

## **BAB 2**

### **LANDASAN TEORI**

Dalam melakukan penelitian ini terdapat beberapa teori pendukung / telaah literatur yang digunakan untuk membangun aplikasi website prediksi penyakit storke, yaitu:

#### **2.1 Stroke**

Stroke merupakan penyakit pembuluh darah di otak. Pada artikel milik Direktorat Pencegahan dan Pengendalian Penyakit Tidak Menular (2018), mengatakan, definisi dari stroke menurut WHO, stroke adalah suatu keadaan dimana ditemukan tanda-tanda klinis yang berkembang cepat berupa defisit neurologik fokal dan global, yang dapat memberat dan berlangsung lama selama 24 jam atau lebih dan atau dapat menyebabkan kematian, tanpa adanya penyebab lain yang jelas selain pembuluh darah [21].

Stroke adalah tantangan kesehatan global yang besar dan terus meningkat. Sekitar 85% stroke adalah iskemik dan disebabkan oleh penyakit pembuluh darah otak kecil, kardioemboli, dan tromboemboli terkait aterosklerosis arteri besar, sedangkan sekitar 15% lainnya disebabkan oleh pendarahan intraserebral, yang bisa dalam atau keseluruhan lobus; 80% diantaranya disebabkan oleh penyakit pembuluh darah otak kecil (arteriopati perforator dalam, angiopati amiloid serebral). Stroke pada orang dewasa dibawah 50 tahun menyumbang 15% kasus dan dapat disebabkan oleh spektrum penyebab yang berbeda dibandingkan dengan individu yang lebih tua, termasuk diseksi ekstrakranial[2].

Ada 2 macam faktor risiko pada penyakit stroke, yaitu faktor risiko yang tidak dapat dimodifikasi dan faktor risiko yang dapat dimodifikasi. Faktor yang tidak dapat dimodifikasi, seperti umur, jenis kelamin, etnis, dan genetik, serta faktor yang dapat dimodifikasi, seperti hipertensi, diabetes melitus, faktor jantung, merokok, hiperlipidemia, konsumsi alkohol dan penyalahgunaan zat, obesitas, perilaku menetap, dan peradangan[2].

#### **2.2 Naïve Bayes**

*Naïve Bayes classifier* adalah teknik prediksi berdasarkan probabilistik sederhana dan penerapan teorema Bayes (atau aturan Bayes) dengan asumsi

independensi yang kuat[22]. Dasar teorema *Naïve Bayes* yang digunakan adalah rumus Bayes sebagai berikut[23] :

$$P(\mathbf{A}|\mathbf{B}) = P(\mathbf{A}) \frac{P(\mathbf{B}|\mathbf{A})}{P(\mathbf{B})} \quad (2.1)$$

Keterangan:

- A : Attributes
- B : Class
- $P(\mathbf{A})$  : Probabilitas kejadian A
- $P(\mathbf{B})$  : Probabilitas kejadian B
- $P(\mathbf{A}|\mathbf{B})$  : Probabilitas kejadian A dengan syarat kejadian B telah terjadi atau Probabilitas posterior
- $P(\mathbf{B}|\mathbf{A})$  : Probabilitas kejadian B dengan syarat kejadian A telah terjadi atau Probabilitas kemungkinan

Pada klasifikasi yang menggunakan teorema *Naïve bayes*, diperlukan beberapa petunjuk untuk proses penentuan kelas yang sesuai bagi data sampel yang dianalisis, sehingga persamaan 2.1 dapat di modifikasi menjadi persamaan berikut [24, 25] :

$$P(\mathbf{y}|x_1, \dots, x_n) = P(\mathbf{y}) \frac{P(x_1, \dots, x_n|\mathbf{y})}{P(x_1, \dots, x_n)} \quad (2.2)$$

Keterangan:

- $P(\mathbf{y}|x_1, \dots, x_n)$  : posterior probability atau peluang adanya karakteristik dari sampel pada kelas y
- $P(\mathbf{y})$  : class prior probability atau peluang kemunculan kelas y
- $P(x_1, \dots, x_n|\mathbf{y})$  : likelihood atau peluang adanya karakteristik dari sampel pada kelas y
- $P(x_1, \dots, x_n)$  : prior probability atau peluang adanya kemunculan karakteristik dari sampel secara umum

Berdasarkan keterangan dari rumus 2.2 diatas maka dapat disimpulkan bahwa posterior probability adalah hasil perkalian antara class prior probability dengan likelihood kemudian dibagi dengan predictor prior probability.

