

Analisis Dampak Anonimisasi LLM terhadap Kinerja Model Klasifikasi Teks dari Dataset Tabular yang Diserialisasi.

Nama Anggota Kelompok:

[202210370311489 - Qinada Farah]
[202210370311368 - Rosita Dwi Yulianti]

Tahap 0 (Poin: 25): Business Objective

Dalam dunia kesehatan, data rekam medis elektronik (EHR) memiliki peran penting untuk analisis dan penelitian berbasis kecerdasan buatan (AI). Namun, tantangan yang sering muncul adalah keragaman format data yang ada serta perlindungan privasi pasien. Data medis umumnya disimpan dalam format terstruktur yang mempersulit analisis berbasis teks alami. Oleh karena itu, penelitian ini bertujuan untuk mengimplementasikan pipeline berbasis Large Language Model (LLM) yang dapat mentransformasi data tabular menjadi narasi teks alami sambil menjaga privasi pasien. Pipeline ini mencakup proses serialisasi, anonimisasi, dan preprocessing teks medis yang akan digunakan untuk analisis lebih lanjut menggunakan teknik machine learning.

Dengan menerapkan pendekatan *hybrid anonymization*, termasuk *pseudonymization*, *generalization*, dan *semantic perturbation*, proyek ini memungkinkan proses transformasi data medis yang terstruktur menjadi bentuk yang lebih ramah untuk analisis semantik, seperti klasifikasi teks atau NLP lainnya. Pendekatan ini diharapkan dapat menjadi solusi dalam menjaga keseimbangan antara pemrosesan data yang efektif dan perlindungan privasi, serta mendorong pengembangan sistem pembelajaran mesin yang lebih aman di bidang kesehatan.

Tahap 1: Original Data (25 Poin)

Pemanfaatan data medis untuk keperluan penelitian atau analisis berbasis AI sering kali terhambat oleh dua faktor utama: pertama, data medis biasanya tersimpan dalam format

terstruktur yang tidak langsung dapat diproses oleh model berbasis NLP. Kedua, data medis mengandung informasi sensitif yang harus dilindungi dengan teknik anonimisasi yang tepat untuk memastikan privasi pasien tetap terjaga.

Sumber Data

| Sumber Data | Link |
|-------------|-------------------------|
| Dataset | Dataset |
| Link Collab | colab |

Tujuan dari data mining task ini adalah untuk mengubah data tabular menjadi teks alami dan kemudian menganonimkan informasi sensitif melalui teknik hybrid anonymization. Proses ini akan memungkinkan analisis lebih lanjut menggunakan model machine learning sambil memastikan bahwa data medis yang digunakan tetap menjaga kerahasiaan pasien. Pemanfaatan data medis untuk keperluan penelitian atau analisis berbasis AI sering kali terhambat oleh dua faktor utama: pertama, data medis biasanya tersimpan dalam format terstruktur yang tidak langsung dapat diproses oleh model berbasis NLP. Kedua, data medis mengandung informasi sensitif yang harus dilindungi dengan teknik anonimisasi yang tepat untuk memastikan privasi pasien tetap terjaga.

Deskripsi Data yang Digunakan

Dataset yang digunakan dalam penelitian ini terdiri dari dua dataset utama:

1. Demographics.csv

Dataset ini berisi informasi demografis pasien yang meliputi data pribadi dasar seperti ID pasien, jenis kelamin, usia, serta status fisik dan kondisi medis yang relevan. Data ini sangat penting untuk memahami latar belakang pasien secara umum.

Jumlah baris: 1541 baris.

Jumlah kolom: 5 kolom.

2. ClinicalData.csv

Dataset ini berisi informasi klinis yang lebih mendalam mengenai prosedur medis yang dijalani pasien, hasil tes laboratorium, riwayat medis, serta pengobatan yang diterima pasien. Data ini berfungsi untuk memberikan gambaran lebih spesifik terkait kondisi medis pasien.

Jumlah baris: 6388 baris.

Jumlah kolom: 5 kolom.

Atribut pada Data

1. Demographics.csv

Tabel 1.1 Fitur Target Pada Dataset Demographic

| Fitur | Deskripsi |
|--------------------|--|
| Patient_ID | ID unik untuk setiap pasien. |
| Gender | Jenis kelamin pasien (misalnya: laki-laki, perempuan). |
| Age | Usia pasien (misalnya: 45 tahun). |
| Diagnosis | Diagnosis medis atau kondisi yang dialami pasien. |
| Hospital_ID | ID rumah sakit tempat pasien dirawat. |

2. ClinicalData.csv

Tabel 1.2 Fitur Target Pada Dataset ClinicalData

| Fitur | Deskripsi |
|-------------------|--|
| Patient_ID | ID unik yang mengidentifikasi pasien untuk menghubungkan data demografis dengan data medis. |
| Procedure_Type | Jenis prosedur medis atau tindakan yang dilakukan pada pasien (misalnya: operasi, pemeriksaan, dll). |
| Anesthesia_Type | Hasil tes medis atau laboratorium yang dilakukan pada pasien. |
| Treatment_Details | Rincian pengobatan yang diterima pasien. |
| Hospital_ID | ID rumah sakit tempat pasien dirawat. |

Tujuan Data Mining Task

Tujuan dari data mining task ini adalah untuk mengubah data tabular menjadi teks alami yang lebih mudah dipahami dan kemudian menganonimkan informasi sensitif melalui teknik *hybrid anonymization*. Proses ini akan memungkinkan analisis lebih lanjut menggunakan model machine learning, sambil memastikan bahwa data medis yang digunakan tetap menjaga kerahasiaan pasien.

Tahap 2: Serialization

Serialization merupakan proses mengubah data tabular menjadi teks alami agar informasi yang ada dapat dianalisis menggunakan teknik pemrosesan bahasa alami (NLP). Dalam tahap ini, data yang sudah ada dalam bentuk tabel (seperti dataset Demographics dan ClinicalData) dikonversi menjadi narasi teks yang lebih mudah dipahami.

Proses Serialization:

- Untuk Demographics, data yang mencakup informasi seperti ID Pasien, jenis kelamin, usia, ras, status pasien, riwayat penyakit, rumah sakit, dan tanggal masuk diubah menjadi narasi teks.
- Setiap baris data pasien diubah menjadi kalimat yang mencakup semua informasi relevan dalam format naratif, contoh:
 - *"Seorang pasien dengan ID Pasien 1, laki-laki berusia 93 tahun dari ras kulit hitam, dirawat di rumah sakit SUNY. Ia masuk pada hari Selasa, 07 April 2020 pukul 15:09. Pasien ini memiliki riwayat hipertensi..."*
- Untuk ClinicalData, data medis seperti nomor identifikasi pasien, umur, jenis kelamin, status fisik ASA, jenis prosedur medis yang dijalani, serta hasil pemeriksaan laboratorium dan riwayat medis, juga diserialisasikan ke dalam narasi teks.
- Contoh:
 - *"Pasien dengan nomor identifikasi 5955 adalah seorang pria berusia 77 tahun dengan tinggi 160.02 cm, berat 67.05 kg, dan BMI 26.03. Status fisik ASA pasien adalah 2..."*

Prompt LLM:

Proses serialization ini dilakukan dengan bantuan model bahasa seperti Gemini 2.5 Flash dari Google. Sebuah prompt digunakan untuk memberikan instruksi kepada model AI

agar menghasilkan narasi teks yang relevan dan sesuai dengan aturan yang ditetapkan, seperti mengganti kata-kata tertentu dengan istilah medis yang tepat dan memastikan informasi pasien terstruktur dalam teks.

Hasil Serialization:

Setelah proses selesai, setiap entri dalam dataset yang awalnya berbentuk tabel kini berubah menjadi teks naratif.

Tahap 3: Data Anonymization

Pada tahap ini, proses anonimisasi diterapkan untuk melindungi identitas pribadi dalam dataset, dengan menggunakan teknik *pseudonymization*, *generalization*, dan *perturbation*. Anonimisasi bertujuan untuk menjaga privasi individu dalam data, sehingga data tetap dapat digunakan untuk analisis tanpa mengungkapkan informasi sensitif.

1. Pseudonymization (Pemberian Token Pseudonim)

Teknik pseudonimisasi dilakukan dengan mengganti ID pasien dan nama rumah sakit dengan token acak yang dihasilkan melalui fungsi hash. Misalnya, "ID Pasien 4461" diubah menjadi token seperti "Patient_4461". Ini dilakukan menggunakan algoritma *hash Blake2b* untuk menghasilkan token unik dan sulit untuk dikembalikan ke data asli.

2. Generalisasi (Mengaburkan Informasi Sensitif)

Dalam generalisasi, informasi seperti usia dan tanggal spesifik diubah menjadi kategori yang lebih umum untuk mengurangi kemungkinan identifikasi individu.

- **Usia** digeneralisasi menjadi kategori seperti "20–39 tahun" atau "40–59 tahun", mengelompokkan usia individu dalam rentang umur tertentu.

- **Tanggal** digeneralisasi untuk menghapus detail spesifik dan mengganti dengan format seperti "Bulan Mei Tahun 2021".

- **Waktu** juga disamarkan, misalnya, "pukul 14:30" digantikan dengan "pukul sekitar jam tersebut".

3. Perturbasi (Modifikasi Semantik untuk L-Diversity)

Teknik perturbasi dilakukan dengan mengganti istilah medis sensitif menggunakan variasi yang lebih umum atau padanan lain, seperti mengganti "hipertensi" menjadi "tekanan darah tinggi" atau "diabetes mellitus" menjadi "kadar gula tinggi". Tujuannya adalah untuk memastikan bahwa kelompok data yang serupa tetap memiliki keragaman (*L-diversity*) agar sulit untuk merekonstruksi informasi pribadi.

4. Implementasi pada Dataset

Anonimisasi diterapkan pada dua dataset: **Demographics** dan **ClinicalData.csv**. Kedua dataset tersebut diolah menggunakan fungsi yang telah dijelaskan di atas, menghasilkan teks anonim yang siap untuk digunakan dalam analisis lebih lanjut tanpa mengungkapkan informasi pribadi.

5. Hasil Anonimisasi

Dataset yang telah dianonimisasi berisi teks yang telah melalui proses *Pseudonymization*, *Generalization*, dan *Perturbation*.

A. Demographic Dataset

Tabel 3.1 Hasil Anonimisasi Demographic Dataset

| Raw Data | Anonimisasi Data |
|--|--|
| Seorang pasien dengan Patient_4461, laki-laki berusia 80-99 tahun, dirawat di Rumah Sakit RS-A | Seorang pasien dengan Patient_4461, laki-laki berusia 80-99 tahun, dirawat di Rumah Sakit RS-A |

| | |
|--|--|
| Seorang pasien dengan Patient_5531, perempuan berusia 60-79 tahun, dirawat di Rumah Sakit RS-B | Seorang pasien dengan Patient_BA75, perempuan berusia 60-79 tahun, dirawat di Rumah Sakit RS-B |
|--|--|

B. Clinical Dataset

Tabel 3.2 Hasil Anonimisasi Clinical Dataset

| Raw Dataset | Anonimisasi Dataset |
|--|---|
| Pasien dengan Patient_7056 adalah seorang pria berusia 40-59 tahun, dirawat dengan diagnosis diabetes mellitus | Pasien dengan Patient_CE17 adalah seorang pria berusia 60-79 tahun, dirawat karena hipertensi |

Tahap 4: Data Pre-processing & Transformation (25 Poin)

Pada tahap pre-processing, dua dataset yang digunakan, yaitu **Demographics** dan **ClinicalData**, mengalami serangkaian langkah pembersihan dan transformasi untuk mempersiapkan data sebelum digunakan dalam model pembelajaran mesin.

1. Data Cleaning

- Demographics:** Dalam dataset ini, dilakukan pembersihan data dengan:
 - Lowercasing:* Semua teks diubah menjadi huruf kecil untuk konsistensi.
 - Penghapusan Tanda Baca:* Tanda baca yang tidak relevan dihapus.
 - Pembersihan Angka:* Angka yang tidak relevan dibersihkan, sementara angka terkait rentang usia dipertahankan.
 - Penghapusan Stopwords:* Kata-kata umum seperti "yang", "dan", "di", dan lainnya dihapus untuk mengurangi kebisingan dalam teks.

