MIXED DISTRIBUTION PADA *NAIVE BAYES* UNTUK DETEKSI PENYAKIT *STROKE*

SKRIPSI

Oleh:
MUHAMMAD KHOIRUL HUDA
NIM. 200605110085

PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM
MALANG
2024

MIXED DISTRIBUTION PADA *NAIVE BAYES* UNTUK DETEKSI PENYAKIT *STROKE*

SKRIPSI

Diajukan kepada:
Fakultas Sains dan Teknologi
Universitas Islam Negeri Maulana Malik Ibrahim Malang
Untuk memenuhi Salah Satu Persyaratan dalam
Memperoleh Gelar Sarjana Komputer (S.Kom)

Oleh:
MUHAMMAD KHOIRUL HUDA
NIM. 200605110085

PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM MALANG
2024

HALAMAN PERSETUJUAN

MIXED DISTRIBUTION* PADA *NAIVE BAYES* UNTUK DETEKSI PENYAKIT *STROKE

SKRIPSI

Oleh:
MUHAMMAD KHOIRUL HUDA
NIM. 200605110085

Telah Diperiksa dan Disetujui untuk Diuji:
Tanggal: 12 Desember 2024

Pembimbing I,	Pembimbing II,
Dr. Irwan Budi Santoso, M.Kom	Dr. Cahyo Crysdian, M.CS
NIP. 19770103 201101 1 004	NIP. 19740424 200901 1 008

Mengetahui,
Ketua Program Studi Teknik Informatika
Fakultas Sains dan Teknologi
Universitas Islam Negeri Maulana Malik Ibrahim Malang

Dr. Ir. Fachrul Kurniawan, M.MT, IPU
NIP. 19771020 200912 1 001

HALAMAN PENGESAHAN

MIXED DISTRIBUTION PADA *NAIVE BAYES* UNTUK DETEKSI PENYAKIT *STROKE*

SKRIPSI

Oleh:
MUHAMMAD KHOIRUL HUDA
NIM. 200605110085

Telah Dipertahankan di Depan Dewan Penguji Skripsi
dan Dinyatakan Diterima Sebagai Salah Satu Persyaratan
Untuk Memperoleh Gelar Sarjana Komputer (S.Kom)
Tanggal: 20 Desember 2024

Susunan Dewan Penguji

Ketua Penguji	: Dr. Zainal Abidin, M.Kom NIP. 19760613 200501 1 004	( )
Anggota Penguji I	: Ajib Hanani, M.T NIP. 19840731 202321 1 013	( )
Anggota Penguji II	: Dr. Irwan Budi Santoso, M.Kom NIP. 19770103 201101 1 004	( )
Anggota Penguji III	: Dr. Cahyo Crysdian, M.CS NIP. 19740424 200901 1 008	( )

Mengetahui dan Mengesahkan,
Ketua Program Studi Teknik Informatika
Fakultas Sains dan Teknologi
Universitas Islam Negeri Maulana Malik Ibrahim Malang

Dr. Fachrul Kurniawan, M.MT, IPU
NIP. 19771020 200912 1 001

PERNYATAAN KEASLIAN TULISAN

Saya yang bertanda tangan di bawah ini:

Nama	: Muhammad Khoirul Huda
NIM	: 200605110085
Fakultas / Program Studi	: Sains dan Teknologi / Teknik Informatika
Judul Skripsi	: *Mixed Distribution* Pada *Naive Bayes* Untuk Deteksi Penyakit *Stroke*.

Menyatakan dengan sebenarnya bahwa Skripsi yang saya tulis ini benar-benar merupakan hasil karya saya sendiri, bukan merupakan pengambil alihan data, tulisan, atau pikiran orang lain yang saya akui sebagai hasil tulisan atau pikiran saya sendiri, kecuali dengan mencantumkan sumber cuplikan pada daftar pustaka.

Apabila dikemudian hari terbukti atau dapat dibuktikan skripsi ini merupakan hasil jiplakan, maka saya bersedia menerima sanksi atas perbuatan tersebut.

Malang, 23 Desember 2024
Yang membuat pernyataan,

Muhammad Khoirul Huda
NIM. 200605110085

MOTTO

"Mereka yang melihat dirinya sebagai orang-orang yang teraliensi dari proses perubahan dunia, akan terperangkap dalam sirkuit gelisa kemelut dan stress. Mereka akan menyalahkan orang lain, bukan dirinya, sebagai penyebabnya"
(KH. HUSEIN MUHAMMAD).

"Ya Allah jangan sibukkan piliranku dengan urusan yang membuat urusanku cemas, Jangan sibukan hatiku dengan orang-orang yyang itdak punya hati kekasih, Jangan sibukkan hari-hariku dengan urusan yang tak berguna"
(KH. HUSEIN MUHAMMAD).

"Seseorang hendaklah mengisi hatinya dengan cahaya akal dan melihat realitas bukan menjadi hamba teks"
(Rumi).

HALAMAN PERSEMBAHAN

Alhamdulillahi Rabbil 'Alamin puji syukur atas kehadirat Allah SWT atas segala nikmat dan karunia-Nya. Sholawat serta salam selalu tercurahkan kepada Nabi Muhammad SAW. Saya Muhammad Khoirul Huda mempersembahkan karya ini kepada:

1. Orang tua (Ibu Tukiyami Lestari dan Bapak Achmad Shaiko).
2. Saudara kandung (Muhammad Saiful Arifin dan Ulfa Nur Mahmudah).
3. Sahabat – sahabat semua.
4. Teman – teman seperjuangan, (“INTEGER”) Teknik Informatika angkatan 2020.

KATA PENGANTAR

Segala Puji dan Syukur penulis panjatkan kepada Allah SWT atas segala karunia-Nya sehingga penulis dapat menyelesaikan penulisan skripsi yang berjudul "*Mixed Distribution* Pada *Naive Bayes* Untuk Deteksi Penyakit *Stroke*". Shalawat serta salam selalu tercurahkan kepada Nabi Muhammad SAW.

Penulis mengucapkan banyak terima kasih kepada seluruh pihak yang telah membantu baik berupa motivasi, bimbingan moril maupun meteril, yang ditujukan kepada:

1. Prof. Dr. H. M. Zainuddin, M.A., selaku Rektor Universitas Islam Negeri Maulana Malik Ibrahim Malang.
2. Prof. Dr. Hj. Sri Hariani, M.Si., selaku Dekan Fakultas Sains dan Teknologi Universitas Islam Negeri Maulana Malik Ibrahim Malang.
3. Dr. Ir. Fachrul Kurniawan, M.MT, IPU, selaku Ketua Program Studi Teknik Informatika dan Dr. Zainal Abidin, M.Kom, selaku Sekretaris Program Studi Teknik Informatika Universitas Islam Negeri Maulana Malik Ibrahim Malang.
4. Dr. Irwan Budi Santoso, M.Kom, selaku dosen pembimbing I dan Dr. Cahyo Crysdian, M.CS, selaku dosen pembimbing II yang telah memberikan bimbingan arahan dan ilmu kepada penulis, sehingga bisa menyelesaikan skripsi.
5. Dr. Zainal Abidin, M.Kom dosen ketua penguji dan Ajib Hanani, M.T selaku dosen penguji I yang telah menguji dan memberikan saran evaluasi kepada penulis sehingga dapat menyelesaikan skripsi dengan baik.

6. Ibu Tukiyami Lestari dan Bapak Achmad Shaiko selaku orang tua penulis yang selalu memberikan doa, dukungan, dan motivasi kepada penulis.
7. Saudara Kandung (Muhammad Saiful Arifin dan Ulfa Nur Mahmudah)
8. Segenap Dosen Program Studi Teknik Informatika Universitas Islam Negeri Maulana Malik Ibrahim Malang.
9. Teman-teman Teknik Informatika 2020 “INTEGER” yang telah memberikan motivasi dan semangat selama penulisan skripsi.
10. Diri sendiri yang telah berjuang hingga berhasil menyelesaikan skripsi ini.

Malang, 23 Desember 2024

Penulis

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

ABSTRAK

Huda, Muhammad Khoirul. 2024. ***Mixed Distribution* Pada *Naive Bayes* Untuk Deteksi Penyakit *Stroke*.** Skripsi. Program Studi Teknik Informatika Fakultas Sains dan Teknologi Universitas Islam Negeri Maulana Malik Ibrahim Malang. Pembimbing: (I) Dr. Irwan Budi Santoso, M.Kom. (II) Dr. Cahyo Crysdian, M.CS.

Kata kunci: *Gaussian Naive Bayes, Mixed Distribution, Multinomial Naive Bayes, Penyakit Stroke.*

Mixed Distribution pada *Naive Bayes* merupakan pendekatan yang mengintegrasikan dua jenis distribusi probabilitas, yaitu distribusi numerik (*kontinu*) dan distribusi kategori (diskrit), dalam algoritma *Naive Bayes*. Metode ini berhasil mengatasi keterbatasan model klasifikasi konvensional dengan mampu mengakomodasi variasi distribusi data dari berbagai variabel prediktor. Pada penelitian ini dilakukan dengan mengumpulkan dataset klinis yang mencakup berbagai variabel risiko *stroke*, seperti *Gender, Age, Hypertension, Heart Disease, Ever Married, Work Type, Residence Type, Avg Glucose Level, BMI Smoking Status, Stroke.* Hasil penelitian menunjukkan bahwa model *Naive Bayes* dengan *Mixed Distribution* memberikan akurasi deteksi *stroke* yang lebih tinggi dibandingkan dengan metode klasifikasi tunggal dengan hasil *accuracy* 76,45%, *precesion* 71,42%, *recall* 87,42%, dan *F1-Score* 78,61%.

ABSTRACT

Huda, Muhammad Khoirul. 2024. **Distribution in Naive Bayes for Stroke Disease Detection**. Undergraduate Thesis. Informatics Engineering Study Program, Faculty of Science and Technology, Maulana Malik Ibrahim State Islamic University Malang. Advisors: (I) Dr. Irwan Budi Santoso, M.K om. (II) Dr. Cahyo Crysdian, M.Cs.

Mixed Distribution in Naive Bayes is an approach that integrates two types of probability distributions, namely numerical distribution (continuous) and categorical distribution (discrete), in the Naive Bayes algorithm. This method successfully overcomes the limitations of conventional classification models by being able to accommodate variations in data distribution from various predictor variables. This study was conducted by collecting clinical datasets that include various stroke risk variables, such as Gender, Age, Hypertension, Heart Disease, Ever Married, Work Type, Residence Type, Avg Glucose Level, BMI Smoking Status, Stroke. The results show that the Naive Bayes model with Mixed Distribution provides higher stroke detection accuracy compared to a single classification method with the results of 76.45% accuracy, 71.42% precesion, 87.42% recall, and 78.61% F1-Score.

Keywords: Gaussian Naive Bayes, Mixed Distribution, Multinomial Naive Bayes, Stroke Disease.

مستخلص البحث

هدى، محمد خويرول. ٢٠٢٤. التوزيع المختلط في نظام باي الساذج للكشف عن مرض السكتة الدماغية. الأطروحة. برنامج دراسة هندسة المعلوماتية، كلية العلوم والتكنولوجيا، الجامعة الإسلامية الحكومية، مولانا مالك إبراهيم مالانج. المشرف: (الأول) د. م. عين اليقين، م. كوم. (الثاني) د. كاهيو كريسديان، ماجستير

التوزيع المختلط، بايز الساذج الغاوسي، بايز الساذج متعدد الحدود، مرض السكتة الدماغية

التوزيع المختلط في بايز الساذج هو نهج يدمج نوعين من التوزيعات الاحتمالية، وهما التوزيع العددي (المستمر) والتوزيع الفئوي (المنفصل)، في خوارزمية بايز الساذج. تتغلب هذه الطريقة بنجاح على قيود نماذج التصنيف التقليدية من خلال قدرتها على استيعاب الاختلافات في توزيع البيانات من مختلف المتغيرات المتنبئة. أُجريت هذه الدراسة من خلال جمع مجموعات بيانات سريرية تتضمن متغيرات مخاطر السكتة الدماغية المختلفة، مثل الجنس، والعمر، وارتفاع ضغط الدم، وأمراض القلب، والزواج من قبل، ونوع العمل، ونوع الإقامة، ومستوى الجلوكوز المتوسط، ومؤشر كتلة الجسم، وحالة التدخين، والسكتة الدماغية. أظهرت النتائج أن نموذج بايز الساذج مع التوزيع المختلط يوفر دقة أعلى في اك تشاف السكتة الدماغية مقارنة بطريقة التصنيف الأحادية حيث بلغت نسبة F1- Score 78.61% الدقة 76.45%، والتوقع 71.42%، والتذكر 87.42%، ودرجة.

BAB I

PENDAHULUAN

1.1 Latar Belakang

Mixed Distribussion atau distribusi campuran pada *Naive Bayes* adalah pendekatan yang menggabungkan distribusi *Gaussian* dan distribusi *Multinomial* dalam satu model *Naive Bayes*. Pendekatan ini memungkinkan kita untuk mengatasi variasi fitur dalam dataset yang terdiri dari campuran fitur numerik, kategorikal. Penerapan *Mixed Distribution*. Dengan menggabungkan metode *Naive Bayes* dengan distribusi *Gaussian* dan *Multinomial* memungkinkan fleksibilitas dalam menangani berbagai jenis fitur pada data klasifikasi. Hal ini memberikan kemampuan untuk menangani data dengan fitur campuran, mengurangi risiko *overfitting*, dan memanfaatkan keunggulan masing-masing distribusi untuk meningkatkan kualitas prediksi model. Selain itu, pendekatan ini relatif mudah diimplementasikan dan dapat mengatasi asumsi independensi dalam *Naive Bayes* dengan menggunakan distribusi yang sesuai untuk setiap jenis fitur dengan ciri-ciri persamaan dan perbedaan menggunakan statistik yang bisa memprediksi probabilitas sebuah kelas (Lestari et al., 2020).

Penyakit *stroke* menjadi masalah kesehatan utama bagi masyarakat dunia di modern saat ini. Hal tersebut karena penyakit *stroke* menjadi penyebab kematian tertinggi nomor dua di dunia dan menjadi penyebab utama kecacatan. Berdasarkan data dari WHO ada 70% angka kematian secara global disebabkan oleh *stroke*, dan 87% kematian akibat penyakit *stroke* yang terjadi di negara-negara berpenghasilan

rendah dan menengah (Byna & Basit, 2020). *Stroke* merupakan salah satu penyakit serius yang melanda hampir seluruh dunia. Kejadian *stroke* yang mendadak bisa berakibat fatal dan mengakibatkan kecacatan fisik dan mental, baik pada individu yang sehat maupun mereka yang berusia lanjut (Sinaga et al., 2023). Masalah *stroke* di Indonesia memerlukan perhatian serius karena jumlah kasusnya terus bertambah, disertai tingginya angka kematian. Berdasarkan laporan Riskesdas 2018, prevalensi *stroke* nasional tercatat cukup tinggi, yaitu 10,9 per mil. Salah satu daerah dengan angka prevalensi *stroke* yang tinggi adalah Provinsi Sulawesi Selatan, mencapai 10,6 per mil. (Byna & Basit, 2020).

Gejala *stroke* dapat bervariasi tergantung pada bagian otak yang terkena dan seberapa parah kerusakan yang terjadi. Gejala awal *stroke* seringkali muncul tiba-tiba dan dapat berakibat kesulitan bicara, kelemahan pada satu sisi tubuh, kehilangan keseimbangan, dan gangguan penglihatan. Pemahaman terkait resiko penyakit *stroke* sangat penting baik dari segi faktor-faktor seperti tekanan darah, diabetes, riwayat keluarga, kebiasaan merokok, tingkat aktivitas fisik, kadar kolesterol, tinggi badan, berat badan, serta riwayat fibrilasi atrium. (Susilawati et al., 2021). Mengenali gejala-gejala ini dengan cepat memungkinkan seseorang untuk segera mencari perawatan medis. Penting untuk segera mendapatkan pertolongan medis saat mengalami gejala *stroke* karena pengobatan yang cepat dapat mengurangi kerusakan otak dan meningkatkan peluang pemulihan. Untuk mendeteksi *stroke* memerluhkan tahapan dan waktu yang relative lama dengan melibatkan evaluasi menyeluruh terhadap gejala, pemeriksaan fisik, pemeriksaan penunjang, dan penilaian faktor risiko. Cara alternatif lain untuk melakukan proses

diagnosis penyakit *stroke* tanpa harus berkonsultasi langsung kepada dokter dan melakukan pengecekan ke laboratorium adalah sebuah sistem yang berperan untuk mendeteksi atau mendiagnosis penyakit *stroke* (Kanggeraldo et al., 2018). Langkah-langkah ini membantu dokter dalam menentukan diagnosis yang akurat dan merencanakan penanganan yang tepat sesuai dengan kondisi pasien. Pentingnya tindakan pencegahan dan penanganan sejak dini seperti yang terkandung dalam Al-Qur'an Surat Yunus Ayat 57:

﴿ يَٰٓأَيُّهَا ٱلنَّاسُ قَدْ جَآءَتْكُم مَّوْعِظَةٌ مِّن رَّبِّكُمْ وَشِفَآءٌ لِّمَا فِى ٱلصُّدُورِ وَهُدًى وَرَحْمَةٌ لِّلْمُؤْمِنِينَ ٥٧ ﴾

"Hai manusia, sesungguhnya telah datang kepadamu pelajaran dari Tuhanmu dan penyembuh bagi penyakit-penyakit (yang berada) dalam dada dan petunjuk serta rahmat bagi orang-orang yang beriman" (Q.S Yunus: 57).

Ibnu Asyur berpendapat bahwa telah datang kepada manusia bahwa telah diturunkan Al-Qur'an dan dibawakan bagi sebagai petunjuk. Pada ayat ini membahas empat poin pokok, yaitu; Pertama, Al-Qur'an berfungsi sebagai nasihat dan pelajaran; kedua, menjadi penyembuh bagi segala penyakit hati dan jiwa; ketiga, berperan sebagai petunjuk (*huda*); dan keempat, merupakan rahmat bagi orang-orang yang beriman (Suhaili et al., 2022).

Meskipun Al-Qur'an memberikan pedoman dan rahmat bagi orang-orang yang beriman, seperti yang diutarakan oleh Ibnu Asyur, terdapat pula tantangan praktis di bidang kesehatan yang perlu diperhatikan. Ada beberapa hambatan yang membuat pasien enggan ke rumah sakit antara lain kurangnya pengetahuan, kesulitan dalam pengambilan keputusan, dan masalah keuangan. Oleh karena itu, dianjurkan agar layanan kesehatan memberikan edukasi tentang penanganan pra-rumah sakit setelah deteksi dini *stroke* kepada masyarakat, serta menyediakan

fasilitas transportasi darurat guna mencegah peningkatan angka kecacatan dan kematian akibat *stroke*. Selain itu, deteksi dini penyakit *stroke* dapat dengan mengenali gejala awal, menilai faktor risiko, dan memulai pengobatan pencegahan lebih awal, deteksi dini membantu dalam meminimalkan dampak jangka panjang dari *stroke*, serta meningkatkan peluang pemulihan yang baik. Kesadaran masyarakat tentang gejala dan faktor risiko *stroke*, serta pemeriksaan rutin dan konsultasi medis secara berkala, juga penting dalam upaya deteksi dini dan pencegahan penyakit ini. Pada saat ini ada banyak penelitian tentang deteksi penyakit *stroke*. Penelitian dilakukan dengan menerapkan *Machine Learning*. Salah satu metode yang sering digunakan pada algoritma klasifikasi dan deteksi adalah metode *Naive Bayes*. Metode ini terkenal karena kemampuannya dalam memproses data dengan cepat dan efisien.

Berdasarkan latarbekang tersebut yang telah dijelaskan diatas, maka penulis mengusulkan sebuah sistem *Mixed Distribution* baik diskrit dan kontinu pada *Naive Bayes* untuk deteksi penyakit *stroke*. Tujuan utama pada penelitian ini adalah untuk meningkatkan *accuracy* deteksi penyakit *stroke* melalui pengembangan metode menggunakan algoritma *Naive Bayes* dengan menerapkan distribusi campuran. Penelitian ini juga bertujuan untuk memberika kontribusi pada bidang ilmu pengetahuan terutama pada bidang kesehatan yang di implementasikan dengan *mechine learning* menggunakan algoritma *Naive Bayes* dengan *Mixed Distribution* yang merupakan gabungan dari *Gaussian* dan *Multinomial*.

1.2 Rumusan Masalah

Bagaimana hasil *accuracy, precesion, recall,* dan *F1-Score* dari penerapan *Mixed Distribution* pada *Naive Bayes* untuk deteksi penyakit *stroke* berdasarkan data medis.

1.3 Batasan Masalah

Adapun batasan masalah pada penelitian ini, yaitu:

1. Data penelitian ini berasal dari *Electronic Health Record* (EHR) oleh *McKinsey* & *Company* yang berasal dari *website Kaggle brain stroke prediction* dataset oleh (Izzet Turkalp Akbasli).
2. Data yang tidak seimbang antara kelas penderita *stroke* dengan kelas tidak penderita *stroke,* maka perlu dilakukan teknik SMOTE data.

1.4 Tujuan Penelitian

Berdasarkan pernyataan masalah yang ada ditujukaan sebelumnya, maka tujuan yang dicapai pada penelitian ini adalah untuk mengetahui hasil *accuracy, precesion, recall,* dan *F1-Score* deteksi penyakit *stroke* dengan menerapkan *Mixed Distribution* pada *Naive Bayes.*

1.5 Manfaat Penelitian

Penelitian ini diharapkan memberikan manfaat sebagai berikut:

1. Kontribusi pada bidang pendidikan, penelitian ini dapat memberikan wawasan, pengetahuan, dan memberikan konstribusi pada penelitian lanjutan tentang penerapan *Mixed Distribution* dengan menggunakan metode *Naive Bayes* untuk deteksi penyakit *stroke*.

2. Kontribusi pada bidang medis, dapat dijadikan sebagai acuan dalam sebuah sistem untuk mendiagnosa penyakit *stroke* dengan parameter-parameter tertentu.