### 2.3 Gaussian Naïve Bayes

Algoritma Gaussian Naïve Bayes digunakan pada saat data bersifat kontinu, dimana diasumsikan nilai kontinu yang berkaitan dengan tiap kelas yang didistribusi sesuai distribusi Gaussian. Data yang dilatih dibagi sesuai kelas dan mean, dimana standar deviasi tiap kelas dihitung. Untuk memperkirakan probabilitas dari data yang kontinu dapat menggunakan persamaan berikut [26, 27]:

$$P(x_i | y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right) \quad (2.3)$$

Keterangan:

$x_i$  : variable

$y$  : class

$\mu$  : mean atau rata-rata

$\sigma^2$  : variance atau (*standardevisi*)<sup>2</sup>

### 2.4 Confusion Matrix

Nilai yang diprediksi dan benar dari keanggotaan kelas dapat diklasifikasi silang dan dihitung dalam confusion matrix[28]. Confusion matrix adalah visualisasi yang ada dimana-mana untuk membantu orang mengevaluasi model pembelajaran mesin, tata letak tabular yang membandingkan label kelas yang diprediksi dengan label kelas yang sebenarnya pada semua sampel data[29].

Tabel 2.1. Tabel Confusion Matrix

		Aktual	
		Positive	Negative
Prediksi	Positive	TP	FP
	Negative	FN	TN

Keterangan:

- True Positive (TP) : Saat sistem memprediksi kalau seorang pasien terkena penyakit stroke dan secara aktual pasien terkena penyakit stroke.
- True Negative (TN) : Saat sistem memprediksi kalau seorang pasien terkena penyakit stroke dan secara aktual pasien tidak terkena penyakit stroke.

- False Positive (FP) : Saat sistem memprediksi kalau seorang pasien tidak terkena penyakit stroke dan secara aktual pasien terkena penyakit stroke.
- False Negative (FN) : Saat sistem memprediksi kalau seorang pasien tidak terkena penyakit stroke dan secara aktual pasien tidak terkena penyakit stroke.

## 2.5 Evaluasi Peforma

- Akurasi

Akurasi didapatkan dengan menghitung nilai dari True Positive ditambah True Negative dibagi dengan penjumlahan dari semua nilai dalam tabel confusion matrix. Berikut adalah persamaan Akurasi[30]:

$$Akurasi = \frac{TP + TN}{TP + TN + FP + FN} \quad (2.4)$$

- Precision

Precision atau presisi didapatkan dengan menghitung nilai dari True Positive dibagi dengan penjumlahan dari True Positive dan False Positive. Berikut adalah persamaan Precision[30]:

$$Precision = \frac{TP}{TP + FP} \quad (2.5)$$

- Recall

Recall atau sensitivitas (sebagaimana disebut dalam Psikologi) merupakan proporsi kasus True Positive yang diprediksi Positif dengan benar, serta digunakan untuk mengukur cakupan kasus True Positive dengan aturan +P (Predicted Positive)[31]. Recall didapat dengan melakukan perhitungan terhadap True Positive dibagi dengan hasil penjumlahan dari True Positive dan False Negative. Berikut adalah persamaan Recall[30]:

$$Recall = \frac{TP}{TP + FN} \quad (2.6)$$

- F1-Score

F1-Score didapat dengan mengalikan precision dengan recall kemudian dibagi dengan penjumlahan precision dan recall setelah itu hasilnya dikalikan

dua. Berikut adalah persamaan F1-Score[30]:

$$Recall = 2 * \frac{precision * recall}{precision + recall} \quad (2.7)$$

## 2.6 Python

Python adalah pemrograman bahasa tingkat tinggi yang dapat digunakan dalam pemrograman untuk banyak kebutuhan umum. Python merupakan bahasa pemrograman yang modern dan berlabel tinggi dengan *dynamic typing* dan sederhana, dan sintaks yang semantik dan konsisten[32]. Masih dalam sumber yang sama dikatakan python bersifat multiplatform, sangat modular, dan cocok untuk digunakan dalam *rapid development* dan pemrograman skala besar.

Python adalah bahasa populer di kalangan programmer belakangan ini. Python dirancang untuk kemudahan penggunaan dan pemahaman, telah mendominasi sebagai bahasa pemrograman yang paling populer, khususnya di kalangan pemula, menggantikan peran Java sebagai bahasa pengantar yang paling diminati[33].

