0	0	1	Private	Rural	Normal	Overweig	formerly smoke	No
Female	44	0	0	1	Govt_job	Urban	Normal	Overweig	Unknown	No

C. *Data Mining*

Pada tahapan ini dilakukan pemodelan data menggunakan hasil dari transformasi data, Adapun metode yang digunakan dalam penelitian ini adalah metode klasifikasi untuk memprediksi dengan menggunakan algoritma *Naïve Bayes* dalam menggunakan algoritma *Naive Bayes* akan melakukan perhitungan prediksi peluang *stroke* pada pasien. Adapun Langkah-langkah dalam melakukan analisis klasifikasi untuk prediksi Stroke yaitu sebagai berikut:

a. Menghitung jumlah kelas/label

Langkah pertama adalah menghitung jumlah kelas berdasarkan klasifikasi yang terbentuk *(Prior Probability)*. Dalam data terdapat 4136 *record* dengan kategori tidak terkena Stroke dan 183 *record* dengan kategori terkena Stroke.

P(Kelas Prediksi Stroke = Yes)
= 4120 / 4319 = 0,954.
P(Kelas Prediksi Stroke = No)
= 199 / 4319 = 0,046.

b. Menghitung jumlah data per kelas/atribut yang ada

Menghitung jumlah data per atribut sama dengan menghitung jumlah kelas. Perhitungan setiap atribut sebagai berikut :

- **P (Age | KelasPrediksiStroke)**
  P (Age = 54 | KelasPrediksiStroke = Yes)
  = 6 / 183 = 0,033.
  P (Age = 54 | KelasPrediksiStroke = No)
  = 72 / 4136 = 0,017.
- **P (Gender | KelasPrediksiStroke)**
  P (Gender = Male | KelasPrediksiStroke = Yes)
  = 74 / 183 = 0,404.
  P (Gender = Male | KelasPrediksiStroke = Yes)
  = 1697 / 4136 = 0,410.
- **P (Hypertension | KelasPrediksiStroke)**
  P (Hypertension= Yes | KelasPrediksiStroke = Yes)
  = 46 / 183 = 0,251.

P (Hypertension= Yes | KelasPrediksiStroke = No) = 391 / 4136 = 0,092.

- **P (Heart Disease | KelasPrediksiStroke)**
  P (HeartDisease= No | KelasPrediksiStroke = Yes) = 158 / 183 = 0,863.
  P (HeartDisease= No | KelasPrediksiStroke = No) = 3937 / 4136 = 0,952.
- **P (Ever Married | KelasPrediksiStroke)**
  P (EverMarried= Yes | KelasPrediksiStroke = Yes) = 23 / 183 = 0,885.
  P (EverMarried= Yes | KelasPrediksiStroke = No) = 1681 / 4136 = 0,662.
- **P (Work Type | KelasPrediksiStroke)**
  P (WorkType = Self-employed | KelasPrediksiStroke = Yes) = 46 / 183 = 0,251.
  P (WorkType = Self-employed | KelasPrediksiStroke = No) = 654 / 4136 = 0,158.
- **P (ResidanceType | KelasPrediksiStroke)**
  P (ResidanceType= Urban | KelasPrediksiStroke = Yes) = 93 / 183 = 0,508
  P (ResidanceType= Urban | KelasPrediksiStroke = No) = 2094 / 4136 = 0,506.
- **P (Avg Glucose Level | KelasPrediksiStroke)**
  P (AvgGlucoseLevel = Diabetes | KelasPrediksiStroke = Yes) = 38 / 183 = 0,208.
  P (AvgGlucoseLevel = Diabetes | KelasPrediksiStroke = No) = 305 / 4136 = 0,074.
- **P (BMI | KelasPrediksiStroke)**
  P (BMI = Normal Weight | KelasPrediksiStroke = Yes) = 35 / 183 = 0,191.
  P (BMI = Normal Weight | KelasPrediksiStroke = No) = 1051 / 4136 = 0,254.
- **P (Smoke Status | KelasPrediksiStroke)**
  P (Smoke Status = Formerly Smoked | KelasPrediksiStroke = Yes) = 49 / 183 = 0,268.
  P (Smoke Status = Formerly Smoked | KelasPrediksiStroke = No) = 697 / 4136 = 0,169.