BAB II

STUDI PUSTAKA

2.1 Penelitian Penyakit *Stroke*

Penelitian yang dilakukan oleh (Susilawati et al., 2021) membahas tentang pengembangan sistem pakar untuk deteksi penyakit *stroke* menggunakan metode *Naive Bayes-Certainty Factor*. Proses deteksi dilakukan dengan cara menghitung nilai probabilitas disetiap kelas dan memilih nilai terbesar sebagai hasil penggunaan metode *Naive Bayes*. Selain itu, metode *Certainty Factor* digunakan untuk menghitung nilai kepastian dari hasil perhitungan *Naive Bayes*. Hasil *accuracy* yang didapat pada penelitian ini adalah 84%. *Accuracy* dihitung berdasarkan hasil deteksi antara sistem dan pakar menggunakan 25 data *training*, di mana 19 data *testing* menghasilkan hasil yang sama antara keduanya.

Penelitian yang dilakukan oleh (Karim et al., 2021) tentang sistem pakar untuk mengidentifikasi jenis penyakit *stroke* berdasarkan gejala-gejala pada pasien metode yang digunakan *Naive Bayes Classifier*. *Dataset* terdiri dari data jenis penyakit *stroke* yaitu *stroke* hemorogik dan *stroke* iskemik. Data gejala penyakit *stroke* yang mencakup 13 gejala berdasarkan pemeriksaan klinis awal. Penggunaan metode *Naive Bayes* digunakan untuk mengklasifikasikan data dan memberikan *output* berupa kemungkinan jenis penyakit *stroke* berdasarkan gejala yang dipilih. Distribusi pada penelitian ini menggunakan distribusi *Gaussian Naive Bayes*. Hasil *accuracy* dari penerapan metode *Naive Bayes* pada aplikasi sistem pakar ini sebesar 80%.

Penelitian yang dilakukan oleh (Mualfah et al., 2022) tentang deteksi penyakit *stroke* menggunakan algoritma *Random Forest* dengan mengatasi ketidakseimbangan data menggunakan teknik SMOTE. *Dataset* yang digunakan terdiri dari 3.400 *record* pasien, di mana 783 di antaranya mengalami *stroke* dengan 12 atribut seperti jenis kelamin, usia, hipertensi, tipe tempat tinggal, rata-rata glukosa darah, indeks massa tubuh, status merokok, dan label *stroke*. Penerapan metode dilakukan tahapan *Handling Missing Values* untuk mengatasi nilai yang hilang dalam dataset. Selanjutnya, dilakukan metode *Resampling* menggunakan teknik SMOTE untuk menangani ketidakseimbangan kelas pada kolom *stroke*. *Splitting* data dilakukan untuk memisahkan dataset menjadi data *training* dan data *testing*. *Modeling* menggunakan algoritma *Random Forest* dengan parameter yang telah ditentukan. Hasil *accuracy* algoritma *Random Forest* tanpa teknik SMOTE *accuracy* sebesar 98% sedangkan algoritma *Random Forest* dengan SMOTE *accuracy* sebesar 91%.

Penelitian yang dilakukan oleh (Sinaga et al., 2023) tentang analisis prediksi penyakit *stroke* dengan pendekatan *Exploratory Data Analysis* (EDA) dengan algoritma *Support Vector Machine* (SVM) dan RFC. Tujuan penelitian ini untuk melakukan analisis hubungan antara kator resiko terjadinya *stroke* dan menganalisis perfoma algoritma *machine learning* dalam memprediksi deteksi *stroke*. Dataset yang digunakan dalam penelitian ini berasal dari *kaggle* dengan judul "*healthcare-dataset-stroke-data*" yang terdiri dari 5.110 observasi (baris) dengan 12 atribut atau kolom yang berkaitan dengan status kesehatan pasien, seperti jenis kelamin, hipertensi, penyakit jantung, status perkawinan, jenis pekerjaan, jenis tempat

tinggal, status merokok, riwayat *stroke*, *age*, BMI, dan kadar glukosa rata-rata. Penerapan metode diawali dengan tahap *Exploratory Data Analysis* (EDA) untuk memahami dataset yang digunakan. Selanjutnya, dilakukan pemodelan prediktif menggunakan algoritma SVM dan RFC. Hasil *accuracy* menunjukan bahwa model RFC memiliki perfoma yang lebih baik daripada model SVM dengan skor *accuracy* 99% dan 98%.

Penelitian yang dilakukan oleh (Praja Utama et al., 2023) tentang penerapan kombinasi algoritma *Naive Bayes* dan *Forward Selection* untuk memprediksi penyakit *stroke*. Data yang dianalisis mencakup 5.110 catatan medis dengan 12 variabel, meliputi: jenis kelamin, usia, hipertensi, penyakit jantung, status pernikahan, jenis pekerjaan, tipe tempat tinggal, tingkat glukosa rata-rata, BMI, dan kebiasaan merokok. Dalam penelitian ini, algoritma *Naive Bayes* yang diterapkan menggunakan dua varian: *Bernoulli Naive Bayes* dan *Gaussian Naive Bayes*. Metode ini bekerja dengan menghitung probabilitas awal dan probabilitas bersyarat untuk setiap kemungkinan klasifikasi. Proses validasi dilakukan dengan membagi dataset menjadi dua bagian: satu bagian untuk data agregat dari data awal dan bagian lainnya untuk validasi. Selanjutnya, klasifikasi dilakukan menggunakan *Forward Selection* dan *Cross Validation* untuk mengukur performa model melalui beberapa metrik, yaitu *accuracy, precesion, recall*, dan *F1-Score*. Hasil pengujian menunjukkan bahwa algoritma *Naive Bayes* mencapai tingkat akurasi sebesar 95.13%.

Penelitian yang dilakukan oleh (Tomasouw & Rumlawang, 2023) membahas penerapan metode *Support Vector Machine* (SVM) untuk deteksi dini

risiko stroke berdasarkan data rekam medis pasien, yang mencakup faktor-faktor seperti tekanan darah, umur, LDL, dan gula darah. Data yang digunakan distandarisasi ke dalam skala [-1, 1] untuk meningkatkan akurasi. Penelitian ini menguji dua skema pembagian data, yaitu 60 data *training* 40 data *testing* dan 70 data *training* 30 data *testing*, dengan hasil akurasi tertinggi 81.25% menggunakan metode SVM linier pada skema 60 data *training* 40 data *testing*. Selain itu, metode SVM *nonlinier* dengan kernel *polinomial* dan RBF menunjukkan hasil yang lebih baik, di mana kernel RBF mencapai akurasi tertinggi sebesar 84.38%. Temuan ini menunjukkan bahwa penggunaan metode SVM, terutama dengan kernel RBF, dapat meningkatkan efektivitas deteksi dini risiko *stroke*.

Tabel 2.1 Penelitian terkait deteksi penyakit *stroke*

No	**Peneliti (tahun)**	**Metode**	**Distribusi**	**Hasil/ *Accuracy***
2	(Susilawati et al., 2021)	*Naive Bayes + Certainty Factor*	*Multinomial*	84%
3	(Karim et al., 2021)	*Naive Bayes*	*Gaussian*	80%
4	(Mualfah et al., 2022)	*Random Forest* SMOTE	*Gaussian + Multinomial*	91%
5	(Sinaga et al., 2023)	SVM dan RVC	*Gaussian + Multinomial*	90% dan 91%
6	(Praja Utama et al., 2023)	*Bernoulli Naive Bayes + Gaussian Naive Bayes* dan *Forward Selection*	*Bernoulli + Gaussian*	95%
7	(Tomasouw & Rumlawang, 2023)	*Support Vector Machine (SVM)*	*Gaussian*	81%

Berdasarkan tabel 2.1 penelitian terkait metode klasifikasi menunjukkan variasi akurasi yang dicapai berdasarkan teknik dan jenis distribusi yang digunakan. Kombinasi *Naive Bayes* dengan *Certainty Factor* oleh Susilawati et al. (2021) memberikan akurasi sebesar 84% dengan distribusi *Multinomial*, sedangkan *Naive Bayes Gaussian* oleh Karim et al. (2021) mencapai 80%. Metode *ensemble* seperti

Random Forest dengan SMOTE oleh Mualfah et al. (2022) menunjukkan performa tinggi dengan akurasi 91% pada distribusi *Gaussian* dan *Multinomial*. Sementara itu, kombinasi SVM dan RVC oleh Sinaga et al. (2023) mencapai akurasi masing-masing 90% dan 91%, dan gabungan *Bernoulli Naive Bayes*, *Gaussian Naive Bayes*, serta *Forward Selection* oleh Praja Utama et al. (2023) mencatat hasil tertinggi, yaitu 95%. Terakhir, penggunaan SVM oleh Tomasouw & Rumlawang (2023) menghasilkan akurasi 81% dengan distribusi *Gaussian*. Secara keseluruhan, pendekatan kombinasi dan teknik pengayaan, seperti SMOTE atau *Forward Selection*, cenderung memberikan hasil yang lebih baik, dengan distribusi *Gaussian* dan *Multinomial* sebagai pilihan utama.

2.2 Penyakit *Stroke*

Korosi Penyakit *stroke* dikenal sebagai *Cerebrovascular Accident* (CVA) merupakan kondisi medis yang terjadi karena kerusakan otak akibat dari gangguan dari suplai darah yang menuju ke otak (Karim et al., 2021) Penyakit *stroke* ditandai dengan gejala-gejala kehilangan fungsi otak karena terhentinya aliran darah menuju ke otak (Sinaga et al., 2023) Penyakit *stroke* disebabkan oleh penyumbatan (*stroke iskemik*) atau pecahnya (*stroke hemoragik*) pembuluh darah yang bertugas mengantarkan oksigen dan nutrisi ke otak. Penyakit *stroke* merupakan kondisi yang terkait dengan gangguan aliran darah ke otak, pada umumnya disebabkan oleh tersumbatnya atau pecahnya pembuluh darah dan pembekuan darah. Ini adalah jenis gangguan sirkulasi darah *non-traumatik* yang dapat menyebabkan berbagai gejala, seperti kelumpuhan pada bagian wajah atau tubuh, kesulitan berbicara, gangguan

penglihatan, serta perubahan dalam tingkat kesadaran (Hikmayanti Handayani et al., 2023).

Terdapat dua jenis *stroke*, yaitu *stroke* iskemik dan *stroke hemorogik*. *Stroke iskemik* merupakan pengurangan atau penyusutan oksigen atau nutrien yang diperlukan oleh otak, yang dapat disebabkan oleh blokir arteri otak atau kondisi lain yang mengganggu sistem saraf. *Stroke* hemoragik merupakan pengeluaran darah atau cairan cair ke dalam otak, yang dapat disebabkan oleh kondisi lain yang mengganggu sistem saraf, seperti aneurysm, tumor, atau thrombosis. Gejala-gejala dan faktor resiko yang mempengaruhi terjadinya penyakit *stroke* yaitu jenis kelamin, usia, tingkat pendidikan, riwayat hypertensi, kadar kolestrol darah obesitas, penyakit jantung koroner, kebiasaan merokok, konsumsi makanan yang mengandung banyak garam, kurang aktivitas olahraga.

2.3 Algoritma *Naive Bayes*

Algoritma *Naive Bayes* menggunakan prinsip *Teorema Bayes* dengan mengambil pendekatan "sederhana" yang menganggap setiap atribut dalam data bersifat independen atau tidak saling mempengaruhi. Walaupun dalam praktiknya asumsi ini jarang terjadi di situasi sebenarnya, *Naive Bayes* masih menjadi metode yang handal dan cepat dalam melakukan klasifikasi, khususnya ketika menghadapi data yang memiliki banyak variabel. bekerja berdasarkan *tebcorema bayes* dengan asumsi "*naif*" bahwa setiap fitur dalam data independen satu sama lain. Meskipun asumsi ini sering kali tidak terpenuhi dalam dunia nyata, *Naive Bayes* tetap menjadi salah satu metode klasifikasi yang efektif dan efisien, terutama dalam kasus data dengan dimensi tinggi (Karim et al., 2021). *Naive Bayes* merupakan suatu algortima

yang berfungsi untuk memperkirakan kemungkinan kejadian di masa depan berdasarkan data atau pengalaman yang telah terjadi sebelumnya. (Mahar et al., 2023). Kelebihan pada *Naive Bayesian* adalah dapat memprediksi probabilitas yang ada didalam suatu class keanggotaan. Pada konteks klasifikasi untuk menghitung probabilitas bahwa suatu kelas maka diketahui fitur objek, X = $\{X_k \mid C_J\}$ Rumus persamaan pada *Naive Bayes* menunjukan probabilitas bersyarat dari vektor fitur X pada kelas C. Secara umum, rumus ini menggambarkan asumsinya bahwa setiap fitur X_k adalah independent.

Rumus *Naive Bayes*:

$$P\,(X|C_i) = \prod_{i=1}^{n} P(\,X_i|C_j) \tag{2.1}$$

Keterangan:

$P\,(X \mid C_j)$	: Probabilitas bersayarat dari fitur-fitur *X* diberi kelas C_j
P	: Probabilitas
n	: Jumlah total fitur
i	: Index mewakili setiap fitur
$(X_i \mid C_j)$	: Probabilitas bersyarat dari fitur X_i diberi kelas C_j

Penggunaan distribusi *Gaussian* dalam *Naive Bayes* diterapkan ketika fitur-fitur dalam data memiliki jenis (tipe) distribusi normal. *Gaussian Naive Bayes* menjadi salah satu metode yang sering digunakan didalam menghitung probabilitas statistika (Nurul. et al., 2022). Algortima *Gaussian Naive Bayes* merupakan salah satu algoritma *Naive Bayes* Classifier yang dihitung berdasarkan distribusi normal. *Gaussian* digunakan pada klasifikasi data numerik dengan distribusi *Gaussian* dan data kategorikal (Pulungan et al., 2023). Pada probabilitas akhir setiap kelas dengan memasukan semua data atau nilai distribusi gaussian yang ada ke dalam satu kelas yang sama (Fadli et al., 2021).

Penggunaan *Multinomial Naive Bayes* didasarkaan pada fitur-fitur yang ada didalam data kategoori. *Multinomial Naive Bayes* adalah sebuah algoritma pembelajaran yang bersifat sederhana, yang menerapkan prinsip *Bayes* untuk melakukan klasifikasi atau menghitung probabilitas suatu kelas data. Dalam proses klasifikasinya, kelas yang memiliki nilai probabilitas paling tinggi akan dipilih sebagai hasil prediksi. Algoritma ini sangat efektif ketika digunakan untuk melakukan klasifikasi pada data yang memiliki fitur-fitur diskrit. (Yuda Lesmana & Andarsyah, 2022). Algoritma *Multinomial Naive Bayes* adalah metode pembelajaran berbasis probabilitas yang menggunakan prinsip *Teorema Bayes* dan sering diaplikasikan dalam pemrosesan bahasa alami (*Natural Language Processing/NLP*). Cara kerja algoritma ini berdasarkan pada frekuensi kemunculan kata dalam suatu dokumen. Dalam penggunaannya, model ini mempertimbangkan dua aspek penting: keberadaan suatu kata dalam dokumen (ada atau tidak ada) dan seberapa sering kata tersebut muncul dalam dokumen tersebut. (Yuyun et al., 2021). *Multinominal Naive Bayes* mengansumsikan inependensi diantara kemunculan data yang ada tanpa memperhitungkan urutan dan konteks informasi pada kalimat tetapi juga memperhitungkan jumlah kemunculan kata (Dhuhita & Zone, 2023).

BAB III

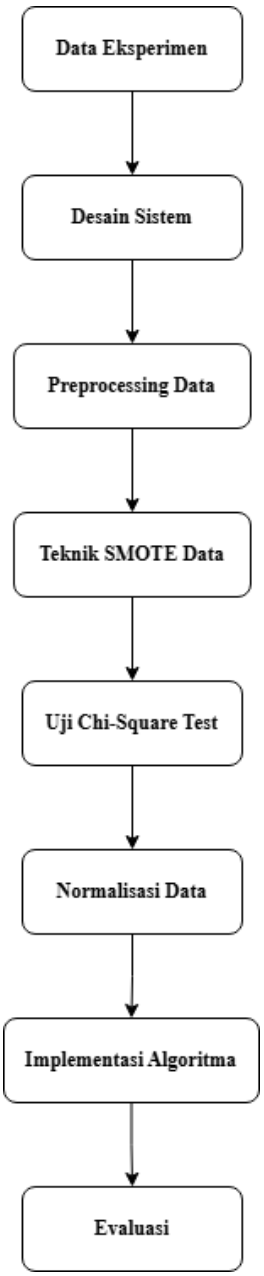
METODOLOGI PENELITIAN

3.1 Prosedur Penelitian

Prosedur penelitian merupakan elemen krusial dalam setiap proses penelitian, karena berfungsi sebagai panduan sistematis untuk mencapai tujuan penelitian secara efektif. Dengan menentukan jenis desain, metode pengumpulan data, *preprocessing*, teknik SMOTE, uji *Chi-Square Test*, normalisasi, dan algoritma yang tepat penelitian dapat menghasilkan model yang valid, akurat, dan memiliki performa yang baik.

Pada gambar 3.1 merupakan prosedur penelitian yang menggambarkan tahapan proses dalam sebuah sistem analisis data. Proses dimulai dengan pengumpulan data yang akan digunakan dalam penelitian atau pengembangan sistem. Selanjutnya, dilakukan tahap desain sistem untuk menentukan struktur dan alur kerja dari sistem yang akan dibangun. Setelah desain selesai, data mentah yang dikumpulkan diproses pada tahap *preprocessing*, yang meliputi pembersihan data dan transformasi agar siap untuk analisis lebih lanjut. Tahap berikutnya adalah menerapkan teknik SMOTE (*Synthetic Minority Over-sampling Technique*), dengan tujuan untuk mengatasi masalah ketidakseimbangan data dengan menambahkan data sintetik pada kelas minoritas. Setelah data seimbang, dilakukan uji *Chi-Square* untuk mengevaluasi relevansi atau hubungan antara variabel dalam dataset. Data kemudian dinormalisasi agar memiliki skala yang konsisten dan dapat digunakan dengan lebih efektif oleh algoritma. Setelah data siap, algoritma yang

telah dirancang diimplementasikan untuk melakukan analisis atau prediksi. Proses ini diakhiri dengan tahap evaluasi, di mana kinerja algoritma diukur berdasarkan hasil yang diperoleh untuk memastikan bahwa sistem bekerja sesuai dengan tujuan yang telah ditetapkan.



Gambar 3.1 Desain Penelitian

3.2 Data Eksperimen

Pada penelitian ini menggunakan data eksperimen yang berasal dari Kaggle. Berikut merupakan contoh dari data eksperimen.

Tabel 3.1 Dataset penyakit *stroke*

Gender	***Age***	***Hypertension***	***Heart Disease***	***Ever Married***	***Work Type***	***Residence Type***	***Avg Glucose Level***	**BMI**	***Smoking Status***	***Stroke***
Male	67.0	0	1	*Yes*	*Private*	*Urban*	228.69	36.6	*formerly smoked*	1
Male	80.0	0	1	*Yes*	*Private*	*Rural*	105.92	32.5	*never smoked*	1
Female	49.0	0	0	*Yes*	*Private*	*Urban*	171.23	34.4	*smokes*	1
Female	79.0	1	0	*Yes*	*Self-employed*	*Rural*	174.12	24.0	*never smoked*	1
Male	81.0	0	0	*Yes*	*Private*	*Urban*	186.21	29.0	*formerly smoked*	1
Male	74.0	1	1	*Yes*	*Private*	*Rural*	70.09	27.4	*never smoked*	1
Female	69.0	0	0	*No*	*Private*	*Urban*	94.39	22.8	*never smoked*	1
Female	78.0	0	0	*Yes*	*Private*	*Urban*	58.57	24.2	*Unknown*	1
Female	81.0	1	0	*Yes*	*Private*	*Rural*	80.43	29.7	*never smoked*	1
Female	61.0	0	1	*Yes*	*Govt_job*	*Rural*	120.46	36.8	*smokes*	1
Male	3.0	0	0	*No*	*children*	*Rural*	95.12	18.0	*Unknown*	0
Male	58.0	1	0	*Yes*	*Private*	*Urban*	87.96	39.2	*never smoked*	0
Female	8.0	0	0	*No*	*Private*	*Urban*	110.89	17.6	*Unknown*	0
Female	70.0	0	0	*Yes*	*Private*	*Rural*	69.04	35.9	*formerly smoked*	0
Female	52.0	0	0	*Yes*	*Private*	*Urban*	77.59	17.7	*formerly smoked*	0
Female	75.0	0	1	*Yes*	*Self-employed*	*Rural*	243.53	27.0	*never smoked*	0
Female	32.0	0	0	*Yes*	*Private*	*Rural*	77.67	32.3	*smokes*	0
Female	79.0	0	0	*Yes*	*Govt_job*	*Urban*	77.08	35.0	*Unknown*	0
Male	79.0	0	1	*Yes*	*Private*	*Urban*	57.08	22.0	*formerly smoked*	0
Female	37.0	0	0	*Yes*	*Private*	*Rural*	162.96	39.4	*never smoked*	0

Pada tabel 3.1 merupakan 20 sampel data *stroke* dan tidak *stroke* dari total 4981 data. Dataset yang digunakan pada penelitian ini adalah *Electronic Health Record* (EHR) oleh *McKinsey* & *Company* yang berasal dari website *Kaggle brain stroke prediction dataset* oleh (Izzet Turkalp Akbasli, 2022). Dataset ini bertipe diskrit dan kontinu yang terdiri dari 29.072 pasien dengan 11 atribut umum. Dari 11 atribut tersebut, 10 di antaranya adalah fitur input termasuk usia, jenis kelamin, status pernikahan, identitas pasien, tipe pekerjaan, tipe tempat tinggal (*urban*/*rural*), kondisi penyakit jantung dalam bentuk biner, indeks massa tubuh, status merokok pasien, tingkat glukosa, dan atribut biner hipertensi yang menunjukkan apakah pasien menderita hipertensi atau tidak. Pada atribut ke 10 pada dataset menjelaskan *output* biner yang menunjukan apakah seorang pasien telah mengalami *stroke* atau tidak.