Tabel 4.1 Hasil Preprocessing Teks Dataset Demographic

| Dataset Sebelum Preprocessing | Dataset Setelah Preprocessing |
|--|---|
| <p>Seorang pasien dengan Patient_4461, laki-laki berusia 80–99 tahun dari ras tidak disebutkan, dirawat di Hospital_4AE2. Ia masuk pada hari Selasa, April 2020 pukul sekitar jam tersebut. Pasien ini memiliki riwayat hipertensi ringan. Penyakit lain seperti hiperlipidemia, kadar gula tinggi, penyakit arteri koroner, gagal jantung, penyakit serebrovaskular, hepatitis, gagal ginjal tahap akhir, penyakit ginjal kronis, asma, penyakit paru obstruktif kronis, demensia ringan, dan kanker tidak tercatat. Status pasien berubah pada hari Sabtu, April 2020 pukul sekitar jam tersebut, dengan outcome deceased.</p> | <p>pasien patienttoken laki-laki berusia 80-99 ras dirawat hospitaltoken masuk selasa april jam pasien memiliki riwayat hipertensi ringan penyakit hiperlipidemia kadar gula penyakit arteri koroner gagal jantung penyakit serebrovaskular hepatitis gagal ginjal tahap penyakit ginjal kronis asma penyakit paru obstruktif kronis demensia ringan kanker tercatat status pasien berubah sabtu april jam outcome deceased</p> |
| <p>Seorang pasien dengan Patient_BA75, perempuan berusia 80–99 tahun dari ras tidak disebutkan, dirawat di Hospital_4AE2. Ia masuk pada hari Senin, Maret 2020 pukul sekitar jam tersebut. Pasien ini memiliki riwayat penyakit arteri koroner dan gangguan kognitif. Penyakit lain seperti tekanan darah tinggi, hiperlipidemia, diabetes mellitus, gagal jantung, penyakit serebrovaskular, hepatitis, gagal ginjal tahap akhir,</p> | <p>pasien patienttoken perempuan berusia 80-99 ras dirawat hospitaltoken masuk senin maret jam pasien memiliki riwayat penyakit arteri koroner gangguan kognitif penyakit tekanan darah hiperlipidemia diabetes mellitus gagal jantung penyakit serebrovaskular hepatitis gagal ginjal tahap penyakit ginjal kronis asma penyakit paru obstruktif kronis kanker tercatat status pasien berubah minggu april jam outcome passed away</p> |

| | |
|--|---|
| <p>penyakit ginjal kronis, asma, penyakit paru obstruktif kronis, dan kanker tidak tercatat. Status pasien berubah pada hari Minggu, April 2020 pukul sekitar jam tersebut, dengan outcome passed away.</p> | |
| <p>Seorang pasien dengan Patient_0967, perempuan berusia 80–99 tahun dari ras tidak disebutkan, dirawat di Hospital_4AE2. Ia masuk pada hari Jumat, April 2020 pukul sekitar jam tersebut. Pasien ini memiliki riwayat tekanan darah tinggi. Penyakit lain seperti hiperlipidemia, diabetes mellitus, penyakit arteri koroner, gagal jantung, penyakit serebrovaskular, hepatitis, gagal ginjal tahap akhir, penyakit ginjal kronis, asma, penyakit paru obstruktif kronis, gangguan kognitif, dan kanker tidak tercatat. Status pasien berubah pada hari Rabu, April 2020 pukul sekitar jam tersebut, dengan outcome passed away.</p> | <p>pasien patienttoken perempuan berusia 80-99 ras dirawat hospitaltoken masuk jumat april jam pasien memiliki riwayat tekanan darah penyakit hiperlipidemia diabetes mellitus penyakit arteri koroner gagal jantung penyakit serebrovaskular hepatitis gagal ginjal tahap penyakit ginjal kronis asma penyakit paru obstruktif kronis gangguan kognitif kanker tercatat status pasien berubah rabu april jam outcome passed away</p> |
| <p>Seorang pasien dengan Patient_88ED, perempuan berusia 80–99 tahun dari ras tidak disebutkan, dirawat di Hospital_4AE2. Ia masuk pada hari Minggu, Maret 2020 pukul sekitar jam tersebut. Pasien ini memiliki riwayat tekanan darah tinggi dan asma. Penyakit lain seperti hiperlipidemia, diabetes</p> | <p>pasien patienttoken perempuan berusia 80-99 ras dirawat hospitaltoken masuk minggu maret jam pasien memiliki riwayat tekanan darah asma penyakit hiperlipidemia diabetes mellitus penyakit arteri koroner gagal jantung penyakit serebrovaskular hepatitis gagal ginjal tahap penyakit ginjal kronis penyakit paru</p> |

| | |
|--|---|
| <p>mellitus, penyakit arteri koroner, gagal jantung, penyakit serebrovaskular, hepatitis, gagal ginjal tahap akhir, penyakit ginjal kronis, penyakit paru obstruktif kronis, demensia ringan, dan kanker tidak tercatat. Status pasien berubah pada hari Selasa, Maret 2020 pukul sekitar jam tersebut, dengan outcome passed away.</p> | <p>obstruktif kronis demensia ringan kanker tercatat status pasien berubah selasa maret jam outcome passed away</p> |
| <p>Seorang pasien dengan Patient_56D5, perempuan berusia 80–99 tahun dari ras tidak disebutkan, dirawat di Hospital_4AE2. Ia masuk pada hari Senin, April 2020 pukul sekitar jam tersebut. Pasien ini memiliki riwayat tekanan darah tinggi, diabetes mellitus, dan penyakit arteri koroner. Penyakit lain seperti hiperlipidemia, gagal jantung, penyakit serebrovaskular, hepatitis, gagal ginjal tahap akhir, penyakit ginjal kronis, asma, penyakit paru obstruktif kronis, demensia ringan, dan kanker tidak tercatat. Status pasien berubah pada hari Rabu, April 2020 pukul sekitar jam tersebut, dengan outcome passed away.</p> | <p>pasien patienttoken perempuan berusia 80-99 ras dirawat hospitaltoken masuk senin april jam pasien memiliki riwayat tekanan darah diabetes mellitus penyakit arteri koroner penyakit hiperlipidemia gagal jantung penyakit serebrovaskular hepatitis gagal ginjal tahap penyakit ginjal kronis asma penyakit paru obstruktif kronis demensia ringan kanker tercatat status pasien berubah rabu april jam outcome passed away</p> |

- b. **ClinicalData:** Proses yang sama diterapkan pada dataset ini, namun dengan tambahan langkah pembersihan:

- Penghapusan Simbol Umum: Simbol unit ukuran (seperti cm, kg, ml) yang tidak diperlukan dalam analisis lanjutan dihapus.
- Normalisasi Tanda Hubung Unicode: Tanda hubung dengan format yang berbeda (seperti “–” dan “—”) disesuaikan menjadi tanda hubung standar (“-”) agar format teks lebih konsisten.

Tabel 4.2 Hasil Preprocessing Teks Clinical Dataset

| Dataset Sebelum Preprocessing | Dataset Setelah Preprocessing |
|--|---|
| <p>Pasien dengan Patient_CE17 adalah seorang pria berusia 60–79 tahun dengan tinggi <angka_tersembunyi> cm, berat <angka_tersembunyi> kg, dan BMI <angka_tersembunyi>. Status fisik ASA pasien adalah 2. Ia menjalani prosedur non-darurat bedah umum untuk kanker saluran pencernaan bawah, yaitu reseksi anterior rendah dengan pendekatan terbuka dan posisi litotomi. Anestesi yang digunakan adalah anestesi umum. Riwayat medis pasien mencakup hipertensi ringan pre-operasi, namun tidak ada diabetes ringan. Hasil EKG pre-operasi menunjukkan irama sinus normal, dan tes fungsi paru juga normal. Pemeriksaan laboratorium pre-operasi menunjukkan hemoglobin <angka_tersembunyi>, platelet <angka_tersembunyi>, PT <angka_tersembunyi>, aPTT</p> | <p>pasien patienttoken pria berusia 60-79 numtoken berat numtoken bmi numtoken status fisik asa pasien menjalani prosedur non-darurat bedah kanker saluran pencernaan reseksi anterior rendah pendekatan terbuka posisi litotomi anestesi anestesi riwayat medis pasien mencakup hipertensi ringan pre-operasi diabetes ringan hasil ekg pre-operasi irama sinus normal tes fungsi paru normal pemeriksaan laboratorium pre-operasi hemoglobin numtoken platelet numtoken pt numtoken aptt numtoken natrium numtoken kalium numtoken glukosa numtoken albumin numtoken ast numtoken alt numtoken bun numtoken kreatinin numtoken penilaian jalan napas cormack jalan napas oral ukuran tabung endotrakeal numtoken akses intravena diperoleh lengan kanan akses arteri radial</p> |

| | | |
|--|--|--|
| <angka_tersembunyi>, <angka_tersembunyi>, <angka_tersembunyi>, <angka_tersembunyi>, <angka_tersembunyi>, <angka_tersembunyi>, <angka_tersembunyi>, <angka_tersembunyi>, <angka_tersembunyi>. Penilaian jalan napas Cormack I dengan jalan napas oral dan ukuran tabung endotrakeal <angka_tersembunyi>. Akses intravena diperoleh di lengan bawah kanan, dan akses arteri radial kiri juga terpasang. Selama operasi, perkiraan kehilangan darah tidak tersedia, dan volume urine yang dikeluarkan adalah <angka_tersembunyi> mL. Pasien tidak menerima transfusi sel darah merah maupun FFP. Cairan kristaloid yang diberikan sebanyak <angka_tersembunyi> mL. Obat-obatan yang digunakan meliputi propofol 120, fentanil 100, rocuronium 70, dan efedrin 10. Pasien tidak menghabiskan hari di ICU dan tidak deceased selama rawat inap. Durasi total kasus adalah <angka_tersembunyi> menit, durasi anestesi <angka_tersembunyi> menit, dan durasi operasi <angka_tersembunyi> menit. | natrium kalium glukosa albumin AST ALT BUN dan kreatinin | kiri terpasang kehilangan darah tersedia dikeluarkan numtoken transfusi sel darah merah kristaloid numtoken propofol fentanil pasien menghabiskan icu inap durasi total numtoken durasi anestesi numtoken durasi operasi numtoken |
|--|--|--|

| | |
|--|--|
| <p>Pasien dengan Patient_7056 adalah seorang pria berusia 40–59 tahun dengan tinggi <angka_tersembunyi> cm, berat <angka_tersembunyi> kg, dan BMI <angka_tersembunyi>. Status fisik ASA pasien adalah 2. Ia menjalani prosedur non-darurat bedah umum untuk kanker saluran pencernaan atas stadium lanjut, yaitu gastrektomi subtotal dengan pendekatan terbuka dan posisi terlentang. Anestesi yang digunakan adalah anestesi umum. Pasien tidak memiliki riwayat tekanan darah tinggi atau kadar gula tinggi pre-operasi. Hasil EKG pre-operasi menunjukkan irama sinus normal, dan tes fungsi paru juga normal. Pemeriksaan laboratorium pre-operasi menunjukkan hemoglobin <angka_tersembunyi>, platelet <angka_tersembunyi>, PT <angka_tersembunyi>, aPTT <angka_tersembunyi>, natrium <angka_tersembunyi>, kalium <angka_tersembunyi>, glukosa <angka_tersembunyi>, albumin <angka_tersembunyi>, AST <angka_tersembunyi>, ALT <angka_tersembunyi>, BUN <angka_tersembunyi>, dan kreatinin <angka_tersembunyi>. Penilaian jalan napas Cormack I dengan jalan napas oral</p> | <p>pasien patienttoken pria berusia 40-59 numtoken berat numtoken bmi numtoken status fisik asa pasien menjalani prosedur non-darurat bedah kanker saluran pencernaan stadium gastrektomi subtotal pendekatan terbuka posisi terlentang anestesi anestesi pasien memiliki riwayat tekanan darah kadar gula pre-operasi hasil ekg pre-operasi irama sinus normal tes fungsi paru normal pemeriksaan laboratorium pre-operasi hemoglobin numtoken platelet numtoken pt numtoken aptt numtoken natrium numtoken kalium numtoken glukosa numtoken albumin numtoken ast numtoken alt numtoken bun numtoken kreatinin numtoken penilaian jalan napas cormack jalan napas oral ukuran tabung endotrakeal numtoken akses intravena diperoleh lengan kiri operasi perkiraan kehilangan darah numtoken volume urine dikeluarkan numtoken pasien menerima transfusi sel darah merah ffp cairan kristaloid numtoken obat-obatan meliputi propofol rocuronium efedrin pasien menghabiskan icu selamat rawat inap durasi total numtoken durasi anestesi numtoken durasi operasi numtoken</p> |
|--|--|

dan ukuran tabung endotrakeal <angka_tersembunyi>. Akses intravena diperoleh di lengan bawah kiri. Selama operasi, perkiraan kehilangan darah adalah <angka_tersembunyi> mL, dan volume urine yang dikeluarkan adalah <angka_tersembunyi> mL. Pasien tidak menerima transfusi sel darah merah maupun FFP. Cairan kristaloid yang diberikan sebanyak <angka_tersembunyi> mL. Obat-obatan yang digunakan meliputi propofol 150, rocuronium 100, dan efedrin 20. Pasien tidak menghabiskan hari di ICU dan tidak selamat selama rawat inap. Durasi total kasus adalah <angka_tersembunyi> menit, durasi anestesi <angka_tersembunyi> menit, dan durasi operasi <angka_tersembunyi> menit.

2. Data Transformation

Setelah data dibersihkan, tahap berikutnya adalah *Feature Extraction* menggunakan tiga pendekatan:

- **BoW:** Teknik ini mengonversi setiap kata menjadi fitur dengan representasi vektor, di mana setiap kata diwakili sebagai kolom dalam matriks fitur.
- **TF-IDF:** Teknik ini memberikan bobot lebih pada kata-kata yang jarang muncul namun penting dalam konteks dokumen.

- **N-Gram:** Teknik ini menangkap hubungan antar kata berturut-turut untuk mempelajari konteks yang lebih kompleks.

Setiap teknik ini menghasilkan matriks fitur yang dapat digunakan dalam pelatihan model pembelajaran mesin, yang merepresentasikan frekuensi kata atau pasangan kata dalam dataset.

3. Data Normalization

Normalisasi dilakukan untuk memastikan bahwa fitur-fitur yang dihasilkan memiliki bobot yang setara, sehingga tidak ada fitur yang mendominasi. Setelah itu, dilakukan seleksi fitur untuk mengurangi fitur yang tidak relevan atau berlebihan, yang dapat mempengaruhi performa model.