Bahasa pemrograman ini digunakan dalam penelitian ini untuk proses prediksi penyakit stroke.

## 2.7 Penelitian Terdahulu

Adapun penelitian terdahulu untuk mendukung penelitian ini disajikan dalam bentuk tabel berikut:

UMMN  
UNIVERSITAS  
MULTIMEDIA  
NUSANTARA

Tabel 2.2. Tabel Penelitian Terdahulu

Judul	Nama Peneliti	Tahun terbit	Masalah	Metode	Dataset	Hasil Penelitian
Multiclass Classification with Iris Dataset using Gaussian Naïve Bayes	Z. Iqbal, dan M. Yadav	2020	Evaluasi akurasi GNB dalam klasifikasi data Iris ke tiga kelas, serta identifikasi metrik kinerja terkait.	Gaussian Naïve Bayes	Dataset Iris dari repositori UCI	Akurasi 95%.
Analyzing the performance of stroke prediction using ML classification algorithms	G. Sailasya, dan G.L.A Kumari	2021	Eksplorasi prediksi stroke otak menggunakan pembelajaran mesin, mengisi kesenjangan penelitian yang terbatas dalam area ini.	Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbors, Support Vector Machine dan Naïve Bayes	Dataset publik di Kaggle	Naïve Bayes akurasi paling tinggi sekitar 82%

Tabel 2.3. Tabel Lanjutan 1 Penelitian Terdahulu

Judul	Nama Peneliti	Tahun terbit	Masalah	Metode	Dataset	Hasil Penelitian
Cerebral infarction classification using the k-nearest neighbor and naive bayes classifier	S.H. Rukmawan, F.R. Aszhari, Z. Rustam, dan J. Pandelaki	2021	Menentukan pengklasifikasi terbaik antara kNN dan NB untuk deteksi infark serebral dengan data pencitraan medis, guna meningkatkan akurasi diagnostik dengan mengatasi keterbatasan masing-masing metode.	K-Nearest Neighbor, Gaussian Naïve Bayes	Dataset dari Departemen Radiologi, rumah sakit dr. Cipto Mangunkusumo (RSCM)	NB akurasi 97%, lebih baik dari kNN 91%

UNIVERSITAS  
MULTIMEDIA  
NUSANTARA

Tabel 2.4. Tabel Lanjutan 2 Penelitian Terdahulu

Judul	Nama Peneliti	Tahun terbit	Masalah	Metode	Dataset	Hasil Penelitian
Sistem Pakar dalam Mengidentifikasi Gejala Stroke Menggunakan Metode Naive Bayes	F. Karim, G.W., Nurcahyo, dan S. Sumijan	2021	Kurangnya kesadaran masyarakat tentang gejala stroke menyebabkan penundaan pemeriksaan medis. Dikembangkan Sistem Pakar web dengan NB untuk membantu mengenali gejala dan mendorong pencarian pertolongan medis.	Naive Bayes	Data dari Rumah Sakit Otak dr. Drs. M. Hatta di Bukit-tinggi	Akurasi 100% dalam identifikasi jenis penyakit stroke dari 10 sampel data yang digunakan

UNIVERSITAS  
MULTIMEDIA  
NUSANTARA



Tabel 2.5. Tabel Lanjutan 3 Penelitian Terdahulu

Judul	Nama Peneliti	Tahun terbit	Masalah	Metode	Dataset	Hasil Penelitian
Stroke Prediction	S, Kowsalya	2021	Tingginya angka kematian akibat stroke, terutama di negara berkembang.. Meningkatkan prediksi risiko stroke dengan membandingkan algoritma klasifikasi untuk mencari metode prediksi terbaik.	Logistic regression, gaussian-NB, linear discriminant analysis	Dataset publik di Kaggle	Akurasi regresi logistik sebesar 0,73%, akurasi gaussian-NB sebesar 0,71%, akurasi analisis diskriminan linear sebesar 0,72%