Untuk penjelasan dari masing-masing atribut atau parameter yang digunakan pada dataset terdapat pada tabel 3.2 di bawah ini:

Tabel 3.2 Penjelasan dari atribut dataset

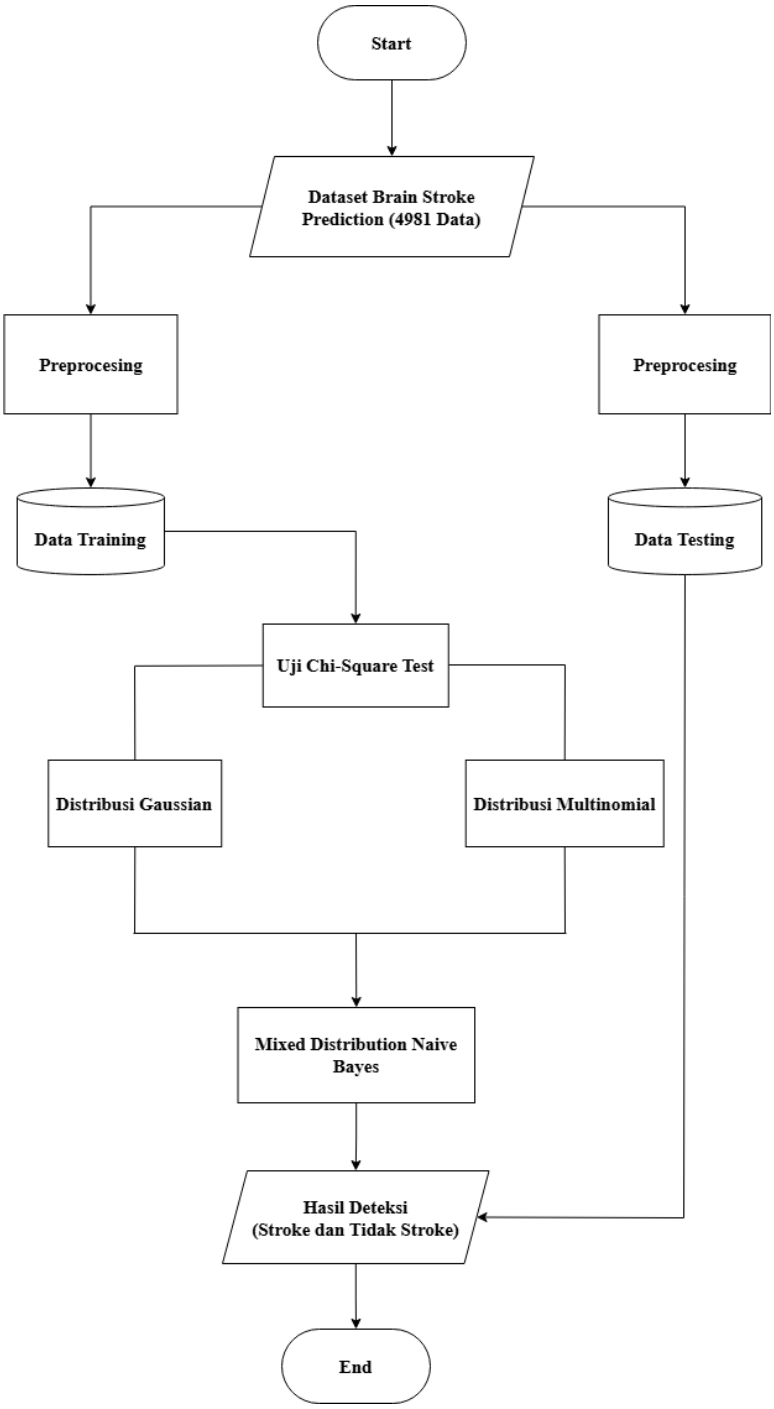
No	Nama Atribut	Deskripsi Atribut
1	*Gender*	Jenis kelamin (Pria “1” dan Wanita “0”).
2	*Age*	Umur
3	*Hypertension*	Memiliki riwayat penyakit Hypertensi dan Tidak
4	*Heart Disease*	Penyakit jantung (Memiliki Penyakit Jantung dan Tidak)
5	*Ever Married*	Faktor pernah menikah (Sudah Menikah dan Belum Menikah)
6	*Work Type*	Anak-anak, Pemerintahan, Tidak pernah berkerja, Swasta, Wiraswasta.
7	*Residence Type*	Perkotaan dan Perdesaan.
8	*Avg Glucose Level*	Rata-rata tingkat glukosa (*Avg Glucose Level*) mengacu pada konsentrasi rata-rata glukosa (gula) dalam darah seseorang selama periode waktu tertentu.
9	BMI	Indeks Massa Tubuh didasarkan pada berat badan dan tinggi badan yang ideal. Berat dengan satuan (kg) sementara tinggi m.
10	*Smoking Status*	"Status Merokok" mengacu pada 4 kategori yang digunakan; yaitu Pernah merokok, Tidak pernah merokok, Merokok, Tidak diketahui.
11	*Stroke*	Status penderita *stroke* (antara penderita *storke* dan tidak penderita *stroke*).

3.3 Desain Sistem

Desain sistem merupakan proses mendefinisikan arsitektur, komponen, modul, antarmuka, dan data untuk sebuah sistem agar memenuhi kebutuhan spesifik. Proses ini mencakup langkah-langkah untuk merancang bagaimana berbagai bagian sistem akan berinteraksi satu sama lain dan bagaimana mereka akan bekerja sama untuk mencapai tujuan yang diinginkan. Dalam tahap desain sistem, berbagai aspek teknis seperti arsitektur sistem, alur data, pemilihan metode, serta integrasi komponen dipertimbangkan dengan cermat untuk memastikan sistem mampu berjalan secara efisien dan memenuhi kebutuhan yang diharapkan. Dengan desain yang baik, implementasi sistem dapat dilakukan dengan lebih terarah dan minim kesalahan, sehingga mendukung pencapaian hasil yang optimal.

Pada gambar 3.2 menggambarkan proses deteksi *stroke* menggunakan metode *Mixed Distribution Naive Bayes* pada dataset prediksi *stroke* otak yang berisi 4981 data. Proses dilakukan dengan mengambil dataset, yang kemudian dibagi menjadi dua bagian, data *training data* dan data pengujian *testing data*. Kedua, bagian ini melalui tahap prapemrosesan untuk memastikan data siap digunakan. Setelah *preprocessing*, data pelatihan diuji menggunakan uji *Chi-Square Test* untuk menentukan relevansi fitur. Berdasarkan hasil uji tersebut, data dikelompokkan ke dalam dua jenis distribusi: *Gaussian* untuk data numerik (seperti usia, tingkat glukosa rata-rata, dan BMI) dan *Multinomial* untuk data kategorikal (seperti jenis kelamin, hipertensi, dan status merokok). Setelah itu, distribusi *Gaussian* dan *Multinomial* ini digabungkan dalam algoritma *Mixed Distribution Naive Bayes*, yang digunakan untuk melatih model. Model ini kemudian diterapkan

pada data pengujian untuk menghasilkan prediksi, yaitu mendeteksi apakah individu termasuk dalam kategori "*stroke*" atau "tidak *stroke*." Proses ini berakhir dengan menampilkan hasil deteksi.



Gambar 3.2 *Flowchart* desain sistem

3.3.1 *Preprocessing* Data

Data *cleaning* (pembersihan data) merupakan tahap kritis dalam proses *preprocessing* data yang bertujuan untuk memastikan bahwa data yang digunakan dalam analisis atau model pembelajaran mesin adalah bersih, konsisten, dan berkualitas tinggi. Proses ini melibatkan beberapa langkah, termasuk mengatasi nilai hilang dengan menggantinya atau menghapus baris yang terpengaruh, menghapus nilai duplikat untuk mencegah bias. Selain itu, pembersihan data mencakup identifikasi dan penanganan *outlier* yang dapat mengganggu analisis, memperbaiki kesalahan data seperti kesalahan penulisan, dan memastikan setiap kolom memiliki tipe data yang sesuai melalui konversi tipe data. Langkah-langkah ini membantu meningkatkan *accuracy* model, mengurangi bias, dan meningkatkan keandalan hasil analisis, sehingga menghasilkan data yang siap digunakan untuk analisis lanjutan dan pembelajaran mesin yang lebih efektif. Berikut beberapal langkah dalam *cleaning* data; mengatasi nilai hilang (*missing values*), mengatasi nilai duplikat (*duplicate values*).

3.3.2 Teknik SMOTE Data

Synthetic Minority Over-sampling Technique (SMOTE) merupakan sebuah metode dalam pengolahan data yang dikembangkan untuk menyelesaikan permasalahan ketimpangan kelas dalam suatu dataset. Ketimpangan kelas adalah kondisi dimana jumlah data pada satu kelas sangat sedikit dibandingkan dengan kelas lainnya. Kondisi ini dapat mengakibatkan model machine learning menjadi bias, dimana model akan lebih baik dalam mengenali kelas dengan data yang banyak namun kurang akurat dalam mengenali kelas dengan data yang sedikit.

3.3.3 Uji Distribusi Data dengan *Chi-Square Test (Goodness of Fit)*

Uji *Chi-Square* digunakan untuk memeriksa apakah distribusi data dari fitur-fitur tersebut sesuai dengan distribusi yang diharapkan. Hal ini penting untuk memastikan bahwa fitur-fitur yang digunakan dalam model benar-benar independen dan sesuai dengan asumsi *Naive Bayes*. Uji *Chi-Square* digunakan untuk menentukan apakah ada hubungan yang siknifikan antara dua variabel kategori. Berikut adalah langkah-langkah dan penjelasan rinci tentang uji *Chi-Square* untuk kesesuaian.

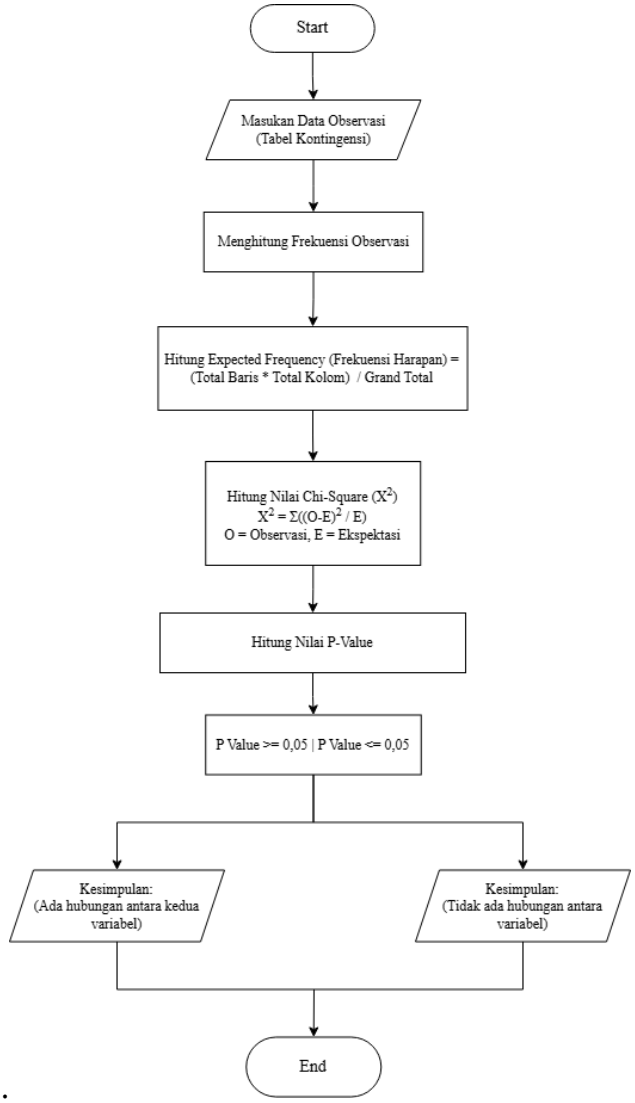
Menghitung nilai *Chi-Square* dengan menggunakan rumus sebagai berikut:

$$x^2 = \sum \frac{(O_i - E_i)^2}{E_i} \tag{3.1}$$

Keterangan:
x^2 : Distribusi *Chi-Square*
O_i : Frekuensi observasi untuk kategori ke-i
E_i : Frekuensi harapan untuk kategori ke-i
$\sum$: (Sigma) Jumlah untuk semua kategori

Pada gambar 3.3 menjelaskan langkah-langkah dalam *melakukan* uji *Chi-Square* (x^2) dalam bentuk diagram. Alur tersebut menjelaskan proses uji Chi-Square untuk menentukan hubungan antara dua variabel dalam sebuah dataset. Proses dimulai dengan memasukkan data observasi ke dalam tabel kontingensi, yang mencatat jumlah kejadian berdasarkan kategori dari masing-masing variabel. Langkah berikutnya adalah menghitung frekuensi observasi untuk setiap kategori dan menentukan frekuensi harapan (*expected frequency*) menggunakan rumus (Total Baris×Total Kolom)/Grand Total. Setelah itu, nilai *Chi-Square* x^2 dihitung dengan rumus 3.1 di mana O adalah frekuensi observasi dan E adalah frekuensi

harapan. Selanjutnya, nilai *p-value* dari hasil *Chi-Square* dihitung untuk menilai signifikansi hubungan antara kedua variabel. Jika nilai p-value kurang dari atau sama dengan 0,05, dapat disimpulkan bahwa terdapat hubungan yang signifikan antara kedua variabel. Sebaliknya, jika nilai p-value lebih besar dari 0,05, maka dapat disimpulkan bahwa tidak terdapat hubungan yang signifikan antara kedua variabel. Proses ini diakhiri dengan kesimpulan yang menunjukkan apakah kedua variabel memiliki hubungan atau tidak.



Gambar 3.3 Flowchart *uji distribusi* data

3.3.4 Normalisasi Data

Normalisasi data adalah proses mengubah skala nilai dari fitur numerik dalam dataset agar berada dalam rentang tertentu, biasanya antara 0 dan 1. Normalisasi ini penting dalam pembelajaran mesin untuk memastikan bahwa setiap fitur memiliki bobot yang sama dalam memengaruhi model, terutama jika perbedaan skala antar fitur cukup besar.

Ada beberapa metode normalisasi, salah satunya adalah *Min-Max Normalization*, yang umumnya digunakan untuk mengubah rentang data menjadi 0 hingga 1. Rumusnya adalah:

$$X_{scaled} = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{3.2}$$

Keterangan:
X : nilai asli yang ingin dinormalisasi
X_{min} : nilai minimum dalam fitur tersebut
X_{max} : nilai maksimum dalam fiturn tersebut
X_{scaled} : nilai hasil normalisasi dalam rentang 0 hingga 1

3.3.5 Gaussian Naive Bayes

Pada *Gaussian Naive Bayes* digunakan ketika fitur-fitur (atau atribut-atribut) dalam data diperlakukan sebagai variabel kontinu dan diasumsikan memiliki distribusi *Gaussian*. Jika fitur pada objek diketahui X = $\{x_1, x_2, \ldots, x_d\}$ dengan setiap fitur distribusi normal (*Gaussian*) maka peluang fitur atau atribut dengan syarat diketahui kelas ke-j (C_j) adalah persamaan sebagai berikut:

$$P(X \mid C_j) = \prod_{k=1}^{d} P(X_k \mid C_j) = \prod_{k=1}^{d} N(x_k; \hat{\mu}_{jk}, \hat{\sigma}_{jk}) \tag{3.3}$$

Keterangan:
$P\ (X \mid C_j)$: Probabilitas bersayarat dari fitur-fitur X diberi kelas C_j
d : Jumlah fitur
k : Index mewakili setiap fitur
$(X_k \mid C_j)$: Probabilitas bersyarat dari fitur X_k diberi kelas C_j

$N(x_k;\hat{\mu}_{jk},\hat{\sigma}_{jk})$: Menunjukan x_k ditribusi secara Normal (*Gaussian*) dengan parameter ($\hat{\mu}_{jk}$, $\hat{\sigma}_{jk}$)

Untuk $\hat{\mu}_{jk}$ dan $\hat{\sigma}_{jk}$ merupakan nilai estimasi parameter rerata dan simpangan baku untuk atribut ke-k dan kelas ke-j dengan menggunakan probabilitas bersyarat (Irwan Budi Santoso et al., 2024). Untuk tahap selanjutnya dapat dilakukan persamaan *Gaussian Naive Bayes* yang dirumuskan sebagai berikut:

$$P(C_j, X) = P(C_j)P(X|C_j) = P(C_j)\prod_{k=1}^{d} N(x_k;\hat{\mu}_{jk},\hat{\sigma}_{jk}) \quad (3.4)$$

Untuk Persamaan N sebagai berikut:

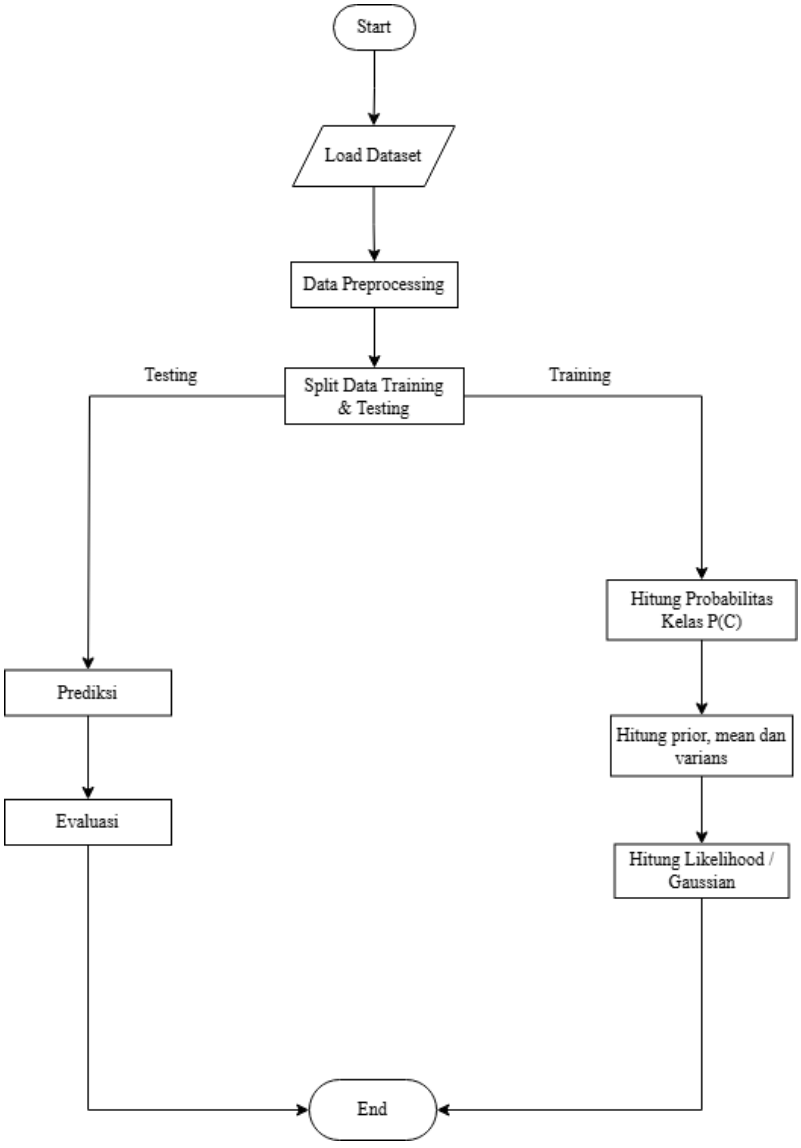
$$N = P(X = \vartheta|C_J) = \frac{1}{\sqrt{2\pi\hat{\sigma}_j^2}} \mathrm{e}^{-\frac{(\vartheta-\hat{\mu}_k)^2}{2\hat{\sigma}_k^2}} \quad (3.5)$$

Keterangan:

N	: Normal (*Gaussian*)
$P(C_j, X)$	: Probabilitas gabungan dari C_j dan fitur X
ϑ	: Nilai fitur yang diamati
$\hat{\sigma}_k^2$	: Variansi dari fitur dalam kelas C_J
$\hat{\mu}_{jk}$	: Mean (rata-rata) dari fitur X_k dalam kelas C_J
$\hat{\sigma}_{jk}$	: Standar deviasi dari fitur X_k dalam kelas C_J
Exp	: Fungsi eksponensial

Pada gambar 3.4 menjelaskan tentang alur proses lengkap dari algoritma *Gaussian Naive Bayes* untuk klasifikasi. Proses dimulai dari tahap "*Start*" sebagai langkah awal. Kemudian, data dimuat dari sumbernya pada tahap "*Load Dataset*". Setelah data dimuat, tahap "Data *Preprocessing*" dilakukan untuk membersihkan atau menyiapkan data agar sesuai untuk model, seperti menangani nilai hilang, mengubah data kategorikal menjadi numerik, atau normalisasi. Selanjutnya, data dibagi menjadi dua bagian: data latih (*training*) dan data uji (*testing*). Pada tahap *training*, dilakukan beberapa perhitungan, yaitu probabilitas kelas P(C), serta nilai

prior, mean, dan varians untuk setiap fitur berdasarkan kelasnya. Kemudian, *likelihood* atau distribusi probabilitas fitur dihitung menggunakan distribusi *Gaussian*, yang menjadi dasar model untuk memprediksi kelas data baru. Setelah model terlatih, data uji digunakan untuk mengukur performa model melalui prediksi dan evaluasi, dengan membandingkan hasil prediksi terhadap label asli. Proses ini diakhiri pada tahap "*End*," menandakan bahwa proses klasifikasi menggunakan *Gaussian Naive Bayes* selesai.



Gambar 3.4 *Flowchart Gaussian Naive Bayes*

3.3.6 Multinomial Naive Bayes

Multinomial merupakan generalisasi dari distribusi *Binomial* untuk lebih dari dua hasil yang mungkin. Distribusi ini sering digunakan untuk model kejadian yang dapat memiliki beberapa kategori. Pada algoritma *Naive Bayes* klasifikasi teks pada setiap data (nilai) direpresentasikan sebagai vektor fitur yang berisi frekuensi kata-kata. *Multinomial Naive Bayes* digunakan untuk memodelkan probabilitas kemunculan kata-kata dalam setiap kategori (kelas). Ketika digunakan bersama dengan asumsi "*Naive*" bahwa setiap fitur (kata) dalam dokumen adalah independen satu sama lain.