4. Feature Selection

Pada tahap ini, dilakukan seleksi fitur menggunakan **Chi-Square Test** untuk memilih fitur terbaik dari berbagai representasi data teks, yaitu BoW (Bag of Words), TF-IDF, dan Bigram (N-Gram). Tujuan dari seleksi fitur ini adalah untuk mengurangi dimensi data dengan memilih fitur yang paling relevan untuk meningkatkan performa model pembelajaran mesin. Untuk dataset Demographic, seleksi fitur dilakukan menggunakan metode Chi-Square dengan Bag of Words (BoW). Dari 178 fitur yang ada, 178 fitur tetap terpilih setelah seleksi, karena tidak ada fitur yang perlu dihilangkan. Pada seleksi fitur dengan TF-IDF untuk dataset Demographic, setelah seleksi, jumlah fitur yang terpilih tetap sebanyak 178, sama seperti sebelum seleksi. Seleksi fitur Bigram (N-Gram) dilakukan pada dataset Demographic dengan memilih hingga 1000 fitur. Untuk dataset ClinicalData, seleksi fitur menggunakan BoW dilakukan pada 1585 fitur, dan hasilnya menghasilkan 1000 fitur terpilih. Seleksi fitur untuk TF-IDF pada dataset ClinicalData juga dilakukan, menghasilkan 1000 fitur terpilih dari 1585 fitur awal. Pada seleksi fitur Bigram (N-Gram) untuk dataset ClinicalData, 1000 fitur terpilih dipilih dari 11423 fitur awal.

Tabe 4.3 Fitur Teratas - Dataset Demographic (BoW, TF-IDF, N-Gram)

| BoW Fitur Teratas | TF-IDF Fitur Teratas | Bigram Fitur Teratas |
|--------------------------|-----------------------------|-----------------------------|
| 'admisi' | 'admisi' | 'amerika berusia' |
| 'amerika' | 'amerika' | 'april pasien' |
| 'april' | 'april' | 'arteri koroner' |

Tabel 4.4 Fitur Teratas - Dataset ClinicalData (BoW, TF-IDF, N-Gram)

| BoW Fitur Teratas | TF-IDF Fitur Teratas | Bigram Fitur Teratas |
|--------------------------|-----------------------------|-----------------------------|
| 'abdomeni' | 'abdomen' | 'admisi detik' |
| 'abdominoperineal' | 'adenoma' | 'admisi pasien' |
| 'abgl' | 'akalkulus' | 'Akses arteri' |

5. QI Weak Labelling

Setelah tahap pre-processing, dataset diberikan label berdasarkan *Quality of Information* (QI) untuk menandakan seberapa besar kemungkinan informasi pribadi bocor. Berikut langkah-langkah yang dilakukan:

a. Definisi QI Patterns:

- **Usia:** Deteksi pola usia, seperti "berusia 60-79 tahun".
- **Jenis Kelamin:** Mencakup kata seperti "laki-laki", "perempuan", dan "pria".
- **Lokasi:** Mencari kata terkait lokasi seperti "kota", "kabupaten", dan "provinsi".
- **Hasil Medis:** Mencari kata seperti "meninggal dunia", "wafat", dan "deceased".

b. Penilaian QI Score:

Setiap entri dalam dataset diberikan skor **QI** berdasarkan jumlah pola yang ditemukan dalam teks. Skor ini mencerminkan potensi kebocoran informasi pribadi. Semakin banyak pola yang ditemukan, semakin tinggi nilai QI Score.

c. Labeling Berdasarkan QI Score:

- **HIGH_RISK:** Data dengan QI score tinggi dianggap **HIGH_RISK** karena memiliki potensi untuk mengungkapkan informasi pribadi yang lebih banyak, seperti usia, jenis kelamin, lokasi, dan hasil medis. Informasi ini bisa digunakan untuk mengidentifikasi individu dengan lebih mudah.
- **LOW_RISK:** Data dengan QI score rendah dianggap **LOW_RISK** karena memiliki lebih sedikit informasi sensitif atau quasi-identifiers, sehingga lebih sulit untuk diidentifikasi atau dikaitkan dengan individu

tertentu.

d. **Optimal Threshold Selection:**

Nilai **threshold** dipilih untuk membagi data dengan proporsi yang seimbang antara **HIGH_RISK** dan **LOW_RISK**. Hasilnya, data diberi label berdasarkan QI score yang dihitung, dengan threshold yang memastikan distribusi label seimbang antara kedua kategori. Dengan demikian, dataset akan terbagi menjadi dua kelompok dengan keseimbangan yang tepat antara risiko kebocoran informasi dan kerahasiaan.

Tabel 4.5 Hasil Transformasi Dataset Demographics

| No | Data Serialisasi | Data Anonimisasi | Data Preprocessing |
|----|---|--|--|
| 1. | Seorang pasien dengan ID Pasien 1, laki-laki berusia 93 tahun dari ras kulit hitam, dirawat di rumah sakit SUNY. Ia masuk pada hari Selasa, 07 April 2020 pukul 15:09. Pasien ini memiliki riwayat hipertensi. Penyakit lain seperti hiperlipidemia, diabetes, penyakit arteri koroner, gagal jantung, penyakit serebrovaskular, hepatitis, gagal ginjal tahap akhir, penyakit ginjal kronis, asma, penyakit paru obstruktif kronis, demensia, dan kanker tidak tertentu. | Seorang pasien dengan Patient_4461, laki-laki berusia 80–99 tahun dari ras tidak disebutkan, dirawat di Hospital_4AE2. Ia masuk pada hari Selasa, April 2020 pukul sekitar jam tersebut. Pasien ini memiliki riwayat hipertensi ringan. Penyakit lain seperti hiperlipidemia, kadar gula tinggi, penyakit arteri koroner, gagal jantung, penyakit serebrovaskular, hepatitis, gagal ginjal tahap akhir, penyakit ginjal kronis, asma, penyakit paru obstruktif kronis, demensia ringan, dan kanker tidak tertentu. | pasien patienttoken laki-laki berusia 80-99 ras dirawat hospitaltoken masuk selasa april jam pasien memiliki riwayat hipertensi ringan penyakit hiperlipidemia kadar gula penyakit arteri koroner gagal jantung penyakit serebrovaskular hepatitis gagal ginjal tahap penyakit ginjal kronis asma penyakit paru obstruktif kronis demensia ringan kanker tercatat status pasien berubah sabtu april jam outcome deceased |

| | | | |
|----|---|---|--|
| | tercatat. Status pasien berubah pada hari Sabtu, 18 April 2020 pukul 09:51, dengan outcome meninggal dunia. | tercatat. Status pasien berubah pada hari Sabtu, April 2020 pukul sekitar jam tersebut, dengan outcome deceased. | |
| 2. | Seorang pasien dengan ID Pasien 2, perempuan berusia 87 tahun dari ras kulit hitam, dirawat di rumah sakit SUNY. Ia masuk pada hari Senin, 30 Maret 2020 pukul 17:51. Pasien ini memiliki riwayat penyakit arteri koroner dan demensia. Penyakit lain seperti hipertensi, hiperlipidemia, diabetes, gagal jantung, penyakit serebrovaskular, hepatitis, gagal ginjal tahap akhir, penyakit ginjal kronis, asma, penyakit paru obstruktif kronis, dan kanker tidak tercatat. Status pasien berubah pada hari Minggu, 05 April 2020 pukul 12:35, dengan | Seorang pasien dengan Patient_BA75, perempuan berusia 80–99 tahun dari ras tidak disebutkan, dirawat di Hospital_4AE2. Ia masuk pada hari Senin, Maret 2020 pukul sekitar jam tersebut. Pasien ini memiliki riwayat penyakit arteri koroner dan demensia ringan. Penyakit lain seperti hipertensi ringan, hiperlipidemia, kadar gula tinggi, gagal jantung, penyakit serebrovaskular, hepatitis, gagal ginjal tahap akhir, penyakit ginjal kronis, asma, penyakit paru obstruktif kronis, dan kanker tidak tercatat. Status pasien berubah pada hari Minggu, April 2020 pukul sekitar jam tersebut, dengan outcome deceased | pasien patienttoken perempuan berusia 80-99 ras dirawat hospitaltoken masuk senin maret jam pasien memiliki riwayat penyakit arteri koroner demensia ringan penyakit hipertensi ringan hiperlipidemia kadar gula gagal jantung penyakit serebrovaskular hepatitis gagal ginjal tahap penyakit ginjal kronis asma penyakit paru obstruktif kronis kanker tercatat status pasien berubah minggu april jam outcome deceased |

| | | | |
|----|---|---|--|
| | outcome meninggal dunia. | deceased. | |
| 3. | Seorang pasien dengan ID Pasien 3, perempuan berusia 92 tahun dari ras kulit hitam, dirawat di rumah sakit SUNY. Ia masuk pada hari Jumat, 10 April 2020 pukul 22:30. Pasien ini memiliki riwayat hipertensi. Penyakit lain seperti hiperlipidemia, diabetes, penyakit arteri koroner, gagal jantung, penyakit serebrovaskular, hepatitis, gagal ginjal tahap akhir, penyakit ginjal kronis, asma, penyakit paru obstruktif kronis, demensia, dan kanker tidak tercatat. Status pasien berubah pada hari Rabu, 15 April 2020 pukul 10:27, dengan outcome meninggal dunia. | Seorang pasien dengan Patient_0967, perempuan berusia 80–99 tahun dari ras tidak disebutkan, dirawat di Hospital_4AE2. Ia masuk pada hari Jumat, April 2020 pukul sekitar jam tersebut. Pasien ini memiliki riwayat hipertensi ringan. Penyakit lain seperti hiperlipidemia, kadar gula tinggi, penyakit arteri koroner, gagal jantung, penyakit serebrovaskular, hepatitis, gagal ginjal tahap akhir, penyakit ginjal kronis, asma, penyakit paru obstruktif kronis, gangguan kognitif, dan kanker tidak tercatat. Status pasien berubah pada hari Rabu, April 2020 pukul sekitar jam tersebut, dengan outcome deceased. | pasien patienttoken perempuan berusia 80-99 ras dirawat hospitaltoken masuk jumat april jam pasien memiliki riwayat hipertensi ringan penyakit hiperlipidemia kadar gula penyakit arteri koroner gagal jantung penyakit serebrovaskular hepatitis gagal ginjal tahap penyakit ginjal kronis asma penyakit paru obstruktif kronis gangguan kognitif kanker tercatat status pasien berubah rabu april jam outcome deceased |
| 4. | Seorang pasien dengan ID Pasien 4, perempuan berusia 83 tahun dari ras kulit hitam, dirawat di rumah sakit SUNY. Ia | Seorang pasien dengan Patient_88ED, perempuan berusia 80–99 tahun dari ras tidak disebutkan, dirawat di Hospital_4AE2. Ia masuk | pasien patienttoken perempuan berusia 80-99 ras dirawat hospitaltoken masuk minggu maret jam pasien memiliki riwayat |

| | | |
|----|--|--|
| | <p>masuk pada hari Minggu, 29 Maret 2020 pukul 21:28. Pasien ini memiliki riwayat hipertensi dan asma. Penyakit lain seperti hiperlipidemia, diabetes, penyakit arteri koroner, gagal jantung, penyakit serebrovaskular, hepatitis, gagal ginjal tahap akhir, penyakit ginjal kronis, penyakit paru obstruktif kronis, demensia, dan kanker tidak tercatat. Status pasien berubah pada hari Selasa, 31 Maret 2020 pukul 21:50, dengan outcome meninggal dunia.</p> | <p>pada hari Minggu, Maret 2020 pukul sekitar jam tersebut. Pasien ini memiliki riwayat tekanan darah tinggi dan asma. Penyakit lain seperti hiperlipidemia, diabetes mellitus, penyakit arteri koroner, gagal jantung, penyakit serebrovaskular, hepatitis, gagal ginjal tahap akhir, penyakit ginjal kronis, penyakit paru obstruktif kronis, demensia ringan, kanker tercatat status pasien berubah selasa maret jam outcome deceased</p> |
| 5. | <p>Seorang pasien dengan ID Pasien 5, perempuan berusia 83 tahun dari ras kulit putih, dirawat di rumah sakit SUNY. Ia masuk pada hari Senin, 06 April 2020 pukul 23:00. Pasien ini memiliki riwayat hipertensi, diabetes, dan penyakit arteri koroner. Penyakit lain seperti hiperlipidemia, gagal</p> | <p>Seorang pasien dengan Patient_56D5, perempuan berusia 80–99 tahun dari ras tidak disebutkan, dirawat di Hospital_4AE2. Ia masuk pada hari Senin, April 2020 pukul sekitar jam tersebut. Pasien ini memiliki riwayat tekanan darah tinggi, kadar gula tinggi, dan penyakit arteri koroner. Penyakit lain seperti hiperlipidemia, gagal jantung, penyakit serebrovaskular, hepatitis, gagal ginjal tahap penyakit ginjal kronis</p> |