c. Mengalikan semua variable setiap kelas.
P (KelasPrediksiStroke =Yes)
= 0,404 * 0,33 * 0,251 * 0,863 * 0,885 * 0,251 * 0,508 * 0,208 * 0,191 * 0,268 * 0,042 = 1,466E-07.
P (KelasPrediksiStroke = No)
= 0,410 * 0,017 * 0,092 * 0,952 * 0,662 * 0,158 * 0,506 * 0,074 * 0,254 * 0,169 * 0,958 = 1,002E-07.

d. Membandingkan hasil perhitungan manual dengan sample sebuah data testing diatas dengan data testing sebagai berikut:
Gender = Male, Age = 54, Hypertension = Yes, Heart Disease = No, *Ever* Married = *No*, Work *Type = Self-employed*, *Residence type = Urban*, *avg glucose level* = Diabetes, BMI = *Normal Weight*, *smoking status* = *formerly smoked*, Prediksi Stroke = Yes.

Pada data testing di atas, telah dilakukan perhitungan manual *Naïve Bayes*. Setelah dilakukan perhitungan, diperoleh hasil prediksi stroke adalah *"Yes"*. Karena nilai hasil perkalian seluruh probabilitas yang tertinggi adalah kelas prediksi *Stroke "Yes"*.

D. Evaluasi

Tahap evaluasi pada penelitian ini menggunakan confusion matrix untuk mendapatkan *accuracy*, *precision*, dan *recall*, *confusion matrix* berisi informasi yang membandingkan hasil klasifikasi yang dilakukan oleh sistem (Tabel 3).

TABEL III. CONFUSION MATRIX

	True No	**True Yes**
Pred No	448	2
Pred Yes	6	24

$$Accuracy = \frac{24 + 448}{24 + 448 + 6 + 2} \text{ x100\%}$$

Accuracy = 0,98 x 100% =98%

$$Recall(R) = \frac{24}{24 + 2}$$

Recall = 0,92 x 100% = 92%

$$Precision(P) = \frac{24}{24 + 6}$$

Precision = 0,80 x 100% = 80%

V. KESIMPULAN

Dari hasil penelitian diatas mengenai analisis prediksi penyakit *Stroke* yang berjumlah 4799 data dan sudah melewati tahap *pre-processing* data yang terdiri dari seleksi data, pembersihan data, transformasi data, dan diproses dengan menggunakan algoritma Naïve Bayes, serta evaluasi data dengan *Confusion Matrix* menggunakan *tools RapidMiner Studio* dapat disimpulkan bahwa akurasi yang dihasilkan menggunakan data testing 10% atau sebanyak 480 data yaitu 98% dengan nilai presisi *(precision)* sebesar 80%, dan tingkat keberhasilan *(recall)* sebesar 92%. Berdasarkan pengujian tersebut maka model algoritma *Naïve Bayes* bisa direkomendasikan untuk prediksi penyakit *Stroke*, karena nilai *Recall*, dan *Precision* tinggi.

REFERENSI

[1] dr. I. G. A. A. Yudha, "Kupas Tuntas Penyakit Stroke, dari Jenis, Gejala, hingga Cara Mengobatinya," 2021.

[2] R. Aprianda, *stroke-dont-be-the-one*. Jakarta: InfoDATIN, 2019.

[3] Anggada Maulana, "Konsep Dasar Data Mining," *Konsep Data Min.*, vol. 1, hal. 1–16, 2018.