Secara matematis, faktorial dari jumlah total kejadian dari semua kategori di simbolkan sebagai ($x_1!,\ x_2!,\ \dots\ x_m!$) yang merupakan probabilitas kelas C_J. dengan probabilitas dari setiap kelas ($\theta_1^{x_1}, \theta_2^{x_2}, \dots \theta_m^{x_m}$) yang dapat dihitung menggunakan *Multinomial* (Santoso, 2013). Berikut merupakan penerapan rumus:

$$\binom{m}{x_1,\ x_2,\ \dots\ x_m} = \frac{m!}{x_1!,\ x_2!,\ \dots\ x_m!}$$

$$\sum_{i=1}^{k} x_i = m$$

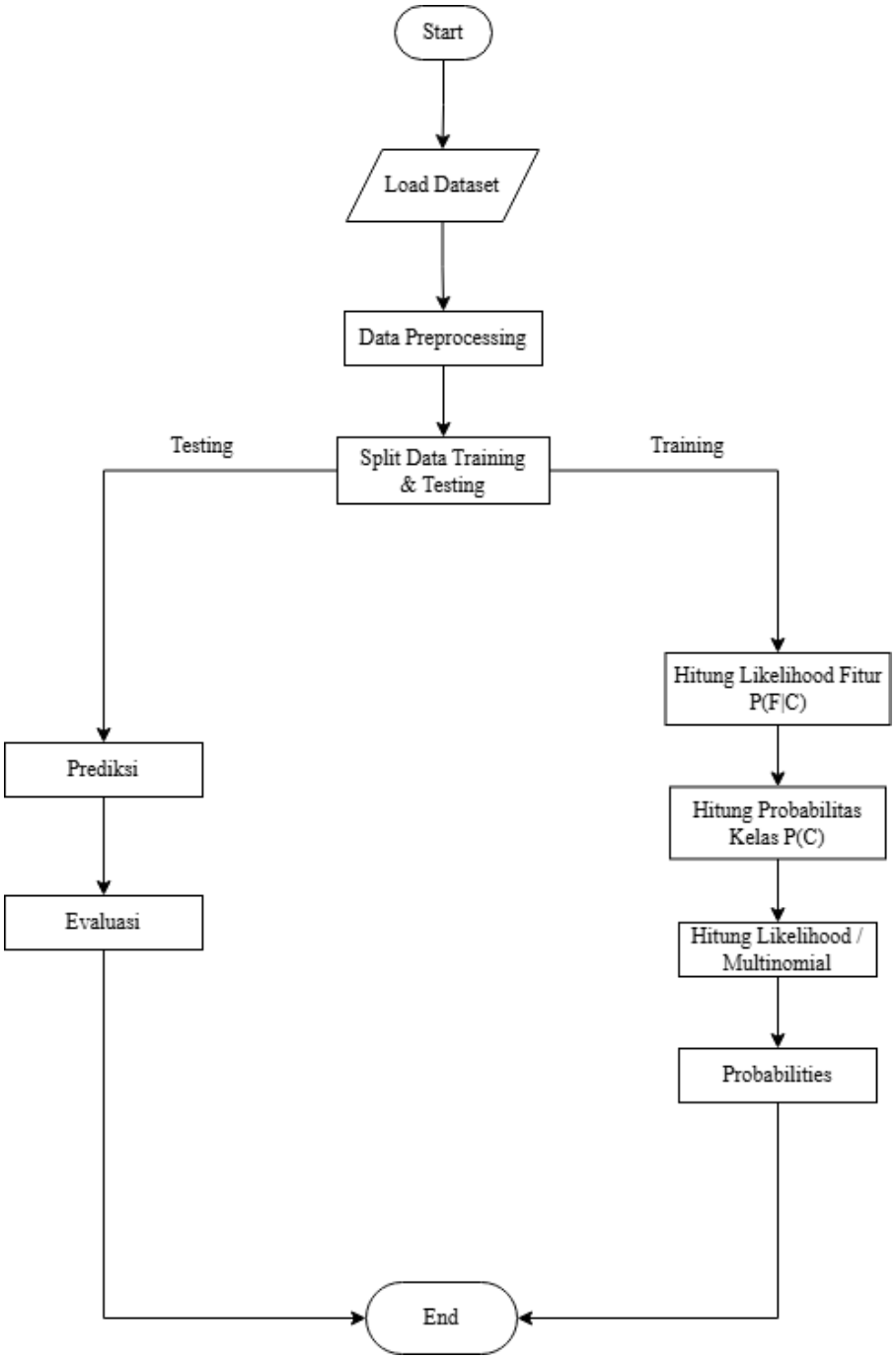
$$\sum_{i=1}^{k} \theta_i = 1$$

$$P(X \mid C_j) = \frac{(\sum_{i=1}^{m} x_i)!}{\prod_{i=1}^{m} x_i!} \prod_{i=1}^{m} \theta_i^{x_i} \tag{3.6}$$

$$P(X \mid C_j) = \prod_{i=1}^{m} \theta(x_i \mid C_j) \tag{3.7}$$

Keterangan:
x_m : Jumlah kemunculan kata (m) dalam dokumen X
$P\ (X \mid C_j)$: Probabilitas bersayarat dari fitur-fitur X diberi kelas C_j

m : Jumlah fitur (kejadian)
i : Index mewakili setiap fitur yang berjumlah (1)
θ_m : probabilitas dari masing-masing kategori atau hasil m
$\theta_i^{x_i}$: probabilitas dari masing-masing kejadian sesuai dengan distribusi



Gambar 3.5 *Flowchart Multinomial Naive Bayes*

Pada gambar 3.5 menjelaskan tentang alur proses kerja algoritma *Multinomial Naive Bayes* dalam membangun dan mengevaluasi model klasifikasi. Proses dimulai dengan *load dataset*, di mana dataset dimuat ke sistem, diikuti oleh tahap data *preprocessing* untuk membersihkan dan mempersiapkan data. Setelah itu, data dibagi menjadi dua bagian dalam langkah *split data training* & *testing*: data pelatihan untuk membangun model dan data pengujian untuk evaluasi. Pada jalur pelatihan, model menghitung *Likelihood* fitur P(F|C) untuk setiap fitur berdasarkan frekuensi kemunculannya dalam setiap kelas, dilanjutkan dengan perhitungan probabilitas kelas P(C), yang merepresentasikan prioritas kelas dalam data. Selanjutnya, perhitungan probabilitas akhir dilakukan menggunakan distribusi multinomial untuk memprediksi kelas berdasarkan fitur yang diberikan. Setelah model selesai dilatih, pada jalur pengujian, model digunakan untuk melakukan prediksi pada data uji dan hasilnya dievaluasi dalam tahap evaluasi menggunakan metrik seperti *accuracy* atau *F1-score*. Akhirnya, proses berakhir setelah evaluasi, menghasilkan model yang dapat digunakan untuk memprediksi data baru dengan probabilitas yang telah dihitung.

3.3.7 *Mixed Distribution* Pada *Naive Bayes*

Mixed Distribution terjadi secara alami ketika variabel acak dengan distribusi kontinu dipotong dengan cara tertentu (*2.1.3 Mixed Distributions*, n.d.). *Mixed Distribution Naive Bayes* digunakan untuk klasifikasi yang memanfaatkan gabungan dari distribusi untuk menghitung probabilitas kelas dengan menggabungkan rumus *Gaussian Naive Bayes* untuk fitur kontinu dengan *Multinomial Naive Bayes* untuk untuk fitur kategorikal. Pendekatan ini tetap

mempertahankan asumsi independensi kondisional dari *Naive Bayes*, sehingga tetap efisien dan sederhana untuk diterapkan.

Persamaan *Mixed Distribution Naive Bayes* sebagai berikut.

$$P(C_j)\left\{\prod_{k=1}^{d} N\left(x_k;\ \hat{\mu}_{jk}, \hat{\sigma}_{jk}\right) \ \times \prod_{d+1}^{m} M\left(x_1, x_2, .., x_m; \theta_1, \theta_2, \ldots . \theta_m\right)\right\} \qquad (3.8)$$

Setelah menggabung rumus *Gaussian Naive Bayes* dengan *Multinomial Naive Bayes* seperti pada rumus 3.8 maka tahap selanjutnya dilakukan proses menghitung probabilitas posterior untuk setiap kelas berdasarkan fitur yang diberikan, kita dapat memprediksi kelas mana yang paling mungkin untuk observasi baru X. Pendekatan ini tetap sederhana dan efisien, mengikuti prinsip dasar dari *Naive Bayes* yang mengasumsikan independensi kondisional antar fitur. Rumus merupakan sebuah fungsi matematika yang digunakan untuk menentukan indeks atau argumen dari elemen terbesar dalam suatu himpunan atau fungsi. Dalam konteks machine learning dan statistik, *Argmax* digunakan untuk menemukan kelas atau nilai yang memiliki probabilitas atau skor tertinggi.

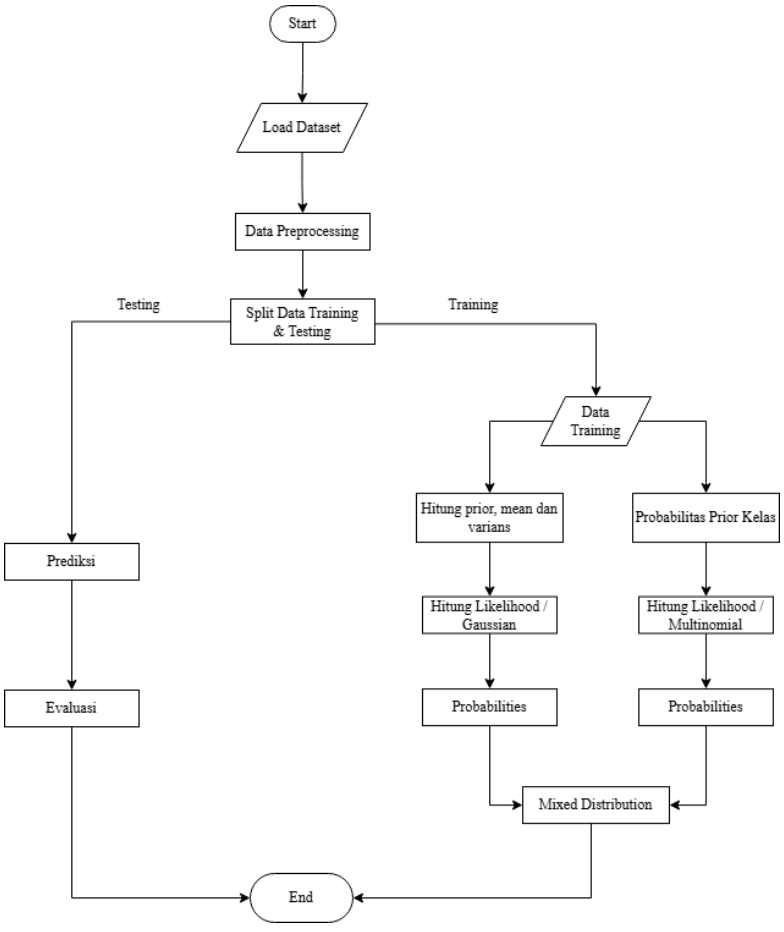
Persamaan *Argmax*:

$$\hat{c} = Argmax\ (\mathrm{P}(C_J)\{\textstyle\prod_{k=1}^{d} N(z_k;\ \hat{\mu}_{jk}, \hat{\sigma}_{jk}) \ \times \prod_{d+1}^{n} M\left(x_1, x_2, .., x_m; \theta_1, \theta_2, .. \theta_n\right)\} \qquad (3.8)$$

Keterangan:

$\hat{c}$	: Prediksi kelas yang memiliki probabilitas tertinggi untuk observasi tertentu *x*.
Argmax	: Operator yang mencari nilai argument yang memaksimalkan ekspresi di dalamnya.
$\mathrm{P}(C_J)$	: Probabilitas prior dari kelas C_J

Pada gambar 3.8 menjelaskan tentang langkah-langkah dalam proses pengembangan model untuk deteksi penyakit menggunakan metode *Naive Bayes* dengan *Mixed Distribution*. Proses dimulai dengan *Load Dataset*, kemudian dilakukan *Data Preprocessing* untuk membersihkan data. Data yang telah diproses dibagi menjadi data pelatihan dan data pengujian. Pada data pelatihan, dihitung nilai *prior*, *mean*, dan *varians*, serta probabilitas prior kelas. Perhitungan *likelihood* dilakukan menggunakan distribusi *Gaussian* untuk data kontinu dan distribusi *multinomial* untuk data kategorikal. Setelah perhitungan probabilitas selesai, model dibangun dan digunakan untuk prediksi pada data pengujian. Terakhir, hasil prediksi dievaluasi untuk menilai performa model sebelum proses berakhir.



Gambar 3.6 Flowchart Mixed Distribution

3.4 Evaluasi

Evaluasi perfoma algoritma pada penelitian ini, merupakan analisis keseluruhan terhadap kinerja dari keseluruhan tahapan-tahapan yang diterapkan. Penilaiaan kemampuan model dalam mengklasifikasikan data untuk mendeteksi penyakit *stroke* dengan *accuracy* yang tinggi. Pada proses evaluasi dibagi menjadi data latihan dan data pengujian Data latihan digunakan untuk melatih model, sementara data pengujian digunakan untuk mengukur performa model yang telah dilatih. Beberapa metrik perfoma yang digunakan pada deteksi penyakit *stroke* meliputi; *accuracy*, *precision*, *recall* (*Sensitivitas), F1-Score*, area di bawah kurva ROC (AUC-ROC). Hasil evaluasi performa model perlu diinterpretasikan untuk menilai apakah model tersebut layak digunakan untuk mendeteksi penyakit stroke dalam sistem. Model yang baik akan memiliki *accuracy* tinggi, *precision*, *recall*, dan AUC-ROC yang mendekati 1. Untuk menetukan nilai evaluasi performa model menggunakan matrix evaluasi dan *confusion matrix*. Berikut beberapa komponen yang digunakan didalam penelitian ini.

TP (*True Positive*): TP mewakili jumlah kasus pasien yang benar-benar menderita *stroke* dan dideteksi secara benar oleh model.

FP (*False Positive*): FP mewakili jumlah kasus pasien yang tidak menderita *stroke* dideteksi secara salah sebagai menderita *stroke* oleh model.

FN (*False Negative*): FN mewakili jumlah kasus pasien yang benar-benar menderita *stroke* tidak terdeteksi sebagai model.

TN (*True Negative*): TN mewakili jumlah kasus pasien yang tidak menderita *stroke* dideteksi dengan benar oleh model.

Tabel 3.3 Representasi *Confusion Matrix*

	Positive	***Negative***
Positive	*True Positive* (TP)	*False Negative* (FN)
Negative	*False Positive* (TP)	*True Negative* (TN)

Accuracy merupakan salah satu matrik yang digunakan untuk mengevaluasi kinerja model atau sistem dalam penelitian, terutama dalam konteks pembelajaran mesin dan statistik. *Accuracy* mengukur seberapa tepat prediksi atau hasil yang dari model yang dibandingkan dengan nilai atau hasil yang sebenarnya. Pada konteks penelitian ini, *accuracy* akan menjadi salah satu metrik kunci untuk menilai seberapa baik model *Naive Bayes* dengan distribusi campuran (*Mixed Distribution*) dalam mendeteksi penyakit *stroke*. Namun, penting untuk mempertimbangkan metrik tambahan untuk mendapatkan gambaran yang lebih komprehensif, terutama dalam kasus data yang tidak seimbang. Menggunakan kombinasi metrik akan membantu memastikan bahwa model tidak hanya akurat secara keseluruhan, tetapi juga efektif dalam mendeteksi pasien yang benar-benar berisiko *stroke*.

$$Accuracy = \frac{TP + TN}{(TP + TN + FP + FN)} \times 100\% \tag{3.9}$$

Precision merupakan salah satu *matrix* evaluasi yang digunakan untuk mengukur kinerja model klasifikasi, terutama dalam konteks pembelajaran data mining. *Precision* mengukur keakuratan dari prediksi positif yang dibuat oleh model. Misalnya dalam diagnostik medis, seperti deteksi penyakit tertentu, *precision* tinggi memastikan bahwa sebagian besar diagnosis positif adalah benar, sehingga mengurangi risiko perawatan yang tidak perlu. Namun, penting untuk

mempertimbangkan *precision* bersama metrik lain seperti *recall* untuk mendapatkan gambaran yang lebih lengkap tentang performa model.

$$Precision = \frac{TP}{(TP + FP)} \times 100\% \quad (3.10)$$

Recall merupakan salah satu metrik evaluasi kinerja model klasifikasi yang juga dikenal sebagai sensitivitas, mengukur proporsi instance positif yang benar-benar teridentifikasi oleh model. *Recall* membantu memastikan bahwa model lebih dapat diandalkan dalam mendeteksi semua kasus positif yang relevan. Semakin tinggi nilai *recall*, semakin baik model dalam mengidentifikasi semua instance positif yang sebenarnya ada.

$$Recall = \frac{TP}{(TP + FN)} \times 100\% \quad (3.11)$$

F1-Score merupakan matrik evaluasi yang mengkombinasikan *precision* dan *recall* untuk memberikan gambaran menyeluruh tentang kinerja model klasifikasi. *F1-Score* digunakan ketika ketidakseimbangan kelas, *precision* dan *recall*, dan dugunakan untuk mengevaluasi seluruh kinerja. Contoh penerapan *F1-Score* dalam deteksi penyakit seperti *stroke*, *F1-Score* membantu mengevaluasi kinerja model dalam mendeteksi penyakit tanpa mengabaikan pentingnya menghindari alarm palsu.

$$F1-Score = 2 \times \frac{(Presisi \times Recall)}{(Presisi + Recall)} \times 100\% \quad (3.12)$$

BAB IV

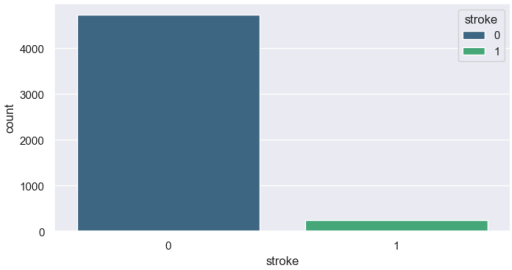
UJI COBA DAN PEMBAHASAN

4.1 Hasil Pengujian Data

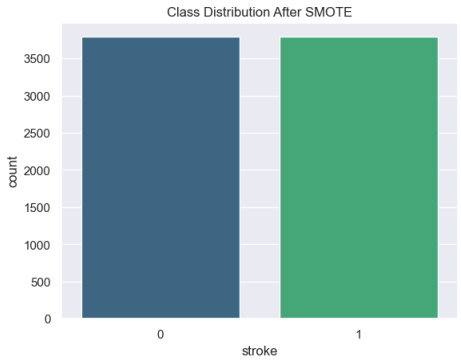
Pada sub bab ini akan menjelaskan terkait dengan hasil teknik SMOTE data. hasil uji *Chi Square*, hasil Normalisasi data untuk variabel kontinu dan hasil uji Skenario.

4.1.1 Hasil SMOTE Data

Pada penelitian ini mempunyai data yang tidak seimbang. Oleh karena itu perlu dilakukan teknik SMOTE data untuk menangani masalah ketidak seimbangan data. Pada gambar 4.1 merupakan perbandingan (*class*) yang belum dilakukan teknik SMOTE data.



Gambar 4.1 Data sebelum dilakukan SMOTE



Gambar 4.2 Data setelah dilakukan SMOTE

Pada gambar 4.2 di atas merupakan hasil penerapan dari teknik SMOTE. Hasilnya adalah data yang seimbang, dengan distribusi antara kelas mayoritas dan minoritas lebih setara. Hal ini sangat penting untuk kasus prediksi *stroke*, di mana kondisi kelas "*stroke*" (minoritas) lebih sedikit dibandingkan kelas "tidak *stroke*" (mayoritas). Jumlah data yang sebelum dilakukan teknik SMOTE berjumlah 4981 data, setelah dilakuan teknik SMOTE menjadi 9466 data.

4.1.2 Hasil *Uji Chi-Square Test* Data *Gaussian*

Uji normalitas *Chi-Square Test* pada data *Gaussian* digunakan pada *Age*, *Avg Glucose Level*, dan BMI dilakukan untuk memeriksa apakah distribusi dari setiap variabel meengikuti distribusi normal.

Tabel 4.1 Uji *Chi-Square Test* pada data *Gaussian*

Variabel	***Chi-square Statistic***	***P-Value***
Age	252,02	0.0
Avg Glucose Level	6435,24	0.0
BMI	2294,63	0.0

Pada tabel 4.1 uji *Chi-Square test* pada data *Gaussian* menunjukan bahwa *p-value* yang sangat kecil dari 0,05 yang menunjukkan bahwa perbedaan antara distribusi yang diamati dan distribusi yang diharapkan (normal) sangat signifikan. Oleh karena itu, kita dapat menyimpulkan bahwa data tersebut tidak berdistribusi normal.

4.1.3 Hasil Uji *Chi-Square Test* Data *Multinomial*

Uji *Chi-Square Test* pada data *Multinomial* untuk variabel kategorikal (*Gender*, *Hypertension*, *Heart Disease*, *Ever Married*, *Work Type*, *Residence Type*,

dan *Smoking Status*), digunakan untuk memeriksa apakah frekuensi setiap kategori dalam variabel-variabel tersebut sesuai dengan distribusi yang diharapkan. Berikut bmerupakan hasil dari uji *Chi-Square test* pada data *Multinomial*.

Tabel 4.2 Uji *Chi-Square Test* pada data *Multinomial*

Variabel	***Chi-square Statistic***	***P-value***
Gender	139,31	0,0
Hypertension	3249,25	0,0
Heart Disease	3941,73	0,0
Ever Married	500,55	0,0
Work Type	2803,53	0,0
Residence Type	1,38	0,24
Smoking Status	625,99	0,0

Berdasarkan tabel 4.2 menunjukan bahwa untuk variabel uji *Gender*, *Hypertension*, *Heart Disease*, *Ever Married*, *Work Type*, *Smoking Status*, untuk p-value adalah 0,0 dalam uji *Chi-Square Goodness of Fit*, itu berarti data sangat tidak sesuai dengan distribusi yang diharapkan, dan kita menolak hipotesis nol bahwa data mengikuti distribusi tersebut. Ini dapat menunjukkan bahwa model distribusi yang diharapkan tidak cocok untuk data yang diuji. Sementara untuk *Residence Type* sesuai dengan uji *Chi-Square Goodness of Fit* karena memiliki nilai *P-value* 0,24 atau lebih dari 0,05.

4.1.4 Hasil Normalisasi Data *Gaussian*

Berdasarkan hasil uji *Chi-Square*, terdapat beberapa fitur yang tidak sesuai dengan distribusi yang diharapkan pada data *Gaussian* maka data tersebut yang berupa variabel *Age*, *Avg Glucose Level*, BMI perluh untuk dilakukan normalisasi atau transformasi data. Untuk hasil pengujian dengan perbandingan data yang belum dilakukan normalisasi dan yang sudah dilakukan normalisasi sebagai berikut.

Tabel 4.3 Sampel data sebelum dan sesudah normalisasi.