| | | | |
|----|--|---|--|
| | jantung, penyakit serebrovaskular, hepatitis, gagal ginjal tahap akhir, penyakit ginjal kronis, asma, penyakit paru obstruktif kronis, demensia, dan kanker tidak tercatat. Status pasien berubah pada hari Rabu, 08 April 2020 pukul 18:17, dengan outcome meninggal dunia. | jantung, penyakit serebrovaskular, hepatitis, gagal ginjal tahap akhir, penyakit ginjal kronis, asma, penyakit paru obstruktif kronis, demensia ringan, dan kanker tidak tercatat. Status pasien berubah pada hari Rabu, April 2020 pukul sekitar jam tersebut, dengan outcome deceased. | asma penyakit paru obstruktif kronis demensia ringan kanker tercatat status pasien berubah rabu april jam outcome deceased |
| 6. | Pasien ID 6, seorang perempuan kulit hitam berusia 70 tahun dari kelompok Expired. Pasien ini memiliki hiperlipidemia dan penyakit arteri koroner. Ia tidak memiliki hipertensi, diabetes, gagal jantung, penyakit serebrovaskular, hepatitis, gagal ginjal tahap akhir, penyakit ginjal kronis, asma, penyakit paru obstruktif kronis, demensia, atau kanker. Ia dirawat di rumah sakit SUNY pada hari Minggu, 05 April 2020 pukul 19:30. | Patient_68BF, seorang perempuan kulit hitam berusia 60–79 tahun dari kelompok Expired. Pasien ini memiliki hiperlipidemia dan penyakit arteri koroner. Ia tidak memiliki hipertensi ringan, diabetes mellitus, gagal jantung, penyakit serebrovaskular, hepatitis, gagal ginjal tahap akhir, penyakit ginjal kronis, asma, penyakit paru obstruktif kronis, gangguan kognitif, atau kanker. Ia dirawat di Hospital_C0E9, April 2020 pukul sekitar jam tersebut. Perubahan status pasien | patienttoken perempuan kulit hitam berusia 60-79 kelompok expired pasien memiliki hiperlipidemia penyakit arteri koroner memiliki hipertensi ringan diabetes mellitus gagal jantung penyakit serebrovaskular hepatitis gagal ginjal tahap penyakit ginjal kronis asma penyakit paru obstruktif kronis gangguan kognitif kanker dirawat hospitaltoken april jam perubahan status pasien sabtu april jam perawatan outcome |

| | | | |
|----|---|--|---|
| | Perubahan status pasien terjadi pada hari Sabtu, 11 April 2020 pukul 02:24, setelah perawatan selama 52.875 hari. Outcome pasien adalah ya. | terjadi pada hari Sabtu, April 2020 pukul sekitar jam tersebut, setelah perawatan selama 52.875 hari. Outcome pasien adalah ya. | pasiennya |
| 7. | Pasien ID 7, seorang perempuan kulit hitam berusia 64 tahun dari kelompok Expired. Pasien ini memiliki hipertensi dan penyakit serebrovaskular. Ia tidak memiliki hiperlipidemia, diabetes, penyakit arteri koroner, gagal jantung, hepatitis, gagal ginjal tahap akhir, penyakit ginjal kronis, asma, penyakit paru obstruktif kronis, demensia, atau kanker. Ia dirawat di rumah sakit SUNY pada hari Minggu, 05 April 2020 pukul 22:32. Perubahan status pasien terjadi pada hari Selasa, 07 April 2020 pukul 01:50, setelah perawatan selama 11.375 hari. Outcome pasien adalah ya. | Patient_B0AE, seorang perempuan kulit hitam berusia 60–79 tahun dari kelompok Expired. Pasien ini memiliki hipertensi ringan dan penyakit serebrovaskular. Ia tidak memiliki hiperlipidemia, penyakit arteri koroner, gagal jantung, hepatitis, gagal ginjal tahap akhir, penyakit ginjal kronis, asma, penyakit paru obstruktif kronis, demensia, atau kanker. Ia dirawat di Hospital_C0E9, April 2020 pukul sekitar jam tersebut. Perubahan status pasien terjadi pada hari Selasa, April 2020 pukul sekitar jam tersebut, setelah perawatan selama 11.375 hari. Outcome pasien adalah ya. | patienttoken perempuan kulit hitam berusia 60-79 kelompok expired pasien memiliki hipertensi ringan penyakit serebrovaskular memiliki hiperlipidemia kadar gula penyakit arteri koroner gagal jantung hepatitis gagal ginjal tahap penyakit ginjal kronis asma penyakit paru obstruktif kronis gangguan kognitif kanker dirawat hospitaltoken april jam perubahan status pasien selasa april jam perawatan outcome pasiennya. |

Tabel 1 memperlihatkan hasil transformasi pada dataset *Demographics* Data tabular berhasil diubah menjadi teks naratif yang deskriptif melalui proses serialisasi. Setiap narasi menggambarkan kondisi pasien secara natural dan informatif. Proses anonimisasi kemudian menghapus informasi identitas pribadi dan menggantinya dengan kode pseudonim, seperti [P001] untuk pasien dan [RS-A] untuk rumah sakit. Tahap preprocessing menghasilkan teks yang lebih bersih dan seragam, siap untuk digunakan dalam tahap analisis fitur.

Tabel 4.6 Hasil Transformasi Dataset ClinicalData

| No | Data Serialisasi | Data Anonimisasi | Data PreProcessing |
|----|--|--|--|
| 1. | Pasien dengan nomor identifikasi 5955 adalah seorang pria berusia 77 tahun dengan tinggi 160.02 cm, berat 67.05 kg, dan BMI 26.03. Status fisik ASA pasien adalah 2. Ia menjalani prosedur non-darurat bedah umum untuk kanker rektal, yaitu reseksi anterior rendah dengan pendekatan terbuka dan posisi litotomi. Anestesi yang digunakan adalah anestesi umum. Riwayat medis pasien mencakup hipertensi pre-operasi, namun tidak ada diabetes mellitus. Hasil EKG pre-operasi menunjukkan irama sinus | Pasien dengan Patient_CE17 adalah seorang pria berusia 60–79 tahun dengan tinggi <angka_tersembunyi> cm, berat <angka_tersembunyi> kg, dan BMI <angka_tersembunyi>. Status fisik ASA pasien adalah 2. Ia menjalani prosedur non-darurat bedah umum untuk tumor rektum, yaitu reseksi anterior rendah dengan pendekatan terbuka dan posisi litotomi. Anestesi yang digunakan adalah anestesi umum. Riwayat medis pasien mencakup tekanan darah tinggi pre-operasi, namun tidak ada kadar gula tinggi. Hasil EKG pre-operasi menunjukkan | pasien patienttoken pria berusia 60-79 numtoken berat numtoken bmi numtoken status fisik asa pasien menjalani prosedur non-darurat bedah tumor rektum reseksi anterior rendah pendekatan terbuka posisi litotomi anestesi anestesi riwayat medis pasien mencakup tekanan darah pre-operasi kadar gula hasil ekg pre-operasi irama sinus normal tes fungsi paru normal pemeriksaan laboratorium pre-operasi hemoglobin numtoken platelet numtoken pt numtoken aptt numtoken natrium numtoken kalium |

| | | | |
|--|---|---|--|
| | <p>normal, dan tes fungsi paru juga normal. Pemeriksaan laboratorium pre-operasi menunjukkan hemoglobin 14.01, platelet 189.0, PT 94.0, aPTT 33.02, natrium 141.0, kalium 3.01, glukosa 134.0, albumin 4.03, AST 18.0, ALT 16.0, BUN 10.0, dan kreatinin 0.057. Penilaian jalan napas Cormack I dengan jalan napas oral dan ukuran tabung endotrakeal 7.05. Akses intravena diperoleh di lengan bawah kanan, dan akses arteri radial kiri juga terpasang. Selama operasi, perkiraan kehilangan darah tidak tersedia, dan volume urine yang dikeluarkan adalah 300 mL. Pasien tidak menerima transfusi sel darah merah maupun FFP. Cairan kristaloid yang diberikan sebanyak 350 mL. Obat-obatan yang digunakan meliputi propofol 120, fentanil 100,</p> | <p>irama sinus normal, dan tes fungsi paru juga normal. Pemeriksaan laboratorium pre-operasi menunjukkan hemoglobin <angka_tersembunyi>, platelet <angka_tersembunyi>, PT <angka_tersembunyi>, aPTT <angka_tersembunyi>, natrium <angka_tersembunyi>, kalium <angka_tersembunyi>, glukosa <angka_tersembunyi>, albumin <angka_tersembunyi>, AST <angka_tersembunyi>, ALT <angka_tersembunyi>, BUN <angka_tersembunyi>, dan kreatinin <angka_tersembunyi>. Penilaian jalan napas Cormack I dengan jalan napas oral dan ukuran tabung endotrakeal</p> | <p>numtoken numtoken numtoken ast numtoken alt numtoken bun numtoken kreatinin numtoken penilaian jalan napas cormack jalan napas oral ukuran tabung endotrakeal numtoken akses intravena diperoleh lengan kanan akses arteri radial kiri terpasang operasi perkiraan kehilangan darah tersedia volume urine dikeluarkan numtoken pasien menerima transfusi sel darah merah ffp cairan kristaloid numtoken obat-obatan meliputi propofol fentanil rocuronium efedrin pasien menghabiskan icu selamat rawat inap durasi total numtoken durasi anestesi numtoken durasi operasi numtoken</p> |
|--|---|---|--|

| | | | |
|----|---|---|---|
| | <p>rocuronium 70, dan efedrin 10. Pasien tidak menghabiskan hari di ICU dan tidak meninggal selama rawat inap. Durasi total kasus adalah 11542 menit, durasi anestesi 11400 menit, dan durasi operasi 8700 menit.</p> | <p>akses arteri radial kiri juga terpasang. Selama operasi, perkiraan kehilangan darah tidak tersedia, dan volume urine yang dikeluarkan adalah <angka_tersembunyi> mL. Pasien tidak menerima transfusi sel darah merah maupun FFP. Cairan kristaloid yang diberikan sebanyak <angka_tersembunyi> mL. Obat-obatan yang digunakan meliputi propofol 120, fentanil 100, rocuronium 70, dan efedrin 10. Pasien tidak menghabiskan hari di ICU dan tidak selamat selama rawat inap. Durasi total kasus adalah <angka_tersembunyi> menit, durasi anestesi <angka_tersembunyi> menit, dan durasi operasi <angka_tersembunyi> menit.</p> | |
| 2. | <p>Pasien dengan nomor identifikasi 2487 adalah seorang pria berusia 54 tahun dengan tinggi 167.03</p> | <p>Pasien dengan Patient_7056 adalah seorang pria berusia 40–59 tahun dengan tinggi <angka_tersembunyi> cm,</p> | <p>pasien patienttoken pria berusia 40-59 numtoken berat numtoken bmi numtoken status fisik asa</p> |

| | | | |
|--|---|---|--|
| | <p>cm, berat 54.08 kg, dan BMI 19.06. Status fisik ASA pasien adalah 2. Ia menjalani prosedur non-darurat bedah umum untuk kanker lambung stadium lanjut, yaitu gastrektomi subtotal dengan pendekatan terbuka dan posisi terlentang. Anestesi yang digunakan adalah anestesi umum. Pasien tidak memiliki riwayat hipertensi atau diabetes mellitus pre-operasi. Hasil EKG pre-operasi menunjukkan irama sinus normal, dan tes fungsi paru juga normal. Pemeriksaan laboratorium pre-operasi menunjukkan hemoglobin 10.02, platelet 251.0, PT 110.0, aPTT 31.09, natrium 143.0, kalium 4.07, glukosa 88.0, albumin 3.08, AST 18.0, ALT 15.0, BUN 14.0, dan kreatinin 0.060. Penilaian jalan napas Cormack I dengan jalan napas oral</p> | <p>berat <angka_tersembunyi> kg, dan BMI <angka_tersembunyi>. Status fisik ASA pasien adalah 2. Ia menjalani prosedur non-darurat bedah umum untuk tumor lambung stadium lanjut, yaitu gastrektomi subtotal dengan pendekatan terbuka dan posisi terlentang. Anestesi yang digunakan adalah anestesi umum. Pasien tidak memiliki riwayat tekanan darah tinggi atau kadar gula tinggi pre-operasi. Hasil EKG pre-operasi menunjukkan irama sinus normal, dan tes fungsi paru juga normal. Pemeriksaan laboratorium pre-operasi menunjukkan hemoglobin <angka_tersembunyi>, platelet <angka_tersembunyi>, PT <angka_tersembunyi>, aPTT <angka_tersembunyi>, natrium <angka_tersembunyi>, kalium</p> | <p>pasien menjalani prosedur non-darurat bedah tumor lambung stadium gastrektomi subtotal pendekatan terbuka posisi terlentang anestesi anestesi pasien memiliki riwayat tekanan darah kadar gula pre-operasi hasil ekg pre-operasi irama sinus normal tes fungsi paru normal pemeriksaan laboratorium pre-operasi hemoglobin numtoken platelet numtoken pt numtoken aptt numtoken natrium numtoken kalium numtoken glukosa numtoken albumin numtoken ast numtoken alt numtoken bun numtoken kreatinin numtoken penilaian jalan napas cormack jalan napas oral ukuran tabung endotrakeal numtoken akses intravena diperoleh lengan kiri operasi perkiraan kehilangan</p> |
|--|---|---|--|