[4] A. Saleh, "Implementasi Metode Klasifikasi Naïve Bayes Dalam Memprediksi Besarnya Penggunaan Listrik Rumah Tangga," *Creat. Inf. Technol. J.*, vol. 2, no. 3, hal. 207–217, 2015.

[5] A. Puspitawuri, E. Santoso, dan C. Dewi, "Diagnosis Tingkat Risiko Penyakit Stroke Menggunakan Metode K-Nearest Neighbor dan Naïve Bayes," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 3, no. 4, hal. 3319–3324, 2019, [Daring]. Tersedia pada: e-issn: 2548-964X http://j-ptiik.ub.ac.id

[6] A. Faisal dan A. Subekti, "Deep Neural Network untuk Prediksi Stroke," *J. Edukasi dan Penelit. Inform.*, vol. 7, no. 3, hal. 443, 2021, doi: 10.26418/jp.v7i3.50094.

[7] dr. G. Florencia, "Stroke," *halodoc.com*, 2020.

[8] M. Muharrom, "Klasifikasi Diagnosa Peradangan Kandung Kemih Menggunakan Metode Algoritma Naïve Bayes," *Indones. J. Bus. Intell.*, vol. 3, no. 2, hal. 31, 2021, doi: 10.21927/ijubi.v3i2.1472.

[9] Karsito dan W. M. Sari, "Prediksi Potensi Penularan Produk Delifrance dengan Metode Naive Bayes di PT. Pangan Lestari," vol. 9, no. September, hal. 67–78, 2018.

[10] M. H. Rifqo dan A. Wijaya, "Implementasi Algoritma Naive Bayes Dalam Penentuan Pemberian Kredit," *Pseudocode*, vol. 4, no. 2, hal. 120–128, 2017, doi: 10.33369/pseudocode.4.2.120-128.

[11] Herdianto, "Prediksi Kerusakan Motor Induksi Menggunakan Metode Jaringan Saraf Tiruan Backpropagation," Universitas Sumatera Utara, 2013.

[12] E. Rahajeng dan S. Tuminah, "Prevalensi Hipertensi dan Determinannya di Indonesia," *Maj Kedokteran Indonesia*, vol. 59. hal. 580–587, 2009.

[13] W. P. Nurmayanti, "Penerapan Naive Bayes dalam Mengklasifikasikan Masyarakat Miskin di Desa Lepak," *Geodika J. Kaji. Ilmu dan Pendidik. Geogr.*, vol. 5, no. 1, hal. 123–132, 2021, doi: 10.29408/geodika.v5i1.3430.

[14] I. W. Saputro dan B. W. Sari, "Uji Performa Algoritma Naïve Bayes untuk Prediksi Masa Studi Mahasiswa," *Creat. Inf. Technol. J.*, vol. 6, no. 1, hal. 1, 2020, doi: 10.24076/citec.2019v6i1.178.

[15] Karsito dan S. Susanti, "Klasifikasi Kelayakan Peserta Pengajuan Kredit Rumah Dengan Algoritma Naïve Bayes Di Perumahan Azzura Residencia," *J. Teknol. Pelita Bangsa*, vol. 9, hal. 43–48, 2019.

[16] S. Salmu dan A. Solichin, "Prediksi Tingkat Kelulusan Mahasiswa Tepat Waktu Menggunakan Naïve Bayes : Studi Kasus UIN Syarif Hidayatullah Jakarta," *Semin. Nas. Multidisiplin Ilmu 2017*, no. April, hal. 701–709, 2017.

[17] D. Sugiyono, *Metode Penelitian Kuantitatif, Kualitatif, dan Tindakan*. 2013.

[18] D. A. Lestari, "Batas Kadar Gula Darah yang Normal dalam Tubuh," *Hallosehat, Kemenkes RI*, 2021.

[19] P. K. R. P2PTM Kemenkes RI, "Klasifikasi Obesitas setelah pengukuran IMT," *Kementrian Kesehatan Republik Indonesia*, 2018.