Sebelum Dinormalisasi			**Sesudah Dinormalisasi**		
Age	***Avg Glucose Level***	**BMI**	***Age***	***Avg Glucose Level***	**BMI**
67.0	228.69	36.6	0.82	0.80	0.65
80.0	105.92	32.5	0.98	0.23	0.53
49.0	171.23	34.4	0.60	0.54	0.58
79.0	174.12	24.0	0.96	0.55	0.29
81.0	186.21	29.0	0.99	0.61	0.43

Pada tabel 4.3 terdapat hasil nilai *Age*, Avg *Glucose Level*, dan BMI telah diubah menjadi nilai antara 0 hingga 1. Dengan menyelaraskan skala fitur-fitur dalam dataset, normalisasi mencegah dominasi fitur dengan skala besar, mempercepat konvergensi dalam algoritma pembelajaran, meningkatkan *accuracy* model. Secara keseluruhan, normalisasi memastikan bahwa model dapat belajar dari data secara optimal tanpa bias yang disebabkan oleh perbedaan skala antar fitur.

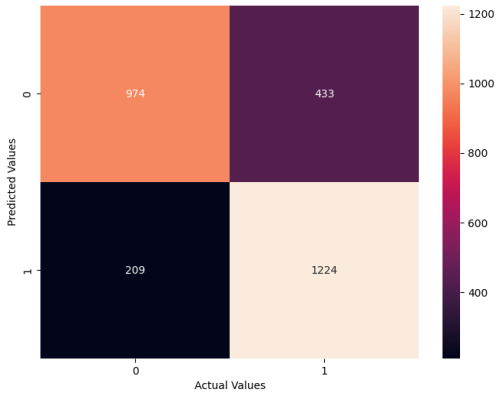
4.2 Hasil Uji Coba

Untuk uji coba dalam penelitian ini dilakukan dengan menerapkan tiga skenario pembagian data. Skenario pertama menggunakan 70% data sebagai data latih dan 30% sebagai data uji. Skenario kedua menggunakan 80% data sebagai data latih dan 20% sebagai data uji. Skenario ketiga menggunakan 90% data sebagai data latih dan 10% sebagai data uji.

4.2.1 Skenario 1 (Data training 70% dan data testing 30%)

Pada gambar 4.3 *confusion matrix* di atas menunjukkan performa model klasifikasi dengan dua kelas, yaitu kelas positif dan kelas negatif. Matriks ini berisi empat elemen utama: *true positives* (TP), *false positives* (FP), *false negatives* (FN),

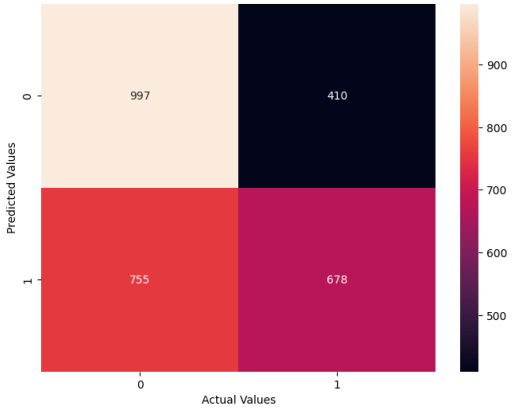
dan *true negatives* (TN). Nilai TP (1224) menunjukkan jumlah prediksi positif yang benar, sedangkan FN (209) menunjukkan jumlah prediksi negatif yang sebenarnya positif. Sebaliknya, FP (433) adalah jumlah prediksi positif yang sebenarnya negatif, dan TN (974) adalah jumlah prediksi negatif yang benar. Matriks ini memberikan gambaran tentang seberapa baik model dalam mengklasifikasikan sampel dengan benar, dan dapat digunakan untuk menghitung berbagai metrik evaluasi, seperti *accuracy*, *precision*, *recall*, dan *F1-score*, untuk memahami kinerja model secara keseluruhan.



Gambar 4.3 *Matrix Gaussian Naive Bayes* skenario 1

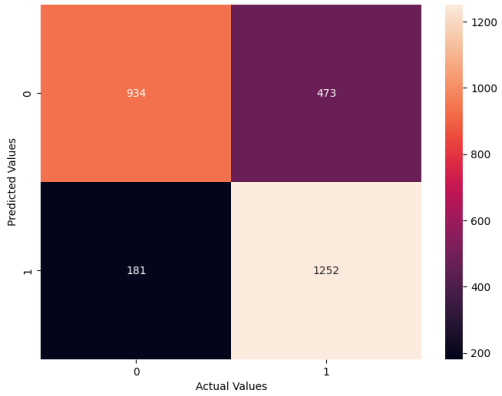
Pada gambar 4.4 merupakan *confusion matrix* tersebut menunjukkan hasil evaluasi performa model klasifikasi. *Matrix* ini menggambarkan jumlah prediksi benar dan salah yang dilakukan oleh model pada dua kelas (kelas positif dan negatif). Setiap baris dan kolom mewakili prediksi dan label sebenarnya dari kelas tertentu. Dalam matriks ini, terdapat 997 prediksi benar untuk kelas negatif (*True Negative*) dan 678 prediksi benar untuk kelas positif (*True Positive*). Sementara itu, terdapat 410 prediksi salah untuk kelas positif yang seharusnya negatif (*False Positive*) dan 755 prediksi salah untuk kelas negatif yang seharusnya positif (*False*

Negative). Dari matriks ini, kita bisa menghitung metrik evaluasi lain seperti *accuracy, precision, recall,* untuk memahami kinerja model secara lebih mendalam.



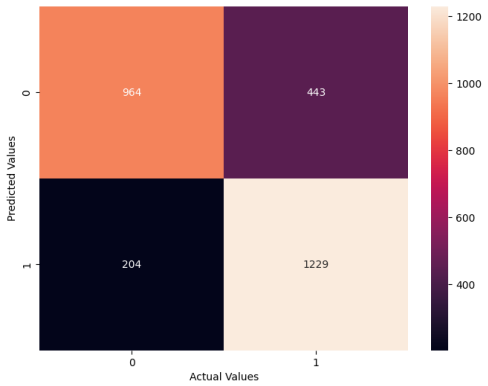
Gambar 4.4 *Matrix Multinomial Naïve Bayes* skenario 1

Pada gambar 4.5 menggambarkan hasil prediksi model klasifikasi dengan dua kelas, positif dan negatif. Untuk nilai 934 menunjukkan jumlah prediksi yang benar untuk kelas negatif (*True Negative*), di mana model dengan tepat memprediksi data negatif. Untuk nilai 473 mewakili kesalahan model dalam memprediksi data negatif sebagai positif (*False Positive*). Untuk nilai 181 menggambarkan kesalahan model dalam memprediksi data positif sebagai negatif (*False Negative*). Sedangkan untuk nilai 1252 menunjukkan jumlah prediksi yang benar untuk kelas positif (*True Positive*), di mana model berhasil mengklasifikasikan data positif dengan benar. Dari *confusion matrix* ini, kita dapat menghitung berbagai metrik evaluasi seperti *accuracy, precision, recall*, dan *F1 score* untuk menilai kinerja model dalam membedakan antara kelas positif dan negatif.



Gambar 4.5 *Matrix Mixed Distribution* 1 skenario 1

Pada gambar 4.6 menjelaskan Matriks konfusi menunjukkan kinerja model klasifikasi dalam mendeteksi *stroke* dengan hasil sebagai berikut: 964 *True Negative* (TN), 1229 *True Positive* (TP), 443 *False Positive* (FP), dan 204 *False Negative* (FN). Berdasarkan ini, *accuracy* model mencapai 77,21%, menunjukkan bahwa sebagian besar prediksi model benar. *Precision* sebesar 73,49% menunjukkan bahwa dari semua prediksi positif, sekitar 73% adalah benar. *Recall* model mencapai 85,76%, mengindikasikan kemampuan yang baik dalam mendeteksi kasus *stroke*. *F1-Score* sebesar 79,17% mencerminkan keseimbangan yang baik antara *precision* dan *recall*.



Gambar 4.6 *Matrix Mixed Distribution* 2 skenario 1

Pada tabel 4.4 menunjukan hasil uji pada skenario 1 menunjukkan kinerja dari tiga model klasifikasi, yaitu *Gaussian Naive Bayes, Multinomial Naive Bayes, Mixed Distribution* dengan menggunakan seluruh variabel *Multinomial*, dan *Mixed Distribution* hanya menggunakan variabel *residence type* yang diukur dengan metrik *accuracy*, *precision, recall*, dan *F1-score*. Pada model *Mixed Distribution* seluruh variabel *Multinomial* memiliki *recall* tertinggi sebesar 87,36%, menunjukkan kemampuannya yang kuat dalam mendeteksi kasus positif *stroke*, dengan *accuracy* 76,97% dan *F1-score* 79,29%. Di sisi lain, model *Mixed Distribution* hanya menggunakan variabel *residence type* menampilkan performa yang lebih seimbang di semua metrik, dengan *accuracy* 77,21%, *precision* 73,50%, *recall* 85,76%, dan *F1-score* 79,16%. Secara keseluruhan, model *Mixed Distribution* dengan seluruh variabel *Multinomial* lebih unggul dalam mendeteksi kasus positif (*recall*), sementara model *Mixed Distribution* hanya menggunakan variabel *residence type* menunjukkan keseimbangan kinerja yang lebih baik di setiap metrik.

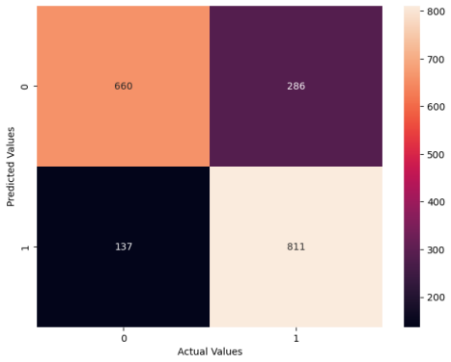
Tabel 4.4 Hasil uji skenario 1

	***Accuracy* (%)**	***Precision* (%)**	***Recall* (%)**	***F1-Score* (%)**
Gaussian Naive Bayes	77,39	73,86	c85,41	79,22
Multinomial Naive Bayes	58,97	62,31	47,31	53,78
***Mixed Distribution* (*All Variabel*)**	76,97	72,57	87,36	79,29
***Mixed Distribution* (*Residence Type*)**	77,21	73,50	85,76	79,16

4.2.2 Skenario 2 (Data *training* 80% dan data *testing* 20%)

Pada gambar *confusion matrix* 4.7 tersebut menggambarkan hasil klasifikasi model dengan dua kelas, yaitu positif dan negatif. Matriks ini menunjukan angka

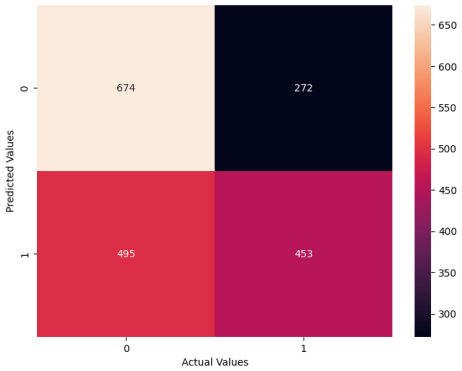
660 menunjukkan jumlah *True Negatives* (TN), yaitu prediksi negatif yang benar. Angka 811 menunjukkan jumlah *True Positives* (TP), yaitu prediksi positif yang benar. Sedangkan angka 286 merupakan *False Positives* (FP), yang artinya model memprediksi positif padahal sebenarnya negatif. Terakhir, angka 137 merupakan *False Negatives* (FN), yaitu ketika model memprediksi negatif padahal sebenarnya positif. *Confusion matrix* ini memberikan informasi tentang seberapa baik model dalam membedakan kedua kelas, serta dapat digunakan untuk menghitung metrik evaluasi seperti *accuracy*, *precision, recall*, dan *F1-score.* Pengujian model data dibagi menjadi 80% untuk pelatihan (*training*) dan 20% untuk pengujian (*testing*). Dalam skenario ini, model dilatih dengan lebih banyak data (80%) dibandingkan Skenario-1, yang seharusnya memberi model lebih banyak informasi untuk belajar pola-pola dalam data dan meningkatkan *accuracy* prediksinya.



Gambar 4.7 *Matrix Gussian Naive Bayes* skenario 2

Pada gambar 4.8 merupakan hasil *confusion matrix* ini menunjukkan evaluasi sebuah model klasifikasi. Matriks berisi empat elemen yang masing-masing menunjukkan jumlah prediksi model dibandingkan dengan nilai aktual. Pada angka 674 menunjukkan *True Positives* (TP), yaitu kasus di mana model

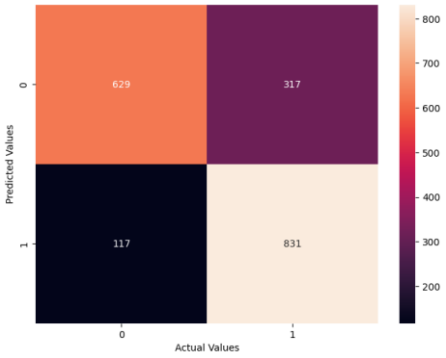
memprediksi positif, dan hasilnya benar. Angka 272 menunjukkan *False Negatives* (FN), yaitu kasus di mana model memprediksi negatif, tetapi hasil sebenarnya positif. Angka 495 adalah *False Positives* (FP), yaitu kasus di mana model memprediksi positif, tetapi hasil sebenarnya negatif. Sedangkan angka 453 adalah *True Negatives (*TN), yaitu kasus di mana model memprediksi negatif dan hasilnya benar. *Confusion matrix* ini membantu kita menilai performa model dengan melihat seberapa baik model memprediksi kelas yang benar, serta berapa banyak kesalahan yang dibuat.



Gambar 4.8 *Matrix Multinomial Naive Bayes* skenario 2

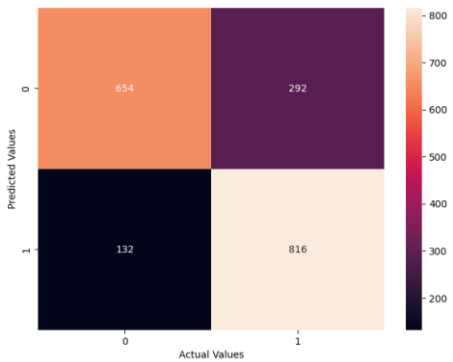
Pada gambar 4.9 menggambarkan hasil prediksi model klasifikasi dengan dua kelas, yaitu kelas positif dan kelas negatif. Angka 629 menunjukkan jumlah prediksi yang benar untuk kelas negatif (*True Negative*), di mana model memprediksi data negatif dengan tepat. Angka 317 menunjukkan kesalahan model dalam memprediksi data negatif sebagai positif (*False Positive*). Angka 117 mengindikasikan kesalahan model dalam memprediksi data positif sebagai negatif (*False Negative*). Sedangkan angka 831 menunjukkan jumlah prediksi yang benar untuk kelas positif (*True Positive*), di mana model berhasil memprediksi data positif dengan tepat. Dari *confusion matrix* ini, kita dapat menghitung metrik evaluasi

seperti *accuracy, precision, recall,* dan *F1 score* untuk menilai seberapa baik model dalam mengklasifikasikan data.



Gambar 4.9 *Matrix Mixed Distribution* 1 skenario 2

Pada gambar 4.10 merupakan *confusion matrix* yaangmenunjukkan hasil prediksi model dengan 654 *True Negative* (TN), 816 True Positive (TP), 292 *False Positive* (FP), dan 132 *False Negative* (FN). Berdasarkan matriks ini, *accuracy*, model mencapai 77,62%, yang berarti sebagian besar prediksi model adalah benar. *precision* sebesar 73,61% menunjukkan bahwa sekitar 74% dari semua prediksi positif adalah benar, sementara *recall* sebesar 86,07% menunjukkan kemampuan model untuk mendeteksi sebagian besar kasus positif. *F1-Score* sebesar 79,31% mencerminkan keseimbangan yang baik antara *precision* dan *recall*.



Gambar 4.10 *Matrix Mixed Distribution* 2 skenario 2

Pada tabel 4.5 menunjukan hasil uji pada skenario 2 menunjukkan kinerja masing-masing model untuk berbagai model *Naive Bayes* dalam deteksi penyakit *stroke* menggunakan metrik *accuracy*, *precision*, *recall*, dan *F1-score*. Model *Gaussian Naive Bayes* menunjukkan performa yang baik dengan *accuracy* sebesar 77,66%, precision 73,92%, *recall* 85,54%, dan *F1-score* 79,31%. Model *Multinomial Naive Bayes* memiliki performa terendah dengan *accuracy* 59,50%, *precision* 62,48%, *recall* 47,78%, dan *F1-score* 54,15%. Model *Mixed Distribution* dengan menggunakan seluruh variabel *Multinomial* memberikan hasil yang cukup baik dengan accuracy 77,08%, precision 72,38%, dan *recall* tertinggi sebesar 87,65%, menghasilkan *F1-score* 79,29%. Pada model *Mixed Distribution* (*Residence Type*) mencatat kinerja keseluruhan terbaik dengan *accuracy* 77,61%, precision 73,64%, *recall* 86,07%, dan *F1-score* tertinggi sebesar 79,37%. Hasil ini menunjukkan bahwa model *Mixed Distribution* dengan fitur *residence type* memiliki performa paling optimal di Skenario 2 untuk deteksi *stroke*.

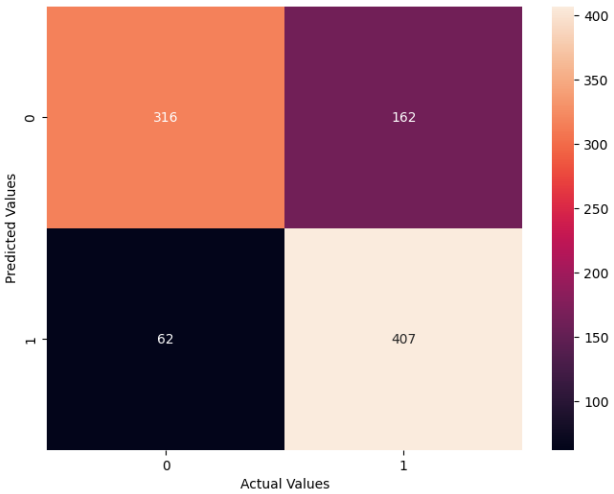
Tabel 4.5 Hasil uji skenario 2

	***Accuracy* (%)**	***Precision* (%)**	***Recall* (%)**	***F1-Score* (%)**
Gaussian Naive Bayes	77,66	73,92	85,54	79,31
Multinomial Naive Bayes	59,50	62,48	47,78	54,15
***Mixed Distribution* (*All*)**	77,08	72,38	87,65	79,29
***Mixed Distribution* (*Residence Type*)**	77,61	73,64	86,07	79,37

4.2.3 Skenario 3 (Data *training* 90% dan data *testing* 10%)

Pengujian Pada gambar 4.11 menunjukkan hasil evaluasi dari model klasifikasi. Dalam matrix ini, terdapat 316 *True Negatives* (TN), yang berarti model dengan benar memprediksi kelas negatif; 407 *True Positives* (TP), yang berarti

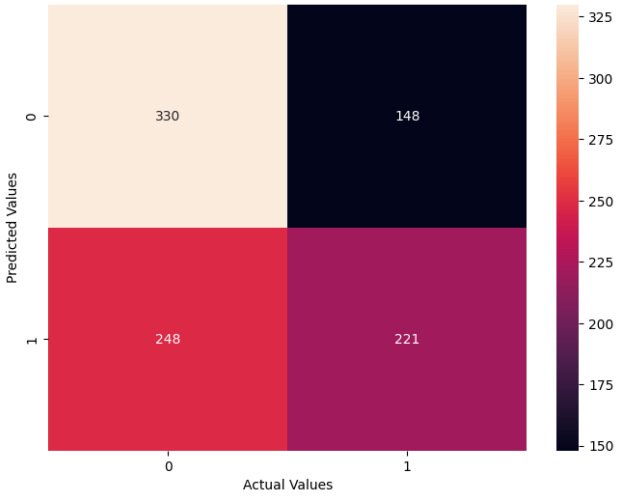
model dengan benar memprediksi kelas positif; 162 *False Positives* (FP), yang berarti model salah memprediksi kelas positif padahal seharusnya negatif; dan 62 *False Negatives* (FN), yang berarti model salah memprediksi kelas negatif padahal seharusnya positif. Berdasarkan nilai-nilai ini, kita dapat menghitung metrik evaluasi seperti accuracy, precision, *recall*, dan *F1-score*. *Accuracy* mengukur seberapa sering model memprediksi dengan benar, sementara precision mengukur seberapa banyak prediksi positif yang benar, dan *recall* mengukur seberapa banyak kelas positif yang berhasil terdeteksi oleh model. *F1-score* adalah rata-rata harmonis dari precision dan *recall*, yang memberikan gambaran umum tentang keseimbangan antara keduanya.



Gambar 4.11 *Matrix Gussian Naive Bayes* skenario 3

Pada gambar 4.12 merupakan *confusion matrix* yang menunjukkan kinerja model klasifikasi biner dengan dua kelas, yaitu kelas positif dan kelas negatif. Matriks ini terdiri dari empat elemen yang menggambarkan hasil prediksi model terhadap data aktual. Elemen pertama, *True Positive* (TP) sebesar 330, menunjukkan jumlah sampel yang diprediksi benar sebagai kelas positif. Elemen

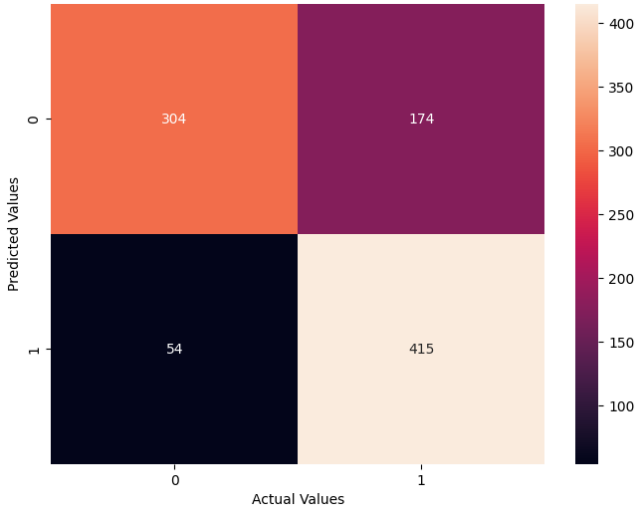
kedua, *False Positive* (FP) sebesar 148, menggambarkan jumlah sampel yang diprediksi salah sebagai kelas positif padahal sebenarnya kelas negatif. Elemen ketiga, *False Negative* (FN) sebesar 248, adalah jumlah sampel yang diprediksi salah sebagai kelas negatif padahal sebenarnya kelas positif. Terakhir, *True Negative* (TN) sebesar 221, menunjukkan jumlah sampel yang diprediksi benar sebagai kelas negatif. Berdasarkan confusion matrix ini, kita dapat menghitung metrik evaluasi seperti *accuracy, precision, recall*, dan *F1-score*, yang membantu dalam menilai kemampuan model untuk memprediksi kelas dengan benar.