| | | | |
|--|---|--|--|
| | <p>dan ukuran tabung endotrakeal 7.05. Akses intravena diperoleh di lengan bawah kiri. Selama operasi, perkiraan kehilangan darah adalah 50 mL, dan volume urine yang dikeluarkan adalah 700 mL. Pasien tidak menerima transfusi sel darah merah maupun FFP. Cairan kristaloid yang diberikan sebanyak 800 mL. Obat-obatan yang digunakan meliputi propofol 150, rocuronium 100, dan efedrin 20. Pasien tidak menghabiskan hari di ICU dan tidak meninggal selama rawat inap. Durasi total kasus adalah 15741 menit, durasi anestesi 15960 menit, dan durasi operasi 12900 menit.</p> | <p><angka_tersembunyi>, glukosa <angka_tersembunyi>, albumin <angka_tersembunyi>, AST <angka_tersembunyi>, ALT <angka_tersembunyi>, BUN <angka_tersembunyi>, dan kreatinin <angka_tersembunyi>. Penilaian jalan napas Cormack I dengan jalan napas oral dan ukuran tabung endotrakeal <angka_tersembunyi>. Akses intravena diperoleh di lengan bawah kiri. Selama operasi, perkiraan kehilangan darah adalah <angka_tersembunyi> mL, dan volume urine yang dikeluarkan adalah <angka_tersembunyi> mL. Pasien tidak menerima transfusi sel darah merah maupun FFP. Cairan kristaloid yang diberikan sebanyak <angka_tersembunyi> mL. Obat-obatan yang digunakan</p> | <p>darah numtoken volume urine dikeluarkan numtoken pasien menerima transfusi sel darah merah ffp cairan kristaloid numtoken obat-obatan meliputi propofol rocuronium efedrin pasien menghabiskan icu deceased rawat inap durasi total numtoken durasi anestesi numtoken durasi operasi numtoken</p> |
|--|---|--|--|

| | | | |
|----|--|--|---|
| | | <p>meliputi propofol 150, rocuronium 100, dan efedrin 20. Pasien tidak menghabiskan hari di ICU dan tidak deceased selama rawat inap. Durasi total kasus adalah <angka_tersembunyi> menit, durasi anestesi <angka_tersembunyi> menit, dan durasi operasi <angka_tersembunyi> menit.</p> | |
| 3. | <p>Pasien dengan nomor identifikasi 2861 adalah seorang pria berusia 62 tahun dengan tinggi 169.01 cm, berat 69.07 kg, dan BMI 24.04. Status fisik ASA pasien adalah 1. Ia menjalani prosedur non-darurat bedah umum untuk batu empedu, yaitu kolesistektomi dengan pendekatan videoskopik dan posisi Trendelenburg terbalik. Anestesi yang digunakan adalah anestesi umum. Pasien tidak memiliki riwayat hipertensi atau diabetes</p> | <p>Pasien dengan Patient_8AC4 adalah seorang pria berusia 60–79 tahun dengan tinggi <angka_tersembunyi> cm, berat <angka_tersembunyi> kg, dan BMI <angka_tersembunyi>. Status fisik ASA pasien adalah 1. Ia menjalani prosedur non-darurat bedah umum untuk kelainan empedu, yaitu kolesistektomi dengan pendekatan videoskopik dan posisi Trendelenburg terbalik. Anestesi yang digunakan adalah anestesi umum. Pasien tidak memiliki</p> | <p>pasien patienttoken pria berusia 60-79 numtoken berat numtoken bmi numtoken status fisik asa pasien menjalani prosedur non-darurat bedah kelainan empedu kolesistektomi pendekatan videoskopik posisi trendelenburg terbalik anestesi anestesi pasien memiliki riwayat hipertensi ringan kadar gula pre-operasi hasil ekg pre-operasi irama sinus normal tes fungsi paru normal pemeriksaan laboratorium pre-operasi</p> |

| | | | |
|--|---|--|---|
| | <p>mellitus pre-operasi. Hasil EKG pre-operasi menunjukkan irama sinus normal, dan tes fungsi paru juga normal. Pemeriksaan laboratorium pre-operasi menunjukkan hemoglobin 14.02, platelet 373.0, PT 103.0, aPTT 30.03, natrium 144.0, kalium 4.09, glukosa 87.0, albumin 4.02, AST 17.0, ALT 34.0, BUN 14.0, dan kreatinin 1.18. Penilaian jalan napas Cormack I dengan jalan napas oral dan ukuran tabung endotrakeal 7.05. Akses intravena diperoleh di lengan bawah kiri. Selama operasi, data perkiraan kehilangan darah dan volume urine tidak tersedia. Pasien tidak menerima transfusi sel darah merah maupun FFP. Cairan kristaloid yang diberikan sebanyak 200 mL. Obat-obatan yang digunakan meliputi</p> | <p>riwayat hipertensi ringan atau kadar gula tinggi pre-operasi. Hasil EKG pre-operasi menunjukkan irama sinus normal, dan tes fungsi paru juga normal. Pemeriksaan laboratorium pre-operasi menunjukkan hemoglobin <angka_tersembunyi>, platelet <angka_tersembunyi>, PT <angka_tersembunyi>, aPTT <angka_tersembunyi>, natrium <angka_tersembunyi>, kalium <angka_tersembunyi>, glukosa <angka_tersembunyi>, albumin <angka_tersembunyi>, AST <angka_tersembunyi>, ALT <angka_tersembunyi>, BUN <angka_tersembunyi>, dan kreatinin <angka_tersembunyi>. Penilaian jalan napas Cormack I dengan jalan napas oral dan ukuran tabung endotrakeal 7.05. Akses intravena diperoleh di lengan bawah kiri. Selama operasi, data perkiraan kehilangan darah dan volume urine tidak tersedia. Pasien tidak menerima transfusi sel darah merah maupun FFP. Cairan kristaloid yang diberikan sebanyak 200 mL. Obat-obatan yang digunakan meliputi</p> | <p>hemoglobin numtoken platelet numtoken pt numtoken aptt numtoken natrium numtoken kalium numtoken glukosa numtoken albumin numtoken ast numtoken alt numtoken bun numtoken kreatinin numtoken penilaian jalan napas cormack jalan napas oral ukuran tabung endotrakeal numtoken akses intravena diperoleh lengan kiri operasi data perkiraan kehilangan darah volume urine tersedia pasien menerima transfusi sel darah merah ffp cairan kristaloid numtoken obat-obatan meliputi rocuronium pasien menghabiskan icu deceased rawat inap durasi total numtoken durasi anestesi numtoken durasi operasi numtoken</p> |
|--|---|--|---|

| | | | |
|----|---|---|---|
| | <p>rocuronium 50. Pasien tidak menghabiskan hari di ICU dan tidak meninggal selama rawat inap. Durasi total kasus adalah 4394 menit, durasi anestesi 4800 menit, dan durasi operasi 1920 menit.</p> | <p>endotrakeal <angka_tersembunyi>. Akses intravena diperoleh di lengan bawah kiri. Selama operasi, data perkiraan kehilangan darah dan volume urine tidak tersedia. Pasien tidak menerima transfusi sel darah merah maupun FFP. Cairan kristaloid yang diberikan sebanyak <angka_tersembunyi> mL. Obat-obatan yang digunakan meliputi rocuronium 50. Pasien tidak menghabiskan hari di ICU dan tidak deceased selama rawat inap. Durasi total kasus adalah <angka_tersembunyi> menit, durasi anestesi <angka_tersembunyi> menit, dan durasi operasi <angka_tersembunyi> menit.</p> | |
| 4. | <p>Pasien dengan nomor identifikasi 1903 adalah seorang pria berusia 74 tahun dengan tinggi 160.06 cm, berat 53 kg, dan BMI 20.05. Status fisik ASA pasien adalah 2. Ia</p> | <p>Pasien dengan Patient_687D adalah seorang pria berusia 60–79 tahun dengan tinggi <angka_tersembunyi> cm, berat <angka_tersembunyi> kg, dan BMI <angka_tersembunyi>.</p> | <p>pasien patienttoken pria berusia 60-79 numtoken berat numtoken bmi numtoken status fisik asa pasien menjalani prosedur non-darurat bedah kanker saluran pencernaan</p> |

| | | | |
|--|---|--|---|
| | <p>menjalani prosedur non-darurat bedah umum untuk kanker lambung stadium lanjut, yaitu gastrektomi distal dengan pendekatan videoskopik dan posisi Trendelenburg terbalik. Anestesi yang digunakan adalah anestesi umum. Riwayat medis pasien mencakup hipertensi pre-operasi, namun tidak ada diabetes mellitus. Hasil EKG pre-operasi menunjukkan irama sinus normal, dan tes fungsi paru juga normal. Pemeriksaan laboratorium pre-operasi menunjukkan hemoglobin 14.04, platelet 275.0, PT 103.0, aPTT 34.05, natrium 141.0, kalium 4.02, glukosa 108.0, albumin 4.01, AST 23.0, ALT 18.0, BUN 10.0, dan kreatinin 0.067. Penilaian jalan napas Cormack I dengan jalan napas oral, namun ukuran tabung endotrakeal tidak</p> | <p>Status fisik ASA pasien adalah 2. Ia menjalani prosedur non-darurat bedah umum untuk kanker saluran pencernaan atas stadium lanjut, yaitu gastrektomi distal dengan pendekatan videoskopik dan posisi Trendelenburg terbalik. Anestesi yang digunakan adalah anestesi umum. Riwayat medis pasien mencakup tekanan darah tinggi pre-operasi, namun tidak ada diabetes ringan. Hasil EKG pre-operasi menunjukkan irama sinus normal, dan tes fungsi paru juga normal. Pemeriksaan laboratorium pre-operasi menunjukkan hemoglobin <angka_tersembunyi>, platelet <angka_tersembunyi>, PT <angka_tersembunyi>, aPTT <angka_tersembunyi>, natrium <angka_tersembunyi>, kalium <angka_tersembunyi>,</p> | <p>stadium gastrektomi distal pendekatan videoskopik posisi trendelenburg terbalik anestesi anestesi riwayat medis pasien mencakup tekanan darah pre-operasi diabetes ringan hasil ekg pre-operasi irama sinus normal tes fungsi paru normal pemeriksaan laboratorium pre-operasi hemoglobin numtoken platelet numtoken pt numtoken aptt numtoken natrium numtoken kalium numtoken glukosa numtoken albumin numtoken ast numtoken alt numtoken bun numtoken kreatinin numtoken penilaian jalan napas cormack jalan napas oral ukuran tabung endotrakeal tercatat akses intravena diperoleh lengan kiri akses arteri radial kanan terpasang operasi perkiraan kehilangan darah tersedia volume</p> |
|--|---|--|---|

| | | | |
|--|---|--|--|
| | <p>tercatat. Akses intravena diperoleh di lengan bawah kiri, dan akses arteri radial kanan juga terpasang. Selama operasi, perkiraan kehilangan darah tidak tersedia, dan volume urine yang dikeluarkan adalah 270 mL. Pasien tidak menerima transfusi sel darah merah maupun FFP. Cairan kristaloid yang diberikan sebanyak 2700 mL. Obat-obatan yang digunakan meliputi propofol 80, fentanil 100, rocuronium 100, dan efedrin 50. Pasien menghabiskan 1 hari di ICU dan tidak meninggal selama rawat inap. Durasi total kasus adalah 20990 menit, durasi anestesi 21000 menit, dan durasi operasi 15300 menit.</p> | <p>glukosa <angka_tersembunyi>, albumin <angka_tersembunyi>, AST <angka_tersembunyi>, ALT <angka_tersembunyi>, BUN <angka_tersembunyi>, dan kreatinin <angka_tersembunyi>. Penilaian jalan napas Cormack I dengan jalan napas oral, namun ukuran tabung endotrakeal tidak tercatat. Akses intravena diperoleh di lengan bawah kiri, dan akses arteri radial kanan juga terpasang. Selama operasi, perkiraan kehilangan darah tidak tersedia, dan volume urine yang dikeluarkan adalah <angka_tersembunyi> mL. Pasien tidak menerima transfusi sel darah merah maupun FFP. Cairan kristaloid yang diberikan sebanyak <angka_tersembunyi> mL. Obat-obatan yang digunakan meliputi propofol 80,</p> | <p>urine numtoken pasien menerima transfusi sel darah merah ffp cairan kristaloid numtoken obat-obatan meliputi propofol fentanil rocuronium efedrin pasien menghabiskan icu selamat rawat inap durasi total numtoken durasi anestesi numtoken durasi operasi numtoken</p> |
|--|---|--|--|

| | | | |
|----|---|--|--|
| | | fentanil 100, rocuronium 100, dan efedrin 50. Pasien menghabiskan 1 hari di ICU dan tidak selamat selama rawat inap. Durasi total kasus adalah <angka_tersembunyi> menit, durasi anestesi <angka_tersembunyi> menit, dan durasi operasi <angka_tersembunyi> menit. | |
| 5. | Pasien dengan nomor identifikasi 4416 adalah seorang pria berusia 66 tahun dengan tinggi 171 cm, berat 59.07 kg, dan BMI 20.04. Status fisik ASA pasien adalah 3. Ia menjalani prosedur darurat bedah umum untuk aneurisma aorta, yaitu perbaikan aneurisma dengan pendekatan terbuka dan posisi tengkurap. Anestesi yang digunakan adalah anestesi umum. Riwayat medis pasien mencakup hipertensi pre-operasi. Hasil EKG pre-operasi menunjukkan | Pasien dengan Patient_2084 adalah seorang pria berusia 60–79 tahun dengan tinggi <angka_tersembunyi> cm, berat <angka_tersembunyi> kg, dan BMI <angka_tersembunyi>. Status fisik ASA pasien adalah 3. Ia menjalani prosedur darurat bedah umum untuk aneurisma aorta, yaitu perbaikan aneurisma dengan pendekatan terbuka dan posisi tengkurap. Anestesi yang digunakan adalah anestesi umum. Riwayat medis pasien mencakup hipertensi ringan pre-operasi. | pasien patienttoken pria berusia 60-79 numtoken berat numtoken bmi numtoken status fisik asa pasien menjalani prosedur darurat bedah aneurisma aorta perbaikan aneurisma pendekatan terbuka posisi tengkurap anestesi anestesi riwayat medis pasien mencakup hipertensi ringan pre-operasi hasil ekg pre-operasi blok fasikular anterior kiri tes fungsi paru normal pemeriksaan laboratorium pre-operasi hemoglobin numtoken platelet numtoken pt |