UNIVERSITAS  
MULTIMEDIA  
NUSANTARA

Tabel 2.6. Tabel Lanjutan 4 Penelitian Terdahulu

Judul	Nama Peneliti	Tahun terbit	Masalah	Metode	Dataset	Hasil Penelitian
Perbandingan Algoritma K-Nearest Neighbor dan Gaussian Naive Bayes pada Klasifikasi Penyakit Diabetes Melitus	P. T. Rahayu, Dayanto, and Q. A'yun	2022	Memilih algoritma klasifikasi terbaik (KNN atau GNB) untuk identifikasi data Diabetes Melitus dari basis data besar	K-Nearest Neighbor, Gaussian Naive Bayes	Dataset publik di Kaggle	GNB memiliki kinerja lebih baik dari KNN dengan peningkatan akurasi = 6,15%, presisi = 6,81%, dan recall = 2,37



Tabel 2.7. Tabel Lanjutan 5 Penelitian Terdahulu

Judul	Nama Peneliti	Tahun terbit	Masalah	Metode	Dataset	Hasil Penelitian
Perbandingan Metode K-Nearest Neighbor dan Gaussian Naive Bayes Untuk Klasifikasi Penyakit Stroke	Rachmad, D.U.M., Oktavianto, H. and Rahman, M.	2022	Pentingnya prediksi cepat dan akurat untuk penyakit stroke guna mengurangi kerusakan otak dan komplikasi, serta membantu praktisi kesehatan dalam pengambilan keputusan.	K-Nearest Neighbors, Gaussian Naive Bayes	Dataset publik di Kaggle	GNB akurasi lebih tinggi (74,45%) dibanding KNN (68,30%). GNB tingkat akurasi, presisi, dan recall lebih baik dari KNN dengan peningkatan akurasi = 6,15%, presisi = 6,81%, dan recall = 2,37%

Tabel 2.8. Tabel Lanjutan 6 Penelitian Terdahulu

Judul	Nama Peneliti	Tahun terbit	Masalah	Metode	Dataset	Hasil Penelitian
A Diabetes Prediction Classifier Model Using Naive Bayes Algorithm	F.M. Okikiola, O. S. Adewale, dan O.O. Obe	2023	Pendekatan deteksi diabetes yang ada menghadapi tantangan signifikan, termasuk ketidakseimbangan data, pemilihan fitur, dan kurangnya kerangka kerja umum untuk klasifikasi diabetes.	Naive Bayes berbasis ontologi	Dataset publik PIMA Indian di Kaggle	F-measure 87% dan evaluasi AUC superior

U N I V E R S I T A S  
M U L T I M E D I A  
N U S A N T A R A

Tabel 2.9. Tabel Lanjutan 7 Penelitian Terdahulu

Judul	Nama Peneliti	Tahun terbit	Masalah	Metode	Dataset	Hasil Penelitian
Comparison of K-Nearest Neighbor, Naive Bayes Classifier, Decision Tree, and Logistic Regression in Classification of Non-Performing Financing	R. Putra dan I. D. Ratih	2023	Pendekatan deteksi diabetes yang ada menghadapi tantangan signifikan, termasuk ketidakseimbangan data, pemilihan fitur, dan kurangnya kerangka kerja umum untuk klasifikasi diabetes.	K-Nearest Neighbors, Naive Bayes, Decision Tree, dan Logistic Regression	Dataset disediakan oleh DOE dan DOS, Malaysia	Hasil klasifikasi terbaik adalah Naive Bayes (80% data latih dan 20% data uji). Akurasi = 84,69%, sensitivitas = 58,25%, dan spesifisitas = 90,16%

UNIVERSITAS  
MULTIMEDIA  
NUSANTARA

Tabel 2.10. Tabel Lanjutan 8 Penelitian Terdahulu

Judul	Nama Peneliti	Tahun terbit	Masalah	Metode	Dataset	Hasil Penelitian
Early Prediction of Lung Cancer Using Gaussian Naive Bayes Classification Algorithm	M. Vedaraj, C.S. Anita, A. Muralidhar, V. Lavanya, K. Balasaranya, dan P. Jagadeesan, P	2023	Mendiagnosis kanker paru-paru secara akurat merupakan tantangan signifikan bagi para radiolog, sehingga prediksi dini sangat penting untuk meningkatkan tingkat kelangsungan hidup pasien.	Gaussian Naïve Bayes, Decision Tree, Random Forest, Support Vector Machine, AdaBoost	Dataset diambil dari Universitas California, Irvivne (UCI)	GNB akurasi 96%