Gambar 4.12 *Matrix Multinomial Naive Bayes skenario 3*

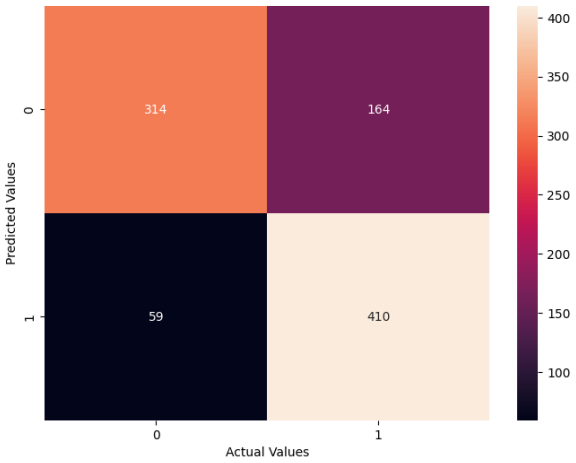
Pada gambar 4.13 baris pertama dan kolom pertama (304) menunjukkan jumlah prediksi benar untuk kelas negatif (*True Negative*). Angka 174 menunjukkan jumlah kesalahan klasifikasi, di mana model memprediksi positif padahal sebenarnya negatif (*False Positive*). Baris kedua dan kolom pertama (54) adalah jumlah kesalahan klasifikasi, di mana model memprediksi negatif padahal sebenarnya positif (*False Negative*). Angka 415 menunjukkan jumlah prediksi benar untuk kelas positif (*True Positive*). Dari *confusion matrix* ini, kita bisa

menghitung beberapa metrik evaluasi lainnya, seperti *accuracy*, *precision*, *recall*, dan *F1-score*, untuk mengukur seberapa baik model dalam memprediksi kedua kelas tersebut. Kemudian juga dapat yang membantu dalam menilai kemampuan model untuk memprediksi kelas dengan benar berdasarkan pada data matrix tersebut.



Gambar 4.13 *Matrix Mixed Distribution* 1 skenario 3

Pada gambar 4.14 merupakan *confusion matrix* yang menunjukkan kinerja model dengan hasil 314 *True Negative* (TN), 410 *True Positive* (TP), 164 *False Positive* (FP), dan 59 *False Negative* (FN). Berdasarkan *matrix* ini, *accuracy* model mencapai 76,46%, yang berarti sebagian besar prediksi model adalah benar. *Precision* sebesar 71,46% menunjukkan bahwa sekitar 71% dari prediksi positif adalah benar-benar positif, sedangkan *recall* sebesar 87,36% mengindikasikan kemampuan model untuk mendeteksi sebagian besar kasus positif. *F1-Score* sebesar 78,64% mencerminkan keseimbangan yang baik antara *precision* dan *recall*.



Gambar 4.14 *Matrix Mixed Distribution* 2 skenario 3

Pada tabel 4.6 menunjukkan hasil uji Skenario 3 untuk berbagai variasi model *Naive Bayes* dalam deteksi penyakit *stroke*, menggunakan metrik *accuracy*, *precision*, *recall*, dan *F1-score*. Model *Gaussian Naive Bayes* memberikan *accuracy* 76,34%, dengan nilai *recall* tertinggi di antara semua model (86,78%) dan *F1-score* sebesar 78,42%. Model *Multinomial Naive Bayes* menunjukkan performa terendah dengan *accuracy* 58,18%, precision 59,89%, dan *recall* 47,12%, yang menghasilkan *F1-score* hanya 52,74%. Model *Mixed Distribution* (*All Variabel Multinomial*) mencapai hasil yang mirip dengan *Gaussian Naive Bayes*, dengan *accuracy* 75,92% dan *recall* tertinggi sebesar 88,48%, namun sedikit lebih rendah pada *F1-score* (78,44%). Sementara itu, model *Mixed Distribution* (*Residence Type*) menghasilkan performa keseluruhan terbaik dengan *accuracy* 76,45%, *precision* 71,42%, *recall* 87,42%, dan *F1-score* tertinggi sebesar 78,61%. Hasil ini menunjukkan bahwa variasi *Mixed Distribution* dengan fitur *tipe residence* memberikan kinerja terbaik dalam mendeteksi *stroke*.

Tabel 4.6 Hasil uji skenario 3

	***Accuracy* (%)**	***Precision* (%)**	***Recall* (%)**	***F1-Score* (%)**
Gaussian Naive Bayes	76,34	71,52	86,78	78,42
Multinomial Naïve Bayes	58,18	59,89	47,12	52,74
Mixed Distribution (All)	75,92	70,45	88,48	78,44
Mixed Distribution (Residence Type)	76,45	71,42	87,42	78,61

4.3 Pembahasan

Pengaruh skenario dalam penelitian *Mixed Distribution* pada *Naive Bayes* untuk deteksi penyakit *stroke* terlihat dari penggunaan berbagai distribusi probabilitas untuk memodelkan data, seperti distribusi *Gaussian* untuk fitur kontinu dan distribusi *Multinomial* untuk fitur kategorikal. Setiap skenario distribusi yang diterapkan memengaruhi hasil deteksi *stroke*, karena distribusi yang berbeda memiliki sensitivitas yang berbeda terhadap jenis data. Dengan menguji beberapa skenario distribusi, peneliti dapat menentukan kombinasi yang paling efektif untuk meningkatkan *accuracy*, *precision*, *recall*, dan *F1-score* dalam memprediksi penyakit *stroke*. Adapun hasil-hasil dari ketiga skenario sebagai berikut.

Tabel 4.7 Rasio hasil seluruh percobaan *Gaussian Naive Bayes*

Skenario	**Rasio Hasil Seluruh Percobaan**			
	***Accuracy* (%)**	***Precision* (%)**	***Recall* (%)**	***F1-score* (%)**
1	77.39	73.86	85.41	79.22
2	77.66	73.92	85.54	79.31
3	76.34	71.52	86.78	78.42

Pada tabel 4.10 menunjukan hasil dari percobaan *Gaussian Naive Bayes* pada tiga rasio pembagian data menunjukkan variasi dalam *accuracy*, *precision*, *recall*, dan *F1-score*. Pada percobaan pertama dengan rasio 70% data pelatihan dan

30% data pengujian, model mencapai accuracy 0.773, precision 0.738, *recall* 0.854, dan *F1-score* 0.792. Pada percobaan kedua, menggunakan 80% data pelatihan dan 20% data pengujian, hasilnya sedikit lebih baik dengan accuracy 0.776, precision 0.739, *recall* 0.855, dan *F1-score* 0.793. Pada percobaan ketiga dengan rasio 90:10, accuracy sedikit menurun menjadi 0.763, precision 0.715, *recall* meningkat menjadi 0.867, dan *F1-score* tercatat 0.784. Hasil ini menunjukkan bahwa precision dan *F1-score* tertinggi tercapai dengan rasio 80:20, sementara *recall* terbaik ditemukan pada rasio 90:10, mengindikasikan bahwa rasio data pelatihan yang lebih besar dapat meningkatkan *recall*, namun sedikit mengorbankan *accuracy* dan *precision*.

Tabel 4.8 Hasil seluruh percobaan *Multinomial*

Skenario	**Rasio Hasil Seluruh Percobaan**			
	***Accuracy* (%)**	***Precision* (%)**	***Recall* (%)**	***F1-score* (%)**
1	58,97	62,31	47,31	53,78
2	59,50	62,48	47,78	54,15
3	58,18	59,89	47,12	52,74

Tabel 4.8 merupakan hasil percobaan *Multinomial Naive Bayes* menunjukkan performa yang relatif stabil namun rendah dibandingkan Gaussian *Naive Bayes*. Pada percobaan pertama dengan rasio 70% data pelatihan dan 30% data pengujian, model menghasilkan *accuracy* 0.589, *precision* 0.623, *recall* 0.473, dan *F1-score* 0.537. Percobaan kedua, dengan rasio 80% data pelatihan dan 20% data pengujian, sedikit meningkatkan *accuracy* menjadi 0.595, *precision* 0.624, *recall* 0.477, dan *F1-score* 0.541. Pada percobaan ketiga dengan rasio 90:10, *accuracy* turun menjadi 0.581, *precision* menjadi 0.598, sedikit menurun ke 0.4712, dan *F1-score* 0.527. Secara keseluruhan, peningkatan proporsi data pelatihan

tampak sedikit menurunkan *accuracy*, *precision*, dan *F1-score*, serta menyebabkan fluktuasi pada nilai *recall*, mengindikasikan bahwa *Multinomial Naive Bayes* mungkin kurang optimal untuk dataset ini dan cenderung menghasilkan performa yang lebih rendah dibandingkan *Gaussian Naive Bayes* dalam klasifikasi data.

Tabel 4.9 Hasil seluruh percobaan *Mixed Distribution* 1

Skenario	**Rasio Hasil Seluruh Percobaan (*All Variabel Multinomial*)**			
	***Accuracy* (%)**	***Precision* (%)**	***Recall* (%)**	***F1-score* (%)**
1	76,97	72,57	87,36	79,29
2	77,08	72,38	87,65	79,29
3	75,92	70,45	88,48	78,44

Hasil percobaan *Mixed Distribution* dengan tiga skenario menunjukkan bahwa model memiliki kinerja yang baik, terutama pada metrik *recall*, meskipun *accuracy* dan *precision* bervariasi. Pada percobaan pertama dengan 6626 data pelatihan dan 2840 data pengujian, model mencapai *accuracy* 76.97%, *precision* 72.57%, *recall* 87.36%, dan *F1-score* 79.29%. Pada percobaan kedua, dengan jumlah data pelatihan yang meningkat menjadi 7572 dan data pengujian sebanyak 1894, *accuracy* sedikit meningkat menjadi 77.08%, dengan *recall* yang juga naik menjadi 87.65%, sementara *precision* sedikit turun ke 72.38%, dan *F1-score* tetap di 79.29%. Percobaan ketiga menggunakan data pelatihan yang lebih besar (8519) dan data pengujian yang lebih kecil (947), menghasilkan *accuracy* yang sedikit lebih rendah, yaitu 75.92%, *precision* turun menjadi 70.45%, namun *recall* meningkat signifikan ke 88.48%, dengan *F1-score* 78.44%.

Pada tabel 4.10 menunjukkan perbandingan kinerja model berdasarkan metrik accuracy, precision, *recall*, dan *F1-score*. Skenario 2 memiliki hasil terbaik dengan sedikit peningkatan pada semua metrik, yaitu accuracy 77,61%, precision

73,64%, *recall* 86,07%, dan *F1-score* 79,37%. Skenario 1 menunjukkan nilai accuracy dan *F1-score* yang tinggi, namun precision dan *recall* sedikit lebih rendah. Skenario 3, meskipun memiliki *recall* tertinggi (87,42%), menunjukkan penurunan pada accuracy dan precision, serta penurunan kecil pada *F1-score*. Secara keseluruhan, skenario 2 memberikan kinerja yang lebih seimbang.

Tabel 4.10 *Mixed distribution* 2

Skenario	**Rasio Hasil Seluruh Percobaan Multinomial *Variabel Residence Type***			
	***Accuracy* (%)**	***Precision* (%)**	***Recall* (%)**	***F1-score* (%)**
1	77,21	73,50	85,76	79,16
2	77,61	73,64	86,07	79,37
3	76,45	71,42	87,42	78,61

Berdasarkan tabel 4.11 penelitian pertama terkait klasifikasi/deteksi penyakit *stroke* menggunakan algoritma *Naive Bayes* dengan menerapkan model *Mixed Distribution* menunjukkan bahwa variabel *Residence type* memiliki pengaruh signifikan dalam memprediksi *stroke*. Ketika model hanya menggunakan variabel *residence type*, *accuracy* mencapai 77,61%, sedikit lebih tinggi dibandingkan dengan model yang menggunakan semua variabel kategori, yang *accuracy* 77,08%. Hal ini menunjukkan bahwa variabel *Residence type* dapat menjadi prediktor kuat dalam model, sehingga berpotensi memberikan hasil prediksi yang optimal meskipun digunakan tanpa variabel lain.

Tabel 4.11 Penelitian dengan data yang sama pada skenario 2

No	**Metode**	***Accuracy* (%)**
1	*Gaussian Naïve Bayes*	77,66
2	*Multinomial Naïve Bayes*	59,50
3	*Mixed Distribution* (*Multinomial All Variabel*)	77,08
4	*Mixed Distribution* (*Multinomial Variabel Residence Type*)	77,61

Pada dasarnya penelitian yang telah dilakukan mengenai *Mixed Distribution* pada *Naive Bayes* untuk deteksi penyakit *stroke* merupakan salah satu wujud dari penerapan nilai-nilai dalam Al-Qur'an yang dijadikan sebagai pedoman, rahmat, dan petunjuk bagi umat manusia dalam segala aspek kehidupan. Ayat ini menyampaikan bahwa Al-Qur'an mengandung penjelasan dan panduan yang komprehensif, yang mencakup nilai-nilai kebaikan, keilmuan, serta rahmat bagi seluruh umat manusia. Dalam konteks penelitian ini, ilmu yang digunakan untuk pendekatan *Mixed Distribution* dan algoritma *Naive Bayes* merupakan salah satu bentuk implementasi dari pengetahuan dan teknologi yang diberikan Allah SWT kepada manusia. Seperti yang terkandung di dalam Surat Shad Ayat 29:

كِتَٰبٌ أَنزَلْنَٰهُ إِلَيْكَ مُبَٰرَكٌ لِّيَدَّبَّرُوٓاْ ءَايَٰتِهِۦ وَلِيَتَذَكَّرَ أُوْلُواْ ٱلْأَلْبَٰبِ

"Ini adalah sebuah kitab yang Kami turunkan kepadamu penuh dengan berkah supaya mereka memperhatikan ayat-ayatnya dan supaya mendapat pelajaran orang-orang yang mempunyai fikiran" (Q.S Shad: 29).

Pada penelitian ini juga berperan sebagai wujud nyata dari ilmu pengetahuan yang diamanahkan Allah SWT untuk kemaslahatan umat manusia. Selain itu, ayat ini mengingatkan bahwa segala bentuk pengetahuan, termasuk teknologi kesehatan, pada dasarnya adalah bagian dari rahmat Allah SWT yang dimaksudkan untuk memelihara kehidupan dan kesejahteraan manusia, serta menunjukkan bagaimana penerapan ilmu pengetahuan yang berlandaskan petunjuk Ilahi dapat memberikan manfaat yang luas bagi masyarakat. Seperti yang terkandung di dalam Surat An-Nahl Ayat 82:

وَنُنَزِّلُ مِنَ ٱلْقُرْءَانِ مَا هُوَ شِفَآءٌ وَرَحْمَةٌ لِّلْمُؤْمِنِينَ ۙ وَلَا يَزِيدُ ٱلظَّٰلِمِينَ إِلَّا خَسَارًا

“Dan Kami turunkan dari Al Quran suatu yang menjadi penawar dan rahmat bagi orang-orang yang beriman dan Al Quran itu tidaklah menambah kepada orang-orang yang zalim selain kerugian.” (QS. Al-Isra’/17; 82).

Tafsir Ringkas Kementerian Agama RI tentang Surat Al-Isra ayat 82: Al-Qur'an yang diturunkan kepada Nabi Muhammad SAW memiliki dua fungsi utama bagi manusia. Bagi orang-orang beriman, Al-Qur'an menjadi obat yang dapat menyembuhkan penyakit hati dan memberikan rahmat ketika mereka mengamalkan ajarannya. Sebaliknya, bagi orang-orang yang zalim, Al-Qur'an justru meningkatkan kerugian mereka karena kekufuran yang semakin bertambah saat mendengar ayat-ayatNya.Sikap manusia terhadap nikmat Allah juga berbeda-beda. Ketika diberi kenikmatan seperti kesehatan atau harta, sebagian manusia justru bersikap sombong dan tidak bersyukur kepada Allah. Namun ketika ditimpa kesulitan seperti sakit atau miskin, mereka mudah putus asa dan kehilangan harapan akan rahmat Allah SWT.Surah Ali Imran Ayat 190:

اِنَّ فِيْ خَلْقِ السَّمٰوٰتِ وَالْاَرْضِ وَاخْتِلَافِ الَّيْلِ وَالنَّهَارِ لَاٰيٰتٍ لِّاُولِى الْاَلْبَابِۚ ١٩٠

"Sesungguhnya dalam penciptaan langit dan bumi serta pergantian malam dan siang terdapat tanda-tanda (kebesaran Allah) bagi orang yang berakal," (QS Ali Imran: 190).

Pada Qur’an Surat Ali Imran ayat 190 menjelaskan tentang ajakan bagi manusia untuk merenungkan tanda-tanda kebesaran Allah yang tampak dalam alam semesta. dan keteraturan pergantian siang dan malam. Penciptaan langit yang luas, penuh dengan benda-benda langit yang beraturan, serta bumi dengan ekosistemnya yang kompleks, adalah bukti nyata kebesaran dan kebijaksanaan Sang Pencipta. Begitu pula, pergantian siang dan malam yang berfungsi untuk mengatur waktu

istirahat dan bekerja adalah tanda lain dari kebijaksanaan tersebut, menunjukkan keseimbangan yang mendukung kehidupan. Tanda-tanda ini akan tampak jelas bagi mereka yang menggunakan akalnya, yang disebut sebagai ulul albab orang-orang yang tidak hanya melihat alam dari segi fisik, tetapi memahami makna mendalam di baliknya. Dengan merenungkan dan meyakini, mereka akan menyadari keesaan Allah SWT dan kekuasaan-Nya, serta semakin tunduk dan bersyukur atas kebesaran-Nya. Seperti pada HR. Abu Dawud dari Abu Darda nomor 3376:

إِنَّ اللَّهَ أَنْزَلَ الدَّاءَ وَالدَّوَاءَ وَجَعَلَ لِكُلِّ دَاءٍ دَوَاءً فَتَدَاوَوْا وَلَا تَدَاوَوْا بِحَرَامٍ

"Sesungguhnya Allah telah menurunkan penyakit dan obat, dan menjadikan bagi setiap penyakit terdapat obatnya, maka berobatlah dan jangan berobat dengan sesuatu yang haram!" (HR. Abu Dawud dari Abu Darda).

Hadis yang menjelaskan bahwa Allah SWT menurunkan penyakit beserta obatnya menekankan bahwa setiap penyakit pasti ada obat yang telah ditetapkan. Ini menunjukkan keadilan dan rahmat Allah SWT kepada hamba-nya, di mana tidak ada penyakit tanpa solusi atau penyembuhan. Islam juga menganjurkan umatnya untuk berusaha mencari kesembuhan melalui pengobatan yang halal dan melarang penggunaan cara-cara yang haram, seperti obat-obatan yang mengandung zat-zat terlarang. Hadis ini mengajarkan bahwa umat Islam tetap harus berusaha dengan pengobatan yang benar, menjaga keseimbangan antara usaha manusia dan tawakal dan berserah diri kepada Allah SWT.

BAB V

KESIMPULAN DAN SARAN

5.1 Kesimpulan

Berdasarkan hasil penelitian yang telah dilakukan mengenai penerapan *Mixed Distribution* pada *Naive Bayes* untuk deteksi penyakit *stroke* menunjukkan hasil yang baik dibandingkan dengan menggunakan metode *Gaussian Naive Bayes* dan *Multinomial Naive Bayes*. *Gaussian Naive Bayes* memberikan performa terbaik dengan *accuracy* 77,66%, *precision* 73,92%, *recall* 86,78%, dan *F1-score* 79,31%, sedangkan *Multinomial Naive Bayes* memiliki performa terendah dengan *accuracy* 59,50%, *precision* 59,89%, *recall* 47,78%, dan *F1-score* 54,15%. Pendekatan *Mixed Distribution* yang menggabungkan data numerik dan kategori menunjukkan hasil yang mendekati *Gaussian Naive Bayes* dengan accuracy 77,08%, precision 72,57%, *recall* 88,48%, dan *F1-score* 79,29%. Selain itu, penggunaan fitur spesifik seperti *Residence Type* dalam *Mixed Distribution* meningkatkan hasil dengan accuracy 77,61%, precision 73,64%, *recall* 87,42%, dan *F1-score* 79,37%. Dengan demikian, pendekatan *Mixed Distribution* pada *Naive Bayes* terbukti efektif untuk meningkatkan kinerja deteksi penyakit stroke melalui integrasi data numerik dan kategori.

5.2 Saran

Pada pengembangan deteksi penyakit *stroke* diperluhkan adanya perbaikan dan pengembangan lebih lanjut agar mencapai hasil yang maksimal. Oleh kerena itu, peneliti memberikan saran sebagai berikut:

1. Penggunaan dataset yang lebih besar, beragam, dan imbang. Hal tersebut dapat mengurangi *overfitting* dan dapat memberikan pemahaman yang lebih baik mengenai kemampuan generalisasi model.
2. Penggunaan teknik *preprocessing* yang lebih komprehensif berupa teknik *Imbalancing*
3. Mempertimbangkan penggunaan Algoritma lain selain *Naive Bayes* seperti *Random Forest*, *Support Vector Machine* (SVM), *Decision Tree* sehingga dapat membandingkan kinerja hasil yang lebih baik yang memungkinkan memiliki nilai *accuracy* yang lebih tinggi.

Penggunaan dataset yang lebih besar, beragam, dan seimbang, serta teknik *preprocessin*g yang lebih komprehensif seperti SMOTE, sangat penting dalam meningkatkan kinerja model prediksi *stroke*. Dataset yang seimbang mengurangi bias dalam pembelajaran model dan meningkatkan kemampuan generalisasi. Selain itu, penggunaan algoritma lain seperti *Random Forest*, *SVM*, dan *Decision Tree* dapat membantu mengeksplorasi berbagai pendekatan dan menemukan metode yang paling efektif untuk mencapai *accuracy* dan stabilitas yang optimal. Dengan demikian, pendekatan yang komprehensif dalam pemilihan data, *preprocessing*, dan eksplorasi model memungkinkan pengembangan sistem prediksi yang lebih akurat dan andal dalam mendeteksi risiko *stroke*.