| | | |
|--|---|--|
| | <p>blok fasikular anterior kiri, dan tes fungsi paru normal. Pemeriksaan laboratorium pre-operasi menunjukkan hemoglobin 10.01, platelet 67.0, PT 73.0, aPTT 36.05, natrium 146.0, kalium 4.04, glukosa 126.0, albumin 2.06, AST 765.0, ALT 77.0, BUN 50.0, dan kreatinin 4.43. Penilaian jalan napas Cormack I dengan jalan napas oral dan ukuran tabung endotrakeal 7.05. Akses intravena diperoleh di lengan bawah kanan, akses arteri radial kanan, dan akses vena sentral di vena jugularis interna kanan juga terpasang. Selama operasi, perkiraan kehilangan darah adalah 2600 mL, dan volume urine yang dikeluarkan adalah 1490 mL. Pasien menerima transfusi 8 unit sel darah merah dan 8 unit FFP. Cairan kristaloid yang diberikan sebanyak 7100</p> | <p>Hasil EKG pre-operasi menunjukkan blok fasikular anterior kiri, dan tes fungsi paru normal. Pemeriksaan laboratorium pre-operasi menunjukkan hemoglobin <angka_tersembunyi>, platelet <angka_tersembunyi>, PT <angka_tersembunyi>, aPTT <angka_tersembunyi>, natrium <angka_tersembunyi>, kalium <angka_tersembunyi>, glukosa <angka_tersembunyi>, albumin <angka_tersembunyi>, AST <angka_tersembunyi>, ALT <angka_tersembunyi>, BUN <angka_tersembunyi>, dan kreatinin <angka_tersembunyi>. Penilaian jalan napas Cormack I dengan jalan napas oral dan ukuran tabung endotrakeal 7.05. Akses intravena diperoleh di lengan kanan akses arteri radial kanan akses vena sentral vena jugularis interna kanan terpasang operasi perkiraan kehilangan darah numtoken volume urine dikeluarkan numtoken pasien menerima transfusi unit sel darah merah unit ffp cairan kristaloid numtoken obat-obatan meliputi rocuronium efedrin fenilefrin kalsium pasien menghabiskan icu deceased rawat inap durasi total numtoken durasi anestesi numtoken durasi</p> <p>Akses intravena diperoleh di</p> |
|--|---|--|

| | | | |
|----|--|---|---|
| | <p>mL. Obat-obatan yang digunakan meliputi rocuronium 160, efedrin 10, fenilefrin 900, dan kalsium 2100. Pasien menghabiskan 13 hari di ICU dan tidak meninggal selama rawat inap. Durasi total kasus adalah 21531 menit, durasi anestesi 23400 menit, dan durasi operasi 17700 menit.</p> | <p>lengan bawah kanan, akses arteri radial kanan, dan akses vena sentral di vena jugularis interna kanan juga terpasang. Selama operasi, perkiraan kehilangan darah adalah <angka_tersembunyi> mL, dan volume urine yang dikeluarkan adalah <angka_tersembunyi> mL. Pasien menerima transfusi 8 unit sel darah merah dan 8 unit FFP. Cairan kristaloid yang diberikan sebanyak <angka_tersembunyi> mL. Obat-obatan yang digunakan meliputi rocuronium 160, efedrin 10, fenilefrin 900, dan kalsium 2100. Pasien menghabiskan 13 hari di ICU dan tidak deceased selama rawat inap. Durasi total kasus adalah <angka_tersembunyi> menit, durasi anestesi <angka_tersembunyi> menit, dan durasi operasi <angka_tersembunyi> menit.</p> | <p>operasi numtoken</p> |
| 6. | <p>Seorang pasien wanita berusia 78 tahun dengan</p> | <p>Seorang pasien wanita berusia 60–79 tahun dengan</p> | <p>pasien wanita berusia 60-79 numtoken berat</p> |

| | | | |
|--|--|--|--|
| | <p>tinggi 150 cm dan berat 54.06 kg, memiliki BMI 24.03. Status fisik menurut ASA adalah 2.0. Pasien ini menjalani operasi darurat di departemen bedah umum untuk kolesistektomi videoscopik dengan posisi telentang, akibat diagnosis polip kandung empedu. Anestesi umum diberikan. Durasi total rawat inap adalah 432000 menit, durasi observasi kasus adalah 5145 menit, durasi anestesi adalah 4800 menit, dan durasi operasi adalah 1800 menit. Pasien tidak memerlukan perawatan di ICU dan tidak meninggal selama rawat inap. Kondisi pra-operasi pasien menunjukkan tidak ada riwayat hipertensi atau diabetes mellitus. Hasil EKG dan fungsi paru (PFT) normal. Hasil laboratorium pra-operasi meliputi hemoglobin 12.03</p> | <p>tinggi <angka_tersembunyi> cm dan berat <angka_tersembunyi> kg, memiliki BMI <angka_tersembunyi>. Status fisik menurut ASA adalah <angka_tersembunyi>. Pasien ini menjalani operasi darurat di departemen bedah umum untuk kolesistektomi videoscopik dengan posisi telentang, akibat diagnosis polip kandung empedu. Anestesi umum diberikan. Durasi total rawat inap adalah <angka_tersembunyi> menit, durasi observasi kasus adalah <angka_tersembunyi> menit, durasi anestesi adalah <angka_tersembunyi> menit, dan durasi operasi adalah <angka_tersembunyi> menit. Pasien tidak memerlukan perawatan di ICU dan tidak deceased selama rawat inap. Kondisi pra-operasi pasien menunjukkan tidak ada riwayat tekanan darah tinggi</p> | <p>numtoken memiliki bmi numtoken status fisik asa numtoken pasien menjalani operasi darurat departemen bedah kolesistektomi videoscopik posisi telentang akibat diagnosis polip kandung empedu anestesi durasi total rawat inap numtoken durasi observasi numtoken durasi anestesi numtoken durasi operasi numtoken pasien perawatan icu deceased rawat inap kondisi pra-operasi pasien riwayat tekanan darah kadar gula hasil ekg fungsi paru pft normal hasil laboratorium pra-operasi meliputi hemoglobin numtoken g dl platelet numtoken ul pt numtoken aptt numtoken detik natrium numtoken meq 1 kalium numtoken meq 1 glukosa numtoken mg dl albumin numtoken g dl ast numtoken u l alt numtoken u l bun</p> |
|--|--|--|--|

| | | | |
|--|--|--|--|
| | <p>g/dL, platelet 144.0 10³/uL, PT 104.0%, aPTT 29.01 detik, natrium 141.0 mEq/L, kalium 3.04 mEq/L, glukosa 105.0 mg/dL, albumin 3.09 g/dL, AST 19.0 U/L, ALT 17.0 U/L, BUN 14.0 mg/dL, dan kreatinin 1.07 mg/dL. Parameter gas darah (pH, HCO₃, BE, PaO₂, PaCO₂, SaO₂) tidak tersedia. Manajemen jalan napas dilakukan secara oral dengan pipa endotrakeal berukuran 7.0, dan skor Cormack I. Akses intravena diperoleh di lengan bawah kiri. Selama operasi, pasien menerima 100.0 mL cairan kristaloid tanpa transfusi sel darah merah, FFP, atau koloid. Obat-obatan yang diberikan meliputi Propofol 70 mg, Fentanyl 100 mcg, Rocuronium 40 mg, dan Ephedrine 5 mg. Midazolam, Vecuronium, Phenylephrine,</p> | <p>atau kadar gula tinggi. Hasil EKG dan fungsi paru (PFT) normal. Hasil laboratorium pra-operasi meliputi hemoglobin <angka_tersembunyi> g/dL, platelet <angka_tersembunyi> 10³/uL, PT <angka_tersembunyi>% aPTT <angka_tersembunyi> detik, natrium <angka_tersembunyi> mEq/L, kalium <angka_tersembunyi> mEq/L, glukosa <angka_tersembunyi> mg/dL, albumin <angka_tersembunyi> g/dL, AST <angka_tersembunyi> U/L, ALT <angka_tersembunyi> U/L, BUN <angka_tersembunyi> mg/dL, dan kreatinin <angka_tersembunyi> mg/dL. Parameter gas darah (pH, HCO₃, BE, PaO₂, PaCO₂, SaO₂) tidak tersedia. Manajemen jalan napas dilakukan secara oral dengan pipa endotrakeal berukuran 7.0, dan skor Cormack I. Akses intravena diperoleh di lengan bawah kiri. Selama operasi, pasien menerima 100.0 mL cairan kristaloid tanpa transfusi sel darah merah, FFP, atau koloid. Obat-obatan yang diberikan meliputi Propofol 70 mg, Fentanyl 100 mcg, Rocuronium 40 mg, dan Ephedrine 5 mg. Midazolam, Vecuronium, Phenylephrine,</p> | <p>numtoken mg dl kreatinin numtoken mg dl parameter gas darah ph hco3 pao2 paco2 sao2 tersedia manajemen jalan napas oral pipa endotrakeal berukuran numtoken skor cormack akses intravena diperoleh lengan kiri operasi pasien menerima numtoken cairan kristaloid transfusi sel darah merah ffp koloid obat-obatan meliputi propofol mg fentanyl mcg rocuronium mg ephedrine mg midazolam vecuronium phenylephrine epinephrine kalsium estimasi kehilangan darah output urin tersedia</p> |
|--|--|--|--|

| | | | |
|----|--|---|--|
| | Epinephrine, dan Kalsium tidak diberikan. Estimasi kehilangan darah dan output urin tidak tersedia. | pipa endotrakeal berukuran <angka_tersembunyi>, dan skor Cormack I. Akses intravena diperoleh di lengan bawah kiri. Selama operasi, pasien menerima <angka_tersembunyi> mL cairan kristaloid tanpa transfusi sel darah merah, FFP, atau koloid. Obat-obatan yang diberikan meliputi Propofol 70 mg, Fentanyl 100 mcg, Rocuronium 40 mg, dan Ephedrine 5 mg. Midazolam, Vecuronium, Phenylephrine, Epinephrine, dan Kalsium tidak diberikan. Estimasi kehilangan darah dan output urin tidak tersedia. | |
| 7. | Seorang pasien wanita berusia 52 tahun dengan tinggi 167.07 cm dan berat 62.03 kg, memiliki BMI 22.02. Status fisik menurut ASA adalah 2.0. Pasien ini menjalani operasi terencana (non-darurat) di departemen bedah toraks untuk lobektomi paru | Seorang pasien wanita berusia 40–59 tahun dengan tinggi <angka_tersembunyi> cm dan berat <angka_tersembunyi> kg, memiliki BMI <angka_tersembunyi>. Status fisik menurut ASA adalah <angka_tersembunyi>. | pasien wanita berusia 40-59 numtoken berat numtoken memiliki bmi numtoken status fisik asa numtoken pasien menjalani operasi terencana non-darurat departemen bedah toraks lobektomi paru videoscopik posisi |

| | | | |
|--|--|---|--|
| | <p>secara videoscopik dalam posisi dekubitus lateral kiri, karena infeksi mikobakteri nontuberkulosa. Anestesi umum diberikan. Durasi total rawat inap adalah 777600 menit, durasi observasi kasus adalah 15770 menit, durasi anestesi adalah 14340 menit, dan durasi operasi adalah 11400 menit. Pasien dirawat di ICU selama 3 hari dan tidak meninggal selama rawat inap. Kondisi pra-operasi pasien menunjukkan tidak ada riwayat hipertensi atau diabetes mellitus. Hasil EKG dan fungsi paru (PFT) normal. Hasil laboratorium pra-operasi tidak tersedia. Manajemen jalan napas dilakukan secara oral dengan pipa double-lumen ukuran Left-35, dan skor Cormack I. Akses intravena diperoleh di lengan bawah</p> | <p>Pasien ini menjalani operasi terencana (non-darurat) di departemen bedah toraks untuk lobektomi paru secara videoscopik dalam posisi dekubitus lateral kiri, karena infeksi mikobakteri nontuberkulosa. Anestesi umum diberikan. Durasi total rawat inap adalah <angka_tersembunyi> menit, durasi observasi kasus adalah <angka_tersembunyi> menit, durasi anestesi adalah <angka_tersembunyi> menit, dan durasi operasi adalah <angka_tersembunyi> menit. Pasien dirawat di ICU selama 3 hari dan tidak meninggal selama rawat inap. Kondisi pra-operasi pasien menunjukkan tidak ada riwayat tekanan darah tinggi atau diabetes ringan. Hasil EKG dan fungsi paru (PFT) normal. Hasil laboratorium pra-operasi tidak tersedia. Manajemen jalan napas dilakukan secara oral dengan pipa double-lumen ukuran</p> | <p>dekubitus lateral kiri infeksi mikobakteri nontuberkulosa anestesi durasi total rawat inap numtoken durasi observasi numtoken durasi anestesi numtoken durasi operasi numtoken pasien dirawat icu selamat rawat inap kondisi pra-operasi pasien riwayat tekanan darah diabetes ringan hasil ekg fungsi paru pft normal hasil laboratorium pra-operasi tersedia manajemen jalan napas oral pipa double-lumen ukuran left-35 skor cormack akses intravena diperoleh lengan kiri akses arteri radial kiri akses vena sentral i JV kanan operasi estimasi kehilangan darah numtoken output urin numtoken pasien menerima numtoken cairan kristaloid transfusi sel darah merah ffp koloid obat-obatan meliputi</p> |
|--|--|---|--|