UNIVERSITAS  
MULTIMEDIA  
NUSANTARA

Tabel 2.11. Tabel Lanjutan 9 Penelitian Terdahulu

Judul	Nama Peneliti	Tahun terbit	Masalah	Metode	Dataset	Hasil Penelitian
Naive Bayes and K-Nearest Neighbor Algorithm Approach in Data Mining Classification of Drugs Addictive Diseases	D. Priyanto, A. R. Iman, dan D. Jollyta	2023	Prevalensi penyalahgunaan narkoba di Nusa Tenggara Barat (NTB) tinggi, yang menghadirkan risiko kesehatan serius bagi pengguna, termasuk penyakit-penyakit yang mengancam jiwa seperti peradangan hati, penyakit jantung, hipertensi, dan stroke.	Naive Bayes dan K-Nearest Neighbors	Dataset berasal dari survei Badan Narkotika Nasional Provinsi NTB	Naive bayes akurasi = 94,5%, KNN = 92,5%

Tabel 2.12. Tabel Lanjutan 10 Penelitian Terdahulu

Judul	Nama Peneliti	Tahun terbit	Masalah	Metode	Dataset	Hasil Penelitian
Perbandingan Metode Algoritma Naive Bayes dan K-Nearest Neighbors untuk klasifikasi penyakit stroke	K. Akmal, A. Faqih, dan F. Dikananda	2023	Kurangnya kesadaran masyarakat Indonesia tentang risiko stroke dan keterbatasan layanan medis, yang membuat deteksi dini sulit. Menggunakan algoritma NB dan KNN untuk klasifikasi penyakit stroke.	Gaussian Naïve Bayes, K-Nearest Neighbors	Dataset publik dari Kaggle	KNN akurasi 94,36%, NB akurasi 90,10%. Hasil perbandingan analisis kurva ROC, NB AUC = 0,817 (klasifikasi Baik), KNN AUC = 0,633 (klasifikasi Cukup)

UNIVERSITAS  
MULTIMEDIA  
NUSANTARA



Tabel 2.13. Tabel Lanjutan 11 Penelitian Terdahulu

Judul	Nama Peneliti	Tahun terbit	Masalah	Metode	Dataset	Hasil Penelitian
Using Machine Learning Algorithm as a Method for Improving Stroke Prediction	N. Alageel, dan R. Alharbi, dan R. Alharbi, dan M. Alsayil, dan L. Alharbi	2023	Meningkatnya kebutuhan akan diagnosis dan pengelolaan stroke serta pemanfaatan teknologi medis untuk meningkatkan perawatan pasien.	Naïve Bayes, SVM, Random Forest, KNN, Decision Tree, Stacking, dan Majority Voting	Dataset publik dari kaggle	NB akurasi terendah 86%, f-1 score 0.93, recall 0.89, presisi 0.97. DT akurasi serupa = 96%, skor f1 0,98, presisi 0,97, dan recall 1

Berdasarkan penelitian terdahulu, beberapa algoritma pembelajaran mesin telah dievaluasi untuk meningkatkan prediksi penyakit, khususnya stroke dan diabetes. Penelitian oleh Iqbal dan Yadav (2020) menunjukkan bahwa Gaussian Naïve Bayes mencapai akurasi 95% dalam mengklasifikasikan data Iris. Sailasya dan Kumari (2021) menemukan bahwa Naïve Bayes adalah algoritma terbaik untuk prediksi stroke dengan akurasi 82%. Rukmawan et al. (2021) membandingkan kNN dan Naïve Bayes untuk deteksi infark serebral dan menemukan NB lebih akurat dengan 97%. Karim et al. (2021) mengembangkan sistem pakar berbasis web untuk meningkatkan kesadaran masyarakat tentang gejala stroke, mencapai akurasi 100%. Penelitian Kowsalya (2021) berfokus pada teknik pembelajaran mesin untuk prediksi risiko stroke dan menemukan bahwa akurasi gaussianNB sebesar 71%,.

Rahayu et al. (2022) menunjukkan Gaussian Naïve Bayes lebih unggul dalam klasifikasi diabetes dibandingkan kNN. Akmal et al. (2023) membandingkan kNN dan Gaussian Naïve Bayes untuk klasifikasi penyakit stroke dan menemukan kNN lebih akurat. Penelitian ini menyoroti pentingnya memilih algoritma yang tepat untuk meningkatkan akurasi prediksi dan diagnosis penyakit.



UMMN

UNIVERSITAS  
MULTIMEDIA  
NUSANTARA