DAFTAR PUSTAKA

2.1.3 Mixed Distributions. (n.d.). 1–3. https://account.coachingactuaries.com/app/P/#/doc/390/2143/3

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. https://doi.org/10.32736/sisfokom.v9i3.1023

Hikmayanti Handayani, H., Ahmad Baihaqi, K., & Buana Perjuangan Karawang, U. (2023). Implementasi Algoritma Logistic Regression Untuk Klasifikasi Penyakit Stroke. *Syntax: Jurnal Informatika*, *12*(01), 15–23.

Irwan Budi Santoso, Shoffin Nahwa Utama, & Supriyono. (2024). Citra Tekstur Terbaik Untuk Gaussian Naïve Bayes Dengan Interpolasi Nearest Neighbor. *Jurnal Nasional Teknik Elektro Dan Teknologi Informasi*, *13*(1), 68–75. https://doi.org/10.22146/jnteti.v13i1.8730

Kanggeraldo, J., Sari, R. P., & Zul, M. I. (2018). Sistem Pakar Untuk Mendiagnosis Penyakit Stroke Hemoragik dan Iskemik Menggunakan Metode Dempster Shafer. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, *2*(2), 498–505. https://doi.org/10.29207/resti.v2i2.268

Karim, F., Nurcahyo, G. W., & Sumijan, S. (2021). Sistem Pakar dalam Mengidentifikasi Gejala Stroke Menggunakan Metode Naive Bayes. *Jurnal Sistim Informasi Dan Teknologi*, 221–226. https://doi.org/10.37034/jsisfotek.v3i4.69

Lestari, S., Akmaludin, A., & Badrul, M. (2020). Implementasi Klasifikasi Naive Bayes Untuk Prediksi Kelayakan Pemberian Pinjaman Pada Koperasi Anugerah Bintang Cemerlang. *PROSISKO: Jurnal Pengembangan Riset Dan Observasi Sistem Komputer*, *7*(1), 8–16. https://doi.org/10.30656/prosisko.v7i1.2129

Mahar, N. M. A., Vihi Atina, & Nugroho Arif Sudibyo. (2023). Pemodelan Prediksi Kelulusan Mahasiswa Dengan Metode Naïve Bayes Di Uniba. *Jurnal Manajemen Informatika Dan Sistem Informasi*, *6*(2), 148–158. https://doi.org/10.36595/misi.v6i2.875

Mualfah, D., Fadila, W., & Firdaus, R. (2022). Teknik SMOTE untuk Mengatasi Imbalance Data pada Deteksi Penyakit Stroke Menggunakan Algoritma Random Forest. *Jurnal CoSciTech (Computer Science and Information Technology)*, *3*(2), 107–113. https://doi.org/10.37859/coscitech.v3i2.3912

Praja Utama, T., Said Haibuan, M., Ilmu Komputer, F., Darmajaya, I., Pagar Alam No, J. Z., Meneng, G., Rajabasa, K., & Bandar Lampung, K. (2023). *PENERAPAN ALGORITMA NAÏVE BAYES DAN FORWARD SELECTION UNTUK PREDIKSI PENYAKIT STROKE* (Vol. 17, Issue 2). https://ejurnal.teknokrat.ac.id/index.php/teknoinfo/index

Santoso, I. B. (2013). *STATISTIKA 1 UNTUK TEKNIK INFORMATIKA*. UIN-MALIKI PRESS (Anggota IKAPI).

Sinaga, S. H., Duha, A. A. M., & Banjarnahor, J. (2023). Analisis Prediksi Deteksi Stroke Dengan Pendekatan Eda Dan Perbandingan Algoritma Machine Learning. *Jurnal Ilmiah Betrik*, *14*(02 AGUSTUS), 355–367. https://ejournal.pppmitpa.or.id/index.php/betrik/article/view/120%0Ahttps://ejournal.pppmitpa.or.id/index.php/betrik/article/download/120/75

Suhaili, A., Tinggi, S., Al, I., Wali, Q., Situbondo, S., Hasan, M., & Azhari, R. (n.d.). *KAJIAN AYAT SYIFA' DALAM AL-QUR'AN DALAM TAFSIR AL-THABARI*. http://ejournal.idia.ac.id/index.php/el-waroqoh

Susilawati, Wibowo, A. H., & Sugiarto, A. (2021). *Sistem Pakar Deteksi Dini Penyakit Stroke Menggunakan Metode Dempster Shafer*. *10*(1), 1–10.

Tomasouw, B. P., & Rumlawang, F. Y. (2023). Penerapan Metode SVM Untuk Deteksi Dini Penyakit Stroke (Studi Kasus : RSUD Dr. H. Ishak Umarella Maluku Tengah dan RS Sumber Hidup-GPM). *Tensor: Pure and Applied Mathematics Journal*, *4*(1), 37–44. https://doi.org/10.30598/tensorvol4iss1pp37-44

LAMPIRAN

Lampiran I (Perhitungan Uji *Chi-Square Test*) dengan 40 Data Uji Chi-Square Test (BMI)

BMI
36,6
32,5
34,4
24
29
27,4
22,8
24,2
29,7
36,8
27,3
28,2
30,9
37,5
25,8
37,8
22,4
48,9
26,6
32,5
27,2
23,5
28,2
28,3
44,2
25,4
22,2
30,5
29,7
26,5
33,7
23,1
32
29,9
23,9
28,5
26,4
20,2
33,6
38,6

n	40	
max	48,9	
min	20,2	
range	28,7	
K (kelas)	6,286797971	6
P (panjang kelas)	4,565122043	5

Kelas	Interval		Interval Kelas	Frekuensi
1	20,2	24,2	20,2 - 24,2	9
2	25,2	29,2	25,2 - 29,2	13
3	30,2	34,2	30,2 - 34,2	10
4	35,2	39,2	35,2 - 39,2	6
5	40,2	44,2	40,2 - 44,2	1
6	45,2	49,2	45,2 - 49,2	1
Total				40

x		fi atau (Oi)	xi	fi.xi	xi-xbar	(xi-xbar)^2	fi.(xi-xbar)^2
20,2	24,2	9	22,2	199,8	-7,5	56,25	506,25
25,2	29,2	13	27,2	353,6	-2,5	6,25	81,25
30,2	34,2	10	32,2	322	2,5	6,25	62,5
35,2	39,2	6	37,2	223,2	7,5	56,25	337,5
40,2	44,2	1	42,2	42,2	12,5	156,25	156,25
45,2	49,2	1	47,2	47,2	17,5	306,25	306,25
		40		1188			1450

rata-rata (xbar)	(Σfi.xi)/(Σfi)	29,7	
standar deviasi	Σfi.(xi-xbar)^2/n	6,020797289	6,020797289

Nilai Observasi			Batas Kelas		Z	
Nilai Praktek		fi atau (Oi)	Bawah	Atas	Bawah	Atas
20,2	24,2	9	19,7	24,7	-1,660909597	-0,8304548
25,2	29,2	13	24,7	29,7	-0,830454799	0
30,2	34,2	10	29,7	34,7	0	0,8304548
35,2	39,2	6	34,7	39,7	0,830454799	1,6609096
40,2	44,2	1	39,7	44,7	1,660909597	2,4913644
45,2	49,2	1	44,7	49,7	2,491364396	3,32181919
		40				

Tabel Z		Pi	Ei	(Oi-Ei)^2/Ei
Bawah	**Atas**	**Proporsi**	**Nilai Harapan**	
0,048365802	0,203140847	0,154775045	6,191001792	1,274506323
0,203140847	0,5	0,296859153	11,87436612	0,106704781
0,5	0,796859153	0,296859153	11,87436612	0,295868285
0,796859153	0,951634198	0,154775045	6,191001792	0,005892695
0,951634198	0,993637323	0,042003126	1,680125024	0,275318826
0,993637323	0,999552837	0,005915514	0,236620541	2,46279633
Total				4,421087241

Rumus uji-*Chi-Square*

$x^2 = \sum \frac{(O_i - E_i)^2}{E_i}$	4,421087241

DK (Derajat Kebebasan)

K-3	3
α = 0,05	
Nilai Tabel X^2	7,814727903

Uji Hipotesis:

Dengan menggunakan rumus uji chi-square (x^2) dengan hasil 4,4218. Sementara nilai kritis tabel Uji-Square.adalah 7,8147

Kesimpulan : Data BMI indeks masa tubuh termasuk data distribusi normal.

Signifikansi Uji:

- Jika nilai x^2 dihitung < nilai x^2 tabel, maka Ho diterima, Ha ditolak
- Jika nilai x^2 dihitung > nilai x^2 tabel, maka Ho ditolak, Ha diterima

Lampiran II (Perhitungan *Gaussian Naive Bayes* (*Stroke*))

Age	Avg Glucose Level	BMI
67	228,69	36,6
80	105,92	32,5
49	171,23	34,4
79	174,12	24
81	186,21	29
74	70,09	27,4
69	94,39	22,8
78	58,57	24,2
81	80,43	29,7
61	120,46	36,8
54	104,51	27,3
79	214,09	28,2
50	167,41	30,9
64	191,61	37,5
75	221,29	25,8
60	89,22	37,8
71	193,94	22,4
52	233,29	48,9
79	228,7	26,6
82	208,3	32,5
71	102,87	27,2
80	104,12	23,5
65	100,98	28,2
69	195,23	28,3
57	212,08	44,2
42	83,41	25,4
82	196,92	22,2
80	252,72	30,5
48	84,2	29,7
82	84,03	26,5
74	219,72	33,7
72	74,63	23,1
58	92,62	32
49	60,91	29,9
78	78,03	23,9
54	71,22	28,5
82	144,9	26,4
60	213,03	20,2
76	243,58	33,6
58	107,26	38,6

40 Data (*Age*) Penderita *Stroke*

Age	$(x_1 - \bar{x})$	$(x_1 - \bar{x})^2$
67	-1,05	1,1025
80	11,95	142,8025
49	-19,05	362,9025
79	10,95	119,9025
81	12,95	167,7025
74	5,95	35,4025
69	0,95	0,9025
78	9,95	99,0025
81	12,95	167,7025
61	-7,05	49,7025
54	-14,05	197,4025
79	10,95	119,9025
50	-18,05	325,8025
64	-4,05	16,4025
75	6,95	48,3025
60	-8,05	64,8025
71	2,95	8,7025
52	-16,05	257,6025
79	10,95	119,9025
82	13,95	194,6025
71	2,95	8,7025
80	11,95	142,8025
65	-3,05	9,3025
69	0,95	0,9025
57	-11,05	122,1025
42	-26,05	678,6025
82	13,95	194,6025
80	11,95	142,8025
48	-20,05	402,0025
82	13,95	194,6025
74	5,95	35,4025
72	3,95	15,6025
58	-10,05	101,0025
49	-19,05	362,9025
78	9,95	99,0025
54	-14,05	197,4025
82	13,95	194,6025
60	-8,05	64,8025
76	7,95	63,2025
58	-10,05	101,0025

(Age)

Rumus Gaussian Naive Bayes :

$$P(\widehat{\vartheta}) = \frac{1}{\sqrt{2\pi\hat{\sigma}_j^2}} exp^{(-\frac{(\vartheta-\hat{\mu}_k)^2}{2\hat{\sigma}_k^2})}$$

Rumus simpangan baku (*standard deviation*)

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_1-\bar{x})^2}{n-1}}$$

n	40
Mean (□□	68,05
Deviasi Standar (σ)	12,01697517

$\sum_{i=1}^{n}(x_1-\bar{x})^2$	$\frac{\sum_{i=1}^{n}(x_1-\bar{x})^2}{n-1}$	$\sqrt{\frac{\sum_{i=1}^{n}(x_1-\bar{x})^2}{n-1}}$
5631,9	144,4076923	12,01697517

Setelah mencari simpangan baku (*standard deviation*) maka

diketahui beberapa data sebagai berikut:

$\boldsymbol{\pi}$	3,14
$\hat{\sigma}_j^2$	12,01697517
$\boldsymbol{v}$	67
$\boldsymbol{\mu}$	68,05

Penerapan rumus Gaussian:

$$P(\widehat{\vartheta}) = \frac{1}{\sqrt{2\pi\hat{\sigma}_j^2}} exp^{(-\frac{(\vartheta-\hat{\mu}_k)^2}{2\hat{\sigma}_k^2})}$$

$$P(67) = \frac{1}{\sqrt{(2\times\pi\times12{,}016975)}} exp^{(-\frac{(67-68{,}05)^2}{2\times12{,}016975})}$$

$2\pi\hat{\sigma}_j^2$	$\sqrt{2\pi\hat{\sigma}_j^2}$	$\frac{1}{\sqrt{2\pi\hat{\sigma}_j^2}}$	$(\vartheta-\hat{\mu}_k)^2$	$2\hat{\sigma}_k^2$	$(-\frac{(\vartheta-\hat{\mu}_k)^2}{2\hat{\sigma}_k^2})$
75,46660409	8,687151667	0,115112529	1,1025	24,03395035	-0,045872609

$P(\widehat{\vartheta}) = \frac{1}{\sqrt{2\pi\hat{\sigma}_j^2}} exp^{(-\frac{(\vartheta-\hat{\mu}_k)^2}{2\hat{\sigma}_k^2})}$
0,109951302

40 Data Avg Glucose Lavel Penderita *Stroke*

Avg Glucose Lavel	$x_1 - \bar{x}$	$(x_1 - \bar{x})^2$
228,69	82,06675	6734,951456
105,92	-40,70325	1656,754561
171,23	24,60675	605,4921456
174,12	27,49675	756,0712606
186,21	39,58675	1567,110776
70,09	-76,53325	5857,338356
94,39	-52,23325	2728,312406
58,57	-88,05325	7753,374836
80,43	-66,19325	4381,546346
120,46	-26,16325	684,5156506
104,51	-42,11325	1773,525826
214,09	67,46675	4551,762356
167,41	20,78675	432,0889756
191,61	44,98675	2023,807676
221,29	74,66675	5575,123556
89,22	-57,40325	3295,133111
193,94	47,31675	2238,874831
233,29	86,66675	7511,125556
228,7	82,07675	6736,592891
208,3	61,67675	3804,021491
102,87	-43,75325	1914,346886
104,12	-42,50325	1806,526261
100,98	-45,64325	2083,306271
195,23	48,60675	2362,616146
212,08	65,45675	4284,586121
83,41	-63,21325	3995,914976
196,92	50,29675	2529,763061
252,72	106,09675	11256,52036
84,2	-62,42325	3896,662141
84,03	-62,59325	3917,914946
219,72	73,09675	5343,134861
74,63	-71,99325	5183,028046
92,62	-54,00325	2916,351011
60,91	-85,71325	7346,761226

(Avg Glucose Level)

$$P(\widehat{\vartheta}) = \frac{1}{\sqrt{2\pi\hat{\sigma}_j^2}} exp^{(-\frac{(\vartheta-\hat{\mu}_k)^2}{2\hat{\sigma}_k^2})}$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_1-\bar{x})^2}{n-1}}$$

n	40
Mean (μ)	146,62325
Deviasi Standar (σ)	63,09507341

$\sum_{i=1}^{n}(x_1-\bar{x})^2$	$\frac{\sum_{i=1}^{n}(x_1-\bar{x})^2}{n-1}$	$\sqrt{\frac{\sum_{i=1}^{n}(x_1-\bar{x})^2}{n-1}}$
155258,5433	3980,988289	63,09507341

Setelah mencari simpangan baku **(*standard deviation*)** maka diketahui beberapa data sebagai berikut:

π	3,14
$\hat{\sigma}_j^2$	63,09507341
v	228,69
μ	146,62325

Penerapan rumus Gaussian:

$$P(\widehat{\vartheta}) = \frac{1}{\sqrt{2\pi\hat{\sigma}_j^2}} exp^{(-\frac{(\vartheta-\hat{\mu}_k)^2}{2\hat{\sigma}_k^2})}$$

$$P(228{,}69) = \frac{1}{\sqrt{(2\times\pi\times 63{,}095)}} exp^{(-\frac{(228{,}69-146{,}623)^2}{2\times 63{,}095})}$$

78,03	-68,59325	4705,033946
71,22	-75,40325	5685,650111
144,9	-1,72325	2,969590562
213,03	66,40675	4409,856446
243,58	96,95675	9400,611371
107,26	-39,36325	1549,465451

$2\pi\hat{\sigma}_j^2$	$\sqrt{2\pi\hat{\sigma}_j^2}$	$\frac{1}{\sqrt{2\pi\hat{\sigma}_j^2}}$	$(\vartheta - \hat{\mu}_k)^2$	$2\hat{\sigma}_k^2$	$(-\frac{(\vartheta - \hat{\mu}_k)^2}{2\hat{\sigma}_k^2})$
270,0469142	16,43310422	0,060852775	6734,951456	126,1901468	-53,37145272

$P(\widehat{\vartheta}) = \frac{1}{\sqrt{2\pi\hat{\sigma}_j^2}} exp^{(-\frac{(\vartheta-\hat{\mu}_k)^2}{2\hat{\sigma}_k^2})}$
4,03044E-25

40 Data BMI Penderita *Stroke*

BMI	$x_1 - \bar{x}$	$(x_1 - \bar{x})^2$
36,6	6,8275	46,61475625
32,5	2,7275	7,43925625
34,4	4,6275	21,41375625
24	-5,7725	33,32175625
29	-0,7725	0,59675625
27,4	-2,3725	5,62875625
22,8	-6,9725	48,61575625
24,2	-5,5725	31,05275625
29,7	-0,0725	0,00525625
36,8	7,0275	49,38575625
27,3	-2,4725	6,11325625
28,2	-1,5725	2,47275625
30,9	1,1275	1,27125625
37,5	7,7275	59,71425625
25,8	-3,9725	15,78075625
37,8	8,0275	64,44075625
22,4	-7,3725	54,35375625
48,9	19,1275	365,8612563
26,6	-3,1725	10,06475625

BMI

$$P(\widehat{\vartheta}) = \frac{1}{\sqrt{2\pi\hat{\sigma}_j^2}} exp^{(-\frac{(\vartheta-\hat{\mu}_k)^2}{2\hat{\sigma}_k^2})}$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_1-\bar{x})^2}{n-1}}$$

n	40
Mean (μ)	29,7725
Deviasi Standar (σ)	6,149129082

$\sum_{i=1}^{n}(x_1 - \bar{x})^2$	$\frac{\sum_{i=1}^{n}(x_1 - \bar{x})^2}{n-1}$	$\sqrt{\frac{\sum_{i=1}^{n}(x_1 - \bar{x})^2}{n-1}}$
1474,65975	37,81178846	6,149129082

32,5	2,7275	7,43925625
27,2	-2,5725	6,61775625
23,5	-6,2725	39,34425625
28,2	-1,5725	2,47275625
28,3	-1,4725	2,16825625
44,2	14,4275	208,1527563
25,4	-4,3725	19,11875625
22,2	-7,5725	57,34275625
30,5	0,7275	0,52925625
29,7	-0,0725	0,00525625
26,5	-3,2725	10,70925625
33,7	3,9275	15,42525625
23,1	-6,6725	44,52225625
32	2,2275	4,96175625
29,9	0,1275	0,01625625
23,9	-5,8725	34,48625625
28,5	-1,2725	1,61925625
26,4	-3,3725	11,37375625
20,2	-9,5725	91,63275625
33,6	3,8275	14,64975625
38,6	8,8275	77,92475625

$\boldsymbol{\pi}$	3,14
$\hat{\sigma}_j^2$	6,149129082
$\boldsymbol{v}$	36,6
$\boldsymbol{\mu}$	29,7725

Penerapan rumus Gaussian:

$$P(\widehat{\vartheta}) = \frac{1}{\sqrt{2\pi\hat{\sigma}_j^2}} exp^{(-\frac{(\vartheta-\hat{\mu}_k)^2}{2\hat{\sigma}_k^2})}$$

$$P(18) = \frac{1}{\sqrt{(2\times\pi\times 6{,}716729)}} exp^{(-\frac{(18-27{,}0875)^2}{2\times 6{,}716729})}$$

$2\pi\hat{\sigma}_j^2$	$\sqrt{2\pi\hat{\sigma}_j^2}$	$\frac{1}{\sqrt{2\pi\hat{\sigma}_j^2}}$	$(\vartheta-\hat{\mu}_k)^2$	$2\hat{\sigma}_k^2$	$(-\frac{(\vartheta-\hat{\mu}_k)^2}{2\hat{\sigma}_k^2})$
38,61653063	6,214220034	0,160921241	46,61475625	12,29825816	-3,790354344

Lampiran III Perhitungan *Gaussian* Naive Bayes (Tidak *Stroke*)

40 Data Age Tidak *Stroke*

Age	$x_1-\bar{x}$	$(x_1-\bar{x})^2$
3	-39,85	1588,0225
58	15,15	229,5225
8	-34,85	1214,5225
70	27,15	737,1225
52	9,15	83,7225
75	32,15	1033,6225
32	-10,85	117,7225
79	36,15	1306,8225
79	36,15	1306,8225
37	-5,85	34,2225
37	-5,85	34,2225
40	-2,85	8,1225

n	40
Mean (μ)	42,85
Deviasi Standar (σ)	23,10128202

$\sum_{i=1}^{n}(x_1-\bar{x})^2$	$\frac{\sum_{i=1}^{n}(x_1-\bar{x})^2}{n-1}$	$\sqrt{\frac{\sum_{i=1}^{n}(x_1-\bar{x})^2}{n-1}}$
20813,1	533,6692308	23,10128202

35	-7,85	61,6225
20	-22,85	522,1225
42	-0,85	0,7225
44	1,15	1,3225
79	36,15	1306,8225
65	22,15	490,6225
49	6,15	37,8225
71	28,15	792,4225
59	16,15	260,8225
25	-17,85	318,6225
67	24,15	583,2225
38	-4,85	23,5225
54	11,15	124,3225
70	27,15	737,1225
27	-15,85	251,2225
47	4,15	17,2225
58	15,15	229,5225
3	-39,85	1588,0225
58	15,15	229,5225
14	-28,85	832,3225
32	-10,85	117,7225
23	-19,85	394,0225
55	12,15	147,6225
17	-25,85	668,2225
59	16,15	260,8225
13	-29,85	891,0225
4	-38,85	1509,3225
16	-26,85	720,9225