| | | | |
|--|--|---|---|
| | <p>kiri, akses arteri di radial kiri, dan akses vena sentral di IJV kanan. Selama operasi, estimasi kehilangan darah adalah 100.0 mL dan output urin 125.0 mL. Pasien menerima 700.0 mL cairan kristaloid tanpa transfusi sel darah merah, FFP, atau koloid. Obat-obatan yang diberikan meliputi Rocuronium 120 mg. Propofol, Midazolam, Fentanyl, Vecuronium, Ephedrine, Phenylephrine, Epinephrine, dan Kalsium tidak diberikan.</p> | <p>Left-35, dan skor Cormack I. Akses intravena diperoleh di lengan bawah kiri, akses arteri di radial kiri, dan akses vena sentral di IJV kanan. Selama operasi, estimasi kehilangan darah adalah <angka_tersembunyi> mL dan output urin <angka_tersembunyi> mL. Pasien menerima <angka_tersembunyi> mL cairan kristaloid tanpa transfusi sel darah merah, FFP, atau koloid. Obat-obatan yang diberikan meliputi Rocuronium 120 mg. Propofol, Midazolam, Fentanyl, Vecuronium, Ephedrine, Phenylephrine, Epinephrine, dan Kalsium tidak diberikan.</p> | <p>rocuronium mg propofol midazolam fentanyl vecuronium ephedrine phenylephrine epinephrine kalsium</p> |
|--|--|---|---|

Tabel 2 menampilkan hasil transformasi pada dataset *ClinicalData* Tahap serialisasi berhasil mengonversi data medis menjadi narasi teks yang natural dan koheren. Hasil anonimisasi menunjukkan bahwa seluruh informasi yang dapat mengarah ke identitas pasien telah dihapus dan diganti dengan kode anonim. Setelah dilakukan preprocessing, teks menjadi lebih ringkas dan konsisten secara format, sehingga memudahkan untuk diproses pada tahap *feature extraction* seperti *TF-IDF* atau model berbasis *embedding*.

Berdasarkan hasil kedua tabel, dapat disimpulkan bahwa pipeline yang dikembangkan mampu

menghasilkan teks medis anonim yang tetap mempertahankan konteks semantik. Proses serialisasi dengan LLM Gemini 2.5 Flash terbukti efektif dalam mengonversi data tabular menjadi narasi natural tanpa kehilangan informasi penting. Sementara itu, tahap anonimisasi berhasil menjaga privasi pasien dengan mengganti semua entitas sensitif menggunakan *pseudonymization* dan *generalization*.

Tahap preprocessing memberikan hasil akhir berupa teks yang bersih dan seragam. Dengan demikian, pipeline yang dihasilkan telah memenuhi dua kriteria utama dalam pengelolaan data medis: *semantic preservation* (keutuhan makna) dan *privacy protection* (perlindungan privasi).

Tahap 5: Data Mining (25 Poin)

Pada tahap *Data Mining*, dilakukan pembangunan model klasifikasi untuk mengelompokkan data pasien ke dalam dua kategori risiko, yaitu HIGH RISK dan LOW RISK. Dua algoritma yang digunakan dalam eksperimen ini adalah *Naive Bayes* (NB) dan *Logistic Regression* (LR). Masing-masing model diuji dengan tiga pendekatan ekstraksi fitur, yaitu *Bag of Words* (BoW), *TF-IDF*, dan *N-Gram* (Bigram). Selain itu, dilakukan *Feature Selection* (Chi-Square) untuk memilih fitur paling relevan dan Benchmarking untuk membandingkan kinerja model antar dataset (*Demographics* dan *ClinicalData*) serta antar kondisi data *original* dan *anonymized*.

- **Algoritma Data Mining yang Digunakan**

1. ***Naive Bayes (NB)*** Merupakan algoritma probabilistik sederhana namun efektif untuk klasifikasi teks berdimensi tinggi. NB menghitung peluang suatu teks termasuk dalam kategori tertentu (HIGH_RISK atau LOW_RISK) berdasarkan distribusi kata pada data latih.
2. ***Logistic Regression (LR)*** Model ini menghitung probabilitas keanggotaan suatu sampel dalam kelas tertentu menggunakan fungsi sigmoid. LR sangat cocok untuk klasifikasi biner seperti pada kasus ini, serta mampu memberikan interpretasi yang jelas terhadap bobot fitur hasil ekstraksi teks.

- **Skenario Eksperimen**

1. **Preprocessing:** Data teks dibersihkan dengan menghapus karakter khusus, stopword, dan menstandarkan data seperti mengubah teks menjadi huruf kecil

serta mengganti angka atau rentang usia dengan token yang sesuai. Data juga ditokenisasi dan fitur medis digeneralisasi.

2. **Serialisasi:** Data tabular yang sudah diproses diubah menjadi narasi teks alami menggunakan Generative AI (Gemini 2.5). Narasi ini berfungsi untuk mempermudah analisis lanjutan dengan model berbasis NLP. Setiap pasien diubah menjadi teks yang mendeskripsikan informasi medisnya.
3. **Anonimisasi:** Teknik hybrid anonymization diterapkan untuk menjaga kerahasiaan informasi sensitif. Data yang telah diserialisasi kemudian dianonimkan dengan pseudonimisasi (untuk ID pasien dan nama rumah sakit), generalisasi (untuk atribut seperti umur dan lokasi), dan perturbasi (untuk informasi sensitif seperti diagnosis penyakit).
4. **Pembagian Data:** Data dibagi menjadi data latih (80%) dan data uji (20%), dengan stratifikasi untuk memastikan distribusi label HIGH_RISK dan LOW_RISK yang seimbang.
5. **Feature Extraction:** Fitur diekstraksi menggunakan BoW (Unigram), TF-IDF (Unigram), dan N-Gram (Bigram) untuk mendapatkan representasi teks yang relevan dengan pendekatan yang berbeda. Hasil vektorisasi menjadi masukan untuk model klasifikasi.
6. **Feature Selection:** Menggunakan metode **Chi-Square** untuk memilih fitur yang paling berkontribusi terhadap target label, sehingga mengurangi dimensi dan meningkatkan efisiensi pelatihan model.
7. **Evaluasi:** Model dievaluasi menggunakan confusion matrix dan classification report untuk menghitung precision, recall, dan f1-score guna mengukur performa model dalam mengklasifikasikan data uji.
8. **Benchmarking:** Setelah seluruh eksperimen dijalankan, hasil performa dibandingkan antara dua dataset (*Demographics* dan *ClinicalData*) serta antara versi *original* dan *anonymized*. Tahap ini bertujuan menilai sejauh mana anonimisasi memengaruhi akurasi model.

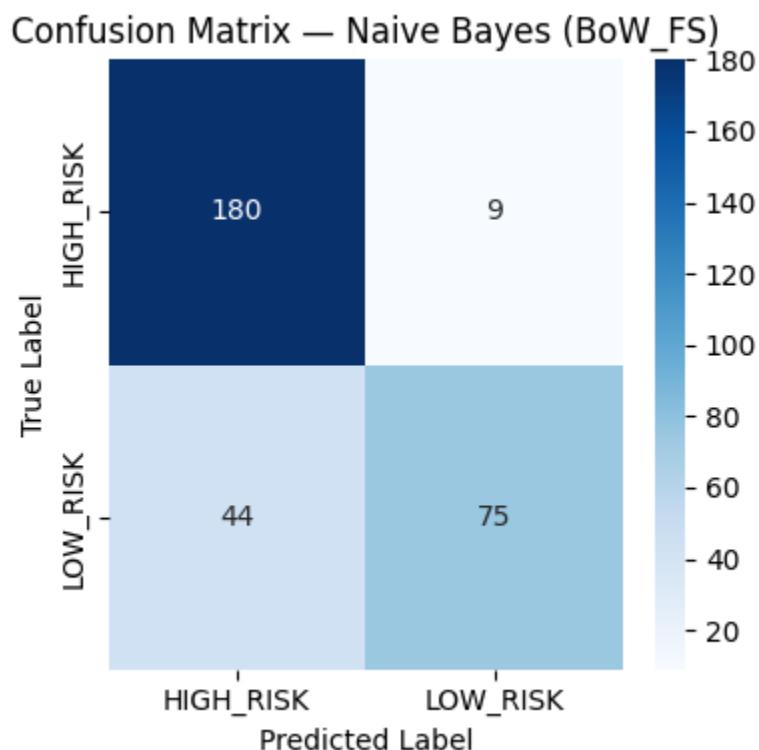
- **Text Representation**

Berikut adalah visualisasi hasil confusion matrix dari Dataset Demographic :

Tabel 5.1 Classification Report Model NB + BoW - Demographics Dataset

| | Precision | Recall | F1-Score | Support |
|-------------------|-----------|--------|----------|---------|
| HIGH_RISK | 0.8036 | 0.9524 | 0.8717 | 189 |
| LOW_RISK | 0.8929 | 0.6303 | 0.7389 | 119 |
| Accuracy | | | 0.8279 | 308 |
| Macro avg | 0.8482 | 0.7913 | 0.8053 | 308 |
| Weight avg | 0.8381 | 0.8279 | 0.8204 | 308 |

Model *Naive Bayes* dengan BoW mencapai akurasi **82.79%**, dengan performa yang kuat pada kelas HIGH_RISK (recall 0.95) tetapi lemah dalam mengenali LOW_RISK. Hasil ini menunjukkan bahwa *Naive Bayes* cenderung lebih sensitif terhadap distribusi kata yang sering muncul pada kelas dominan, dan kurang optimal dalam menangani variasi kontekstual pada kelas minoritas.

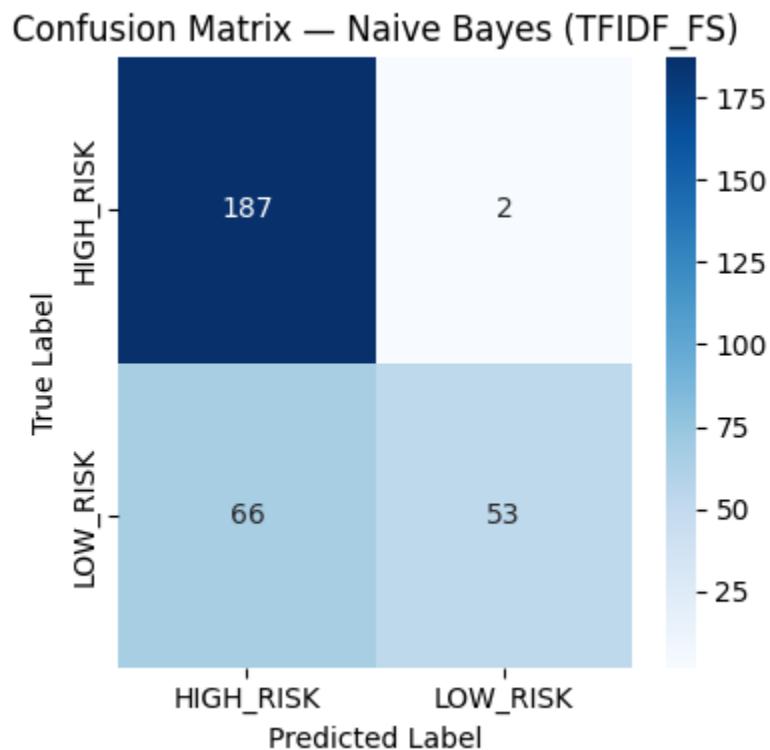


Gambar 5.1 Confusion matrix NB + BoW Pada Dataset Demographics

Tabel 5.2 Classification Report Model NB + TF-IDF - Demographics Dataset

| | Precision | Recall | F1-Score | Support |
|-------------------|-----------|--------|----------|---------|
| HIGH_RISK | 0.7391 | 0.9894 | 0.8462 | 189 |
| LOW_RISK | 0.9636 | 0.4454 | 0.6092 | 119 |
| Accuracy | | | 0.7792 | 308 |
| Macro avg | 0.8514 | 0.7174 | 0.7277 | 308 |
| Weight avg | 0.8259 | 0.7792 | 0.7546 | 308 |

Model *Naive Bayes* dengan TF-IDF menghasilkan akurasi **77.92%**, menunjukkan bahwa model ini kesulitan menyesuaikan diri dengan pembobotan kata berbasis frekuensi invers. Recall untuk HIGH_RISK sangat tinggi (0.99), namun model gagal mempertahankan keseimbangan antar kelas, dengan recall untuk LOW_RISK hanya 0.44. Ini menandakan bias kuat terhadap kelas mayoritas, yang umum terjadi pada *Naive Bayes* dengan distribusi kata tidak seragam.