$\boldsymbol{\pi}$	3,14
$\hat{\sigma}_j^2$	23,10128202
$\boldsymbol{v}$	3
$\boldsymbol{\mu}$	42,85

$$P(\widehat{\vartheta}) = \frac{1}{\sqrt{2\pi\hat{\sigma}_j^2}} exp^{(-\frac{(\vartheta-\hat{\mu}_k)^2}{2\hat{\sigma}_k^2})}$$

$$P(3) = \frac{1}{\sqrt{(2\times\pi\times 23{,}10128)}} exp^{(-\frac{(3-42{,}85)^2}{2\times 23{,}10128})}$$

$2\pi\hat{\sigma}_j^2$	$\sqrt{2\pi\hat{\sigma}_j^2}$	$\frac{1}{\sqrt{2\pi\hat{\sigma}_j^2}}$	$(\vartheta-\hat{\mu}_k)^2$	$2\hat{\sigma}_k^2$	$(-\frac{(\vartheta-\hat{\mu}_k)^2}{2\hat{\sigma}_k^2})$
145,0760511	12,04475201	0,08302371	1588,0225	46,20256403	-34,37087385

$P(\widehat{\vartheta}) = \frac{1}{\sqrt{2\pi\hat{\sigma}_j^2}} exp^{(-\frac{(\vartheta-\hat{\mu}_k)^2}{2\hat{\sigma}_k^2})}$
9,82022E-17

40 Data (Avg Glucose Level) Tidak *Stroke*

Avg Glucose Lavel	$x_1 - \bar{x}$	$(x_1 - \bar{x})^2$
95,12	-4,57925	20,96953056
87,96	-11,73925	137,8099906
110,89	11,19075	125,2328856
69,04	-30,65925	939,9896106
77,59	-22,10925	488,8189356
243,53	143,83075	20687,28465
77,67	-22,02925	485,2878556
77,08	-22,61925	511,6304706
57,08	-42,61925	1816,400471
162,96	63,26075	4001,922491
73,5	-26,19925	686,4007006
95,04	-4,65925	21,70861056
85,37	-14,32925	205,3274056
84,62	-15,07925	227,3837806
82,67	-17,02925	289,9953556
57,33	-42,36925	1795,153346
67,84	-31,85925	1015,011811
75,7	-23,99925	575,9640006
60,22	-39,47925	1558,611181
198,21	98,51075	9704,367866
109,82	10,12075	102,4295806
60,84	-38,85925	1510,041311
94,61	-5,08925	25,90046556
97,49	-2,20925	4,880785563
206,72	107,02075	11453,44093
214,45	114,75075	13167,73463
82,9	-16,79925	282,2148006
103,26	3,56075	12,67894056
55,78	-43,91925	1928,900521
73,74	-25,95925	673,8826606
149,75	50,05075	2505,077576
82,34	-17,35925	301,3435606
62,6	-37,09925	1376,354351
94,09	-5,60925	31,46368556
55,42	-44,27925	1960,651981
82,18	-17,51925	306,9241206
117,92	18,22075	331,9957306
114,84	15,14075	229,2423106
79,17	-20,52925	421,4501056
110,63	10,93075	119,4812956

n	40
Mean (μ)	99,69925
Deviasi Standar (σ)	45,86528777

$\sum_{i=1}^{n}(x_1 - \bar{x})^2$	$\frac{\sum_{i=1}^{n}(x_1 - \bar{x})^2}{n-1}$	$\sqrt{\frac{\sum_{i=1}^{n}(x_1 - \bar{x})^2}{n-1}}$
82041,36028	2103,624623	45,86528777

$\boldsymbol{\pi}$	3,14
$\hat{\sigma}_j^2$	45,86528777
$\boldsymbol{v}$	95,12
$\boldsymbol{\mu}$	99,69925

$$P(\widehat{\vartheta}) = \frac{1}{\sqrt{2\pi\hat{\sigma}_j^2}} exp^{(-\frac{(\vartheta-\hat{\mu}_k)^2}{2\hat{\sigma}_k^2})}$$

$$P(95{,}12) = \frac{1}{\sqrt{(2\times\pi\times 45{,}8652)}} exp^{(-\frac{(95{,}12-99{,}69925)^2}{2\times 45{,}8652})}$$

$2\pi\hat{\sigma}_j^2$	$\sqrt{2\pi\hat{\sigma}_j^2}$	$\frac{1}{\sqrt{2\pi\hat{\sigma}_j^2}}$	$(\vartheta - \hat{\mu}_k)^2$	$2\hat{\sigma}_k^2$	$(-\frac{(\vartheta - \hat{\mu}_k)^2}{2\hat{\sigma}_k^2})$
196,3034317	14,01083265	0,071373346	20,96953056	91,73057555	-0,228599139

$P(\widehat{\vartheta}) = \frac{1}{\sqrt{2\pi\hat{\sigma}_j^2}} exp^{(-\frac{(\vartheta-\hat{\mu}_k)^2}{2\hat{\sigma}_k^2})}$
0,056788018

40 Data (BMI) Tidak *Stroke*

BMI	$x_1 - \bar{x}$	$(x_1 - \bar{x})^2$
18	-9,0875	82,58265625
39,2	12,1125	146,7126563
17,6	-9,4875	90,01265625
35,9	8,8125	77,66015625
17,7	-9,3875	88,12515625
27	-0,0875	0,00765625
32,3	5,2125	27,17015625
35	7,9125	62,60765625
22	-5,0875	25,88265625
39,4	12,3125	151,5976563
26,1	-0,9875	0,97515625
42,4	15,3125	234,4726563
33	5,9125	34,95765625
19,7	-7,3875	54,57515625
22,5	-4,5875	21,04515625
24,6	-2,4875	6,18765625
25,2	-1,8875	3,56265625
41,8	14,7125	216,4576563
31,5	4,4125	19,47015625
27,3	0,2125	0,04515625
23,7	-3,3875	11,47515625
24,5	-2,5875	6,69515625
28,4	1,3125	1,72265625
26,9	-0,1875	0,03515625
26,7	-0,3875	0,15015625
31,2	4,1125	16,91265625
25	-2,0875	4,35765625
25,4	-1,6875	2,84765625
27,5	0,4125	0,17015625
16	-11,0875	122,9326563

n	40
Mean μ	27,0875
Deviasi Standar (σ)	6,716729496

$\sum_{i=1}^{n}(x_1 - \bar{x})^2$	$\frac{\sum_{i=1}^{n}(x_1 - \bar{x})^2}{n-1}$	$\sqrt{\frac{\sum_{i=1}^{n}(x_1 - \bar{x})^2}{n-1}}$
1759,46375	45,11445513	6,716729496

π	3,14
$\hat{\sigma}_j^2$	6,716729496
v	18
μ	27,0875

$$P(\widehat{\vartheta}) = \frac{1}{\sqrt{2\pi\hat{\sigma}_j^2}} exp^{(-\frac{(\vartheta-\hat{\mu}_k)^2}{2\hat{\sigma}_k^2})}$$

$$P(18) = \frac{1}{\sqrt{(2\times\pi\times 6{,}716729)}} exp^{(-\frac{(18-27{,}0875)^2}{2\times 6{,}716729})}$$

27	-0,0875	0,00765625
31,6	4,5125	20,36265625
25,1	-1,9875	3,95015625
30,9	3,8125	14,53515625
24,8	-2,2875	5,23265625
23,4	-3,6875	13,59765625
29,4	2,3125	5,34765625
18,3	-8,7875	77,22015625
20	-7,0875	50,23265625
19,5	-7,5875	57,57015625

$2\pi\hat{\sigma}_j^2$	$\sqrt{2\pi\hat{\sigma}_j^2}$	$\frac{1}{\sqrt{2\pi\hat{\sigma}_j^2}}$	$(\vartheta - \hat{\mu}_k)^2$	$2\hat{\sigma}_k^2$	$(-\frac{(\vartheta - \hat{\mu}_k)^2}{2\hat{\sigma}_k^2})$
42,18106124	6,494694853	0,153971822	82,58265625	13,43345899	-6,147534771

$P(\widehat{\vartheta}) = \frac{1}{\sqrt{2\pi\hat{\sigma}_j^2}} exp^{(-\frac{(\vartheta-\hat{\mu}_k)^2}{2\hat{\sigma}_k^2})}$
0,000329307

Data Multinomial (Stroke)

Gender	Hypertension	Heart Disease	Ever Married	Work Type	Residence Type	Smoking Status
Male	0	1	Yes	Private	Urban	formerly smoked
Male	0	1	Yes	Private	Rural	never smoked
Female	0	0	Yes	Private	Urban	smokes
Female	1	0	Yes	Self-employed	Rural	never smoked
Male	0	0	Yes	Private	Urban	formerly smoked
Male	1	1	Yes	Private	Rural	never smoked
Female	0	0	No	Private	Urban	never smoked
Female	0	0	Yes	Private	Urban	Unknown
Female	1	0	Yes	Private	Rural	never smoked
Female	0	1	Yes	Govt_job	Rural	smokes
Female	0	0	Yes	Private	Urban	smokes
Female	0	1	Yes	Private	Urban	never smoked
Female	1	0	Yes	Self-employed	Rural	never smoked
Male	0	1	Yes	Private	Urban	smokes
Male	1	0	Yes	Private	Urban	smokes
Female	0	0	No	Private	Urban	never smoked
Female	0	0	Yes	Govt_job	Rural	smokes
Female	1	0	Yes	Self-employed	Urban	never smoked
Female	0	0	Yes	Self-employed	Urban	never smoked

Male	0	1	Yes	Private	Rural	Unknown
Male	0	0	Yes	Private	Urban	formerly smoked
Male	0	0	Yes	Self-employed	Rural	never smoked
Female	0	0	Yes	Private	Rural	formerly smoked
Male	0	1	Yes	Self-employed	Urban	smokes
Male	1	0	Yes	Private	Urban	smokes
Male	0	0	Yes	Private	Rural	Unknown
Female	1	0	Yes	Self-employed	Urban	never smoked
Male	0	1	Yes	Self-employed	Urban	formerly smoked
Male	0	0	No	Govt_job	Urban	never smoked
Female	1	1	No	Private	Rural	formerly smoked
Male	0	0	Yes	Private	Rural	formerly smoked
Female	1	0	Yes	Private	Rural	formerly smoked
Male	0	0	No	Private	Rural	Unknown
Female	0	0	Yes	Private	Urban	never smoked
Male	0	0	Yes	Private	Rural	formerly smoked
Male	0	0	Yes	Private	Urban	never smoked
Male	0	1	Yes	Private	Urban	smokes
Male	1	0	Yes	Govt_job	Urban	smokes
Male	1	0	Yes	Private	Rural	never smoked
Female	0	0	Yes	Private	Urban	formerly smoked

Lampiran IV Perhitungan Multinomial Naive Bayes (Stroke)

1 Gender (Perhitungan Multinomial Untuk Variabel Gender (40 Data Sample))

Kriteria	E1 (Male)	E2 (Female)	Total
Frekuensi (X)	21	19	40 = n
Peluang (P)	0,525	0,475	1

$P(x_1,x_2,x_3,.....x_k)$	$\frac{n!}{x_1!,x_2!,x_3!,...x_k!}$	$p_1^{x_1}$	$p_2^{x_2}$	Total
$P(x_1,x_2,x_3,.....x_k)$	1,31282E+11	1,32845E-06	7,19745E-07	0,125525075
$P(x_1) = "0"$	1	0,525	1	0,525

2 Hypertension (Perhitungan Multinomial Untuk Variabel Hypertension (40 Data Sample))

Kriteria	E1 ("0")	E2 ("1")	Total
Frekuensi (X)	28	12	40 = n
Peluang (P)	0,7	0,3	1

$P(x_1,x_2,x_3,.....x_k)$	$\frac{n!}{x_1!,x_2!,x_3!,...x_k!}$	$p_1^{x_1}$	$p_2^{x_2}$	Total
$P(x_1,x_2,x_3,.....x_k)$	5586853480	4,59987E-05	5,31441E-07	0,136573821
$P(x_1) = "0"$	1	0,7	1	0,7

3 Heart Disease (Perhitungan Multinomial Untuk Variabel Heart Disease (40 Data Sample))

Kriteria	E1 ("1")	E2 ("0")	Total
Frekuensi (X)	11	29	40 = n
Peluang (P)	0,275	0,725	1

$P(x_1,x_2,x_3,.....x_k)$	$\frac{n!}{x_1!,x_2!,x_3!,...x_k!}$	$p_1^{x_1}$	$p_2^{x_2}$	Total
$P(x_1,x_2,x_3,.....x_k)$	2311801440	6,80236E-07	8,90845E-05	0,140091683
$P(x_1) = "1"$	1	0,275	1	0,275

4 Ever Married (Perhitungan Multinomial Untuk Variabel Ever Married (40 Data Sample))

Kriteria	E1 ("Yes")	E2 ("No")	Total
Frekuensi	35	5	40 = n
Peluang	0,875	0,125	1

$P(x_1,x_2,x_3,.....x_k)$	$\frac{n!}{x_1!,x_2!,x_3!,...x_k!}$	$p_1^{x_1}$	$p_2^{x_2}$	Total
$P(x_1,x_2,x_3,.....x_k)$	658008	0,009338602	3,05176E-05	0,187526693
$P(x_1) =$ "Yes"	1	0,875	1	0,875

5 Work Type (Perhitungan Multinomial Untuk Variabel Work Type (40 Data Sample))

Kriteria	E1 ("Private")	E2 ("Self-employed")	E3 ("Govt_job")	E4 ("children")	Total
Frekuensi (X)	28	8	4	0	40 = n
Peluang (P)	0,7	0,2	0,1		1

$P(x_1,x_2,x_3,.....x_k)$	$\frac{n!}{x_1!,x_2!,x_3!,...x_k!}$	$p_1^{x_1}$	$p_2^{x_2}$	$p_3^{x_3}$	$p_4^{x_4}$	Total
$P(x_1,x_2,x_3,.....x_k)$	2,76549E+12	4,59987E-05	0,00000256	0,0001		0,032565486
	1	0,7	1	1		0,7

6 Residence Type (Perhitungan Multinomial Untuk Variabel Residence Type (40 Data Sample))

Kriteria	E1 ("Urban")	E2 ("Rural")	Total
Frekuensi (F)	23	17	40 = n
Peluang (P)	0,575	0,425	1

$P(x_1,x_2,x_3,.....x_k)$	$\frac{n!}{x_1!,x_2!,x_3!,...x_k!}$	$p_1^{x_1}$	$p_2^{x_2}$	Total
$P(x_1,x_2,x_3,.....x_k)$	88732378800	2,96729E-06	4,81517E-07	0,126781056
$P(x_1) =$ "Urban"	1	0,575	1	0,575

7 Smoking Status (Perhitungan Multinomial Untuk Variabel Smoking Status (40 Data Sample))

Kriteria	E1 ("formerly smoked")	E2 ("never smoked")	E3 ("smokes")	E4 ("Unknown")	Total
Frekuensi (X)	10	16	10	4	40
Peluang (P)	0,25	0,4	0,25	0,1	1

$P(x_1,x_2,x_3,.....x_k)$	$\frac{n!}{x_1!, x_2!, x_3!, \dots x_k!}$	$p_1^{x_1}$	$p_2^{x_2}$	$p_3^{x_3}$	$p_4^{x_4}$	Total
$P(x_1,x_2,x_3,.....x_k)$	1,23392E+20	9,53674E-07	4,29497E-07	9,53674E-07	0,0001	0,004820013
	1	0,25	1	1	1	0,25

Data Multinomial (Tidak Stroke)

Gender	Hypertension	Heart Disease	Ever Married	Work Type	Residence Type	Smoking Status
Male	0	0	No	children	Rural	Unknown
Male	1	0	Yes	Private	Urban	never smoked
Female	0	0	No	Private	Urban	Unknown
Female	0	0	Yes	Private	Rural	formerly smoked
Female	0	0	Yes	Private	Urban	formerly smoked
Female	0	1	Yes	Self-employed	Rural	never smoked
Female	0	0	Yes	Private	Rural	smokes
Female	0	0	Yes	Govt_job	Urban	Unknown
Male	0	1	Yes	Private	Urban	formerly smoked
Female	0	0	Yes	Private	Rural	never smoked
Female	0	0	Yes	Private	Rural	formerly smoked
Female	0	0	Yes	Private	Rural	never smoked
Male	0	0	No	Private	Rural	never smoked
Female	0	0	No	Private	Urban	smokes
Female	0	0	Yes	Private	Rural	never smoked
Female	0	0	Yes	Govt_job	Urban	smokes
Female	0	1	Yes	Self-employed	Urban	smokes
Female	1	0	Yes	Private	Rural	Unknown
Female	0	0	Yes	Private	Rural	smokes
Male	0	0	Yes	Private	Urban	formerly smoked
Female	0	0	Yes	Private	Urban	never smoked
Female	0	0	Yes	Private	Urban	never smoked
Female	0	0	Yes	Govt_job	Rural	smokes
Female	0	0	No	Private	Rural	never smoked
Female	0	0	Yes	Private	Rural	never smoked
Female	0	0	Yes	Self-employed	Rural	never smoked
Male	0	0	Yes	Self-employed	Urban	Unknown

Female	0	0	Yes	Private	Urban	Unknown
Male	1	0	No	Private	Urban	smokes
Female	0	0	No	children	Urban	Unknown
Female	0	0	Yes	Private	Urban	Unknown
Male	0	0	No	Govt_job	Urban	Unknown
Female	0	0	Yes	Private	Rural	formerly smoked
Female	0	0	No	Private	Urban	never smoked
Female	0	0	Yes	Private	Urban	Unknown
Female	0	0	No	Self-employed	Urban	Unknown
Male	0	0	Yes	Private	Urban	smokes
Male	0	0	No	children	Urban	Unknown
Male	0	0	No	children	Rural	Unknown
Female	0	0	No	children	Rural	Unknown

Lampiran V Perhitungan Multinomial Naive Bayes (Tidak *Stroke*)

1 Gender Perhitungan Multinomial Untuk Variabel Gender (40 Data Sample)

Kriteria	E1 (Male)	E2 (Female)	Total
Frekuensi (X)	11	29	40 = n
Peluang (P)	0,275	0,725	1

$P(x_1,x_2,x_3,.....x_k)$	$\frac{n!}{x_1!,x_2!,x_3!,...x_k!}$	$p_1^{x_1}$	$p_2^{x_2}$	Total
$P(x_1,x_2,x_3,.....x_k)$	2311801440	6,80236E-07	8,90845E-05	0,140091683
	1	0,275	1	0,275

2 Hypertension Perhitungan Multinomial Untuk Variabel Hypertension (40 Data Sample)

Kriteria	E1 ("0")	E2 ("1")	Total
Frekuensi (X)	37	3	40 = n
Peluang (P)	0,925	0,075	1

$P(x_1,x_2,x_3,.....x_k)$	$\frac{n!}{x_1!,x_2!,x_3!,...x_k!}$	$p_1^{x_1}$	$p_2^{x_2}$	Total
$P(x_1,x_2,x_3,.....x_k)$	9880	0,055878419	0,000421875	0,232908236
$P(x_1) = "0"$	1	0,925	1	0,925

3 Heart Disease Perhitungan Multinomial Untuk Variabel Heart Disease (40 Data Sample)

Kriteria	E1 ("1")	E2 ("0")	Total
Frekuensi (X)	3	37	40 = n
Peluang (P)	0,075	0,925	1

$P(x_1,x_2,x_3,.....x_k)$	$\frac{n!}{x_1!,x_2!,x_3!,...x_k!}$	$p_1^{x_1}$	$p_2^{x_2}$	Total
$P(x_1,x_2,x_3,.....x_k)$	9880	0,000421875	0,055878419	0,232908236
$P(x_1) = "0"$	1	1	0,925	0,925

4 Ever Married (Perhitungan Multinomial Untuk Variabel Ever Married (40 Data Sample))

Kriteria	E1 ("Yes")	E2 ("No")	Total
Frekuensi	27	13	40 = n
Peluang	0,675	0,325	1

$P(x_1,x_2,x_3,.....x_k)$	$\frac{n!}{x_1!,x_2!,x_3!,...x_k!}$	$p_1^{x_1}$	$p_2^{x_2}$	Total
$P(x_1,x_2,x_3,.....x_k)$	12033222880	2,46151E-05	4,51319E-07	0,133680333
	1	1	0,325	0,325

5 Work Type (Perhitungan Multinomial Untuk Variabel Work Type (40 Data Sample))

Kriteria	E1 ("Private")	E2 ("Self-employed")	E3 ("Govt_job")	E4 ("children")	Total
Frekuensi (X)	26	5	4	5	40 = n
Peluang (P)	0,65	0,125	0,1	0,125	1

$P(x_1,x_2,x_3,.....x_k)$	$\frac{n!}{x_1!,x_2!,x_3!,...x_k!}$	$p_1^{x_1}$	$p_2^{x_2}$	$p_3^{x_3}$	$p_4^{x_4}$	Total
$P(x_1,x_2,x_3,.....x_k)$	5,85399E+15	1,36693E-05	3,05176E-05	0,0001	3,052E-05	244,20
	1	1	1	1	0,125	0,125

6 Residence Type (Perhitungan Multinomial Untuk Variabel Residence Type (40 Data Sample))

Kriteria	E1 ("Urban")	E2 ("Rural")	Total
Frekuensi (F)	22	18	40 = n
Peluang (P)	0,55	0,45	1

$P(x_1,x_2,x_3,.....x_k)$	$\frac{n!}{x_1!,x_2!,x_3!,...x_k!}$	$p_1^{x_1}$	$p_2^{x_2}$	Total
$P(x_1,x_2,x_3,.....x_k)$	1,1338E+11	1,94079E-06	5,72566E-07	0,125991682
$P(x_1) =$ "Rural"	1	1	0,45	0,45

7 Smoking Status (Perhitungan Multinomial Untuk Variabel Smoking Status (40 Data Sample))

Kriteria	E1 ("formerly smoked")	E2 ("never smoked")	E3 ("smokes")	E4 ("Unknown")	Total
Frekuensi (X)	6	12	8	14	40 = n
Peluang (P)	0,15	0,3	0,2	0,35	1

$P(x_1,x_2,x_3,.....x_k)$	$\frac{n!}{x_1!,x_2!,x_3!,...x_k!}$	$p_1^{x_1}$	$p_2^{x_2}$	$p_3^{x_3}$	$p_4^{x_4}$	Total
$P(x_1,x_2,x_3,.....x_k)$	6,73049E+20	1,13906E-05	5,31441E-07	0,000 00256	4,14E-07	0,004317595
$P(x_1)$"Unknown"	1	1	1	1	0,35	0,35