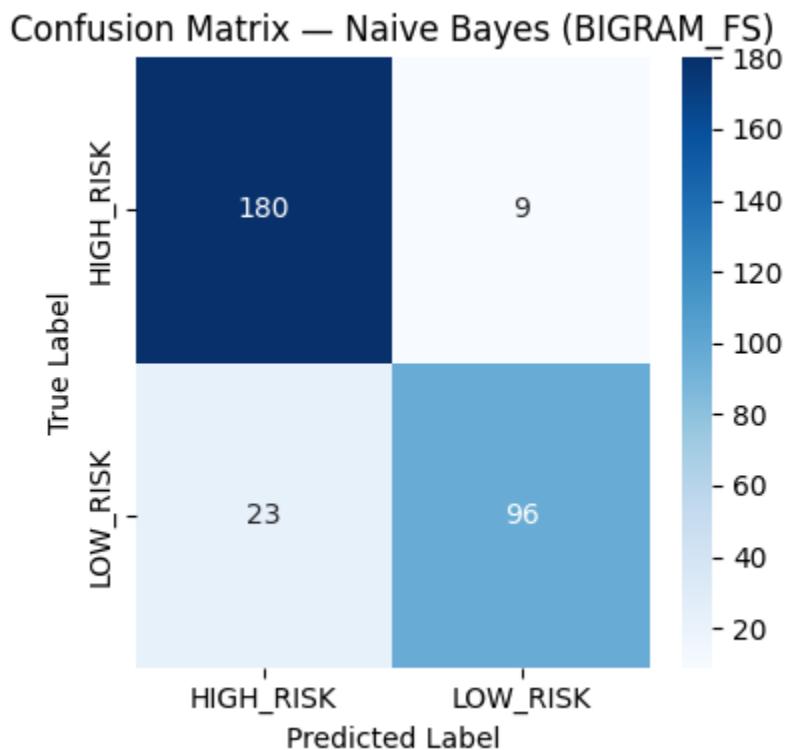


Gambar 5.2 Confusion matrix NB + TF-IDF Pada Dataset Demographics

Tabel 5.3 Classification Report untuk Model NB + N-Gram (Bigram) - Demographics Dataset

| | Precision | Recall | F1-Score | Support |
|-------------------|-----------|--------|----------|---------|
| HIGH_RISK | 0.8867 | 0.9524 | 0.9184 | 189 |
| LOW_RISK | 0.9143 | 0.8067 | 0.8571 | 119 |
| Accuracy | | | 0.8961 | 308 |
| Macro avg | 0.9005 | 0.8796 | 0.8878 | 308 |
| Weight avg | 0.8974 | 0.8961 | 0.8947 | 308 |

Model *Naive Bayes* dengan N-Gram memberikan hasil akurasi **89.61%**, lebih baik dari TF-IDF dan BoW. Konteks dua kata yang diperhitungkan dalam Bigram meningkatkan kemampuan model mengenali kombinasi kata kunci, meskipun masih kalah dibanding *Logistic Regression* yang mampu menangani dependensi fitur lebih kompleks.



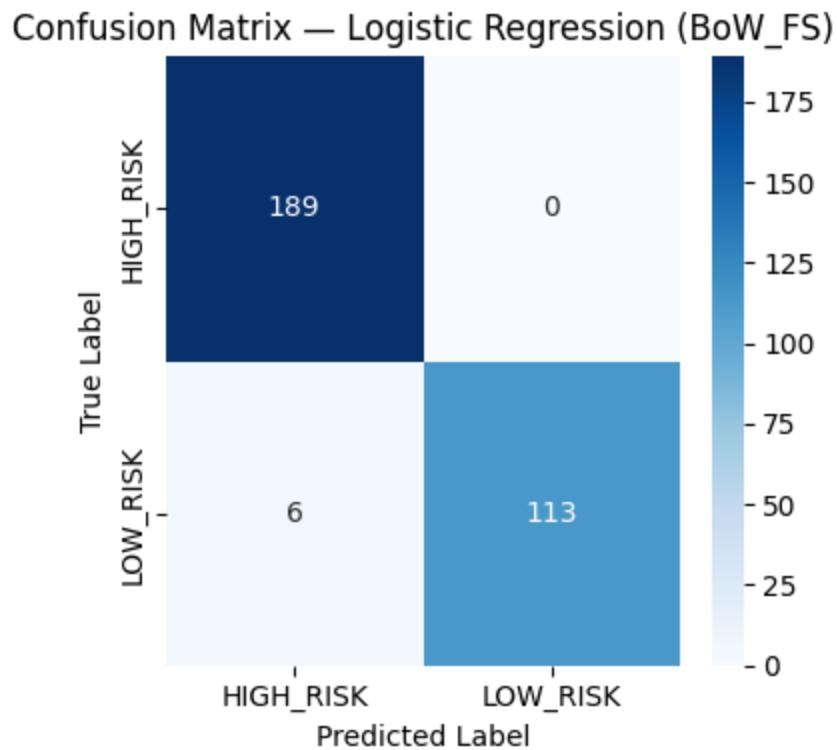
Gambar 5.3 Confusion matrix NB + N-Gram Pada Dataset Demographics

Tabel 5.4 Classification Report untuk Model Logistic Regression + BoW – Demographics

Dataset

| | Precision | Recall | F1-Score | Support |
|-------------------|-----------|--------|----------|---------|
| HIGH_RISK | 0.9692 | 1.0000 | 0.9844 | 189 |
| LOW_RISK | 1.0000 | 0.9496 | 0.9741 | 119 |
| Accuracy | | | 0.9805 | 308 |
| Macro avg | 0.9846 | 0.9748 | 0.9793 | 308 |
| Weight avg | 0.9811 | 0.9805 | 0.9804 | 308 |

Gambar ini menampilkan confusion matrix untuk model *Logistic Regression* menggunakan representasi BoW. Dengan akurasi **98.05%**, model menunjukkan kemampuan klasifikasi yang sangat baik dan stabil. Precision mencapai 1.00 untuk kelas LOW_RISK, menandakan tidak ada prediksi false positive. Hanya terdapat 6 kesalahan dari 308 data uji, menunjukkan bahwa representasi BoW mampu menangkap kata-kata penting yang berhubungan dengan risiko secara efektif.



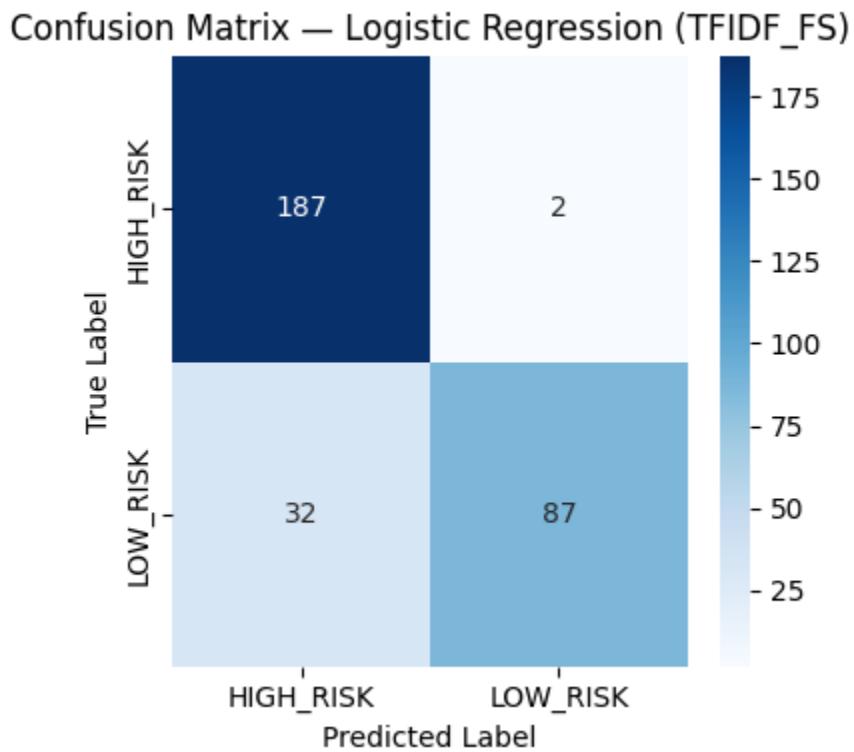
Gambar 5.4 Confusion matrix Logistic Regression + BoW – Demographics Dataset

Tabel 5.5 Classification Report untuk Model Logistic Regression + TF-IDF– Demographics Dataset

| | Precision | Recall | F1-Score | Support |
|-------------------|-----------|--------|----------|---------|
| HIGH_RISK | 0.8539 | 0.9894 | 0.9167 | 189 |
| LOW_RISK | 0.9775 | 0.7311 | 0.8365 | 119 |
| Accuracy | | | 0.8896 | 308 |
| Macro avg | 0.9157 | 0.8603 | 0.8766 | 308 |
| Weight avg | 0.9017 | 0.8896 | 0.8857 | 308 |

Model *Logistic Regression* dengan TF-IDF memiliki akurasi **88.96%**. Meskipun performanya menurun dibanding BoW, model ini menunjukkan kemampuan tinggi dalam mendeteksi kelas HIGH_RISK dengan recall mencapai **0.99**. Hal ini menunjukkan bahwa pembobotan kata-kata langka yang lebih informatif membantu mengenali kasus risiko tinggi, namun ada trade-off

berupa penurunan kinerja pada kelas LOW_RISK karena distribusi nilai TF-IDF yang lebih tidak seimbang.



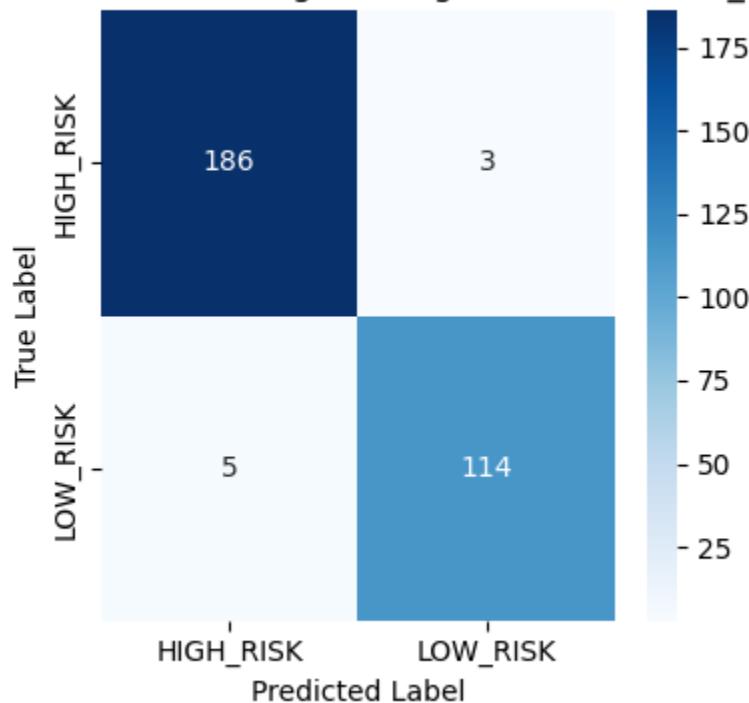
Gambar 5.5 Confusion matrix Logistic Regression + TF-IDF – Demographics Dataset

Tabel 5.6 Classification Report untuk Model Logistic Regression + N-Gram (Bigram) – Demographics Dataset

| | Precision | Recall | F1-Score | Support |
|-------------------|-----------|--------|----------|---------|
| HIGH_RISK | 0.9738 | 0.9841 | 0.9789 | 189 |
| LOW_RISK | 0.9744 | 0.9580 | 0.9661 | 119 |
| Accuracy | | | 0.9740 | 308 |
| Macro avg | 0.9741 | 0.9711 | 0.9725 | 308 |
| Weight avg | 0.9740 | 0.9740 | 0.9740 | 308 |

Model *Logistic Regression* dengan N-Gram (Bigram) menunjukkan akurasi **97.40%**, dengan keseimbangan precision dan recall di atas 0.97 untuk kedua kelas. Penggunaan Bigram membantu model memahami konteks pasangan kata seperti “berusia lanjut” atau “penyakit kronis”, yang sering berkorelasi dengan kategori HIGH_RISK. Ini menjadikannya salah satu model paling stabil dan andal pada dataset demografis.

Confusion Matrix — Logistic Regression (BIGRAM_FS)



Gambar 5.6 Confusion matrix Logistic Regression + N-Gram (Bigram) – Demographics Dataset

Tabel 5.7 Ringkasan Evaluasi Model – Demographic Dataset

| Model | Represen tasi | Accuracy | Precision | Recall | F1-Score | Total Salah |
|-------------|---------------|----------|-----------|--------|----------|-------------|
| Naive Bayes | BoW_FS | 0.8636 | 0.8697 | 0.8636 | 0.8597 | 53 |
| Naive Bayes | TFIDF_F S | 0.7695 | 0.8197 | 0.7695 | 0.7417 | 68 |
| Naive Bayes | BIGRAM | 0.8896 | 0.8923 | 0.8896 | 0.8876 | 32 |

| | _FS | | | | | |
|---------------------|-----------|--------|--------|--------|--------|----|
| Logistic Regression | BoW_FS | 0.9838 | 0.9842 | 0.9838 | 0.9837 | 6 |
| Logistic Regression | TFIDF_FS | 0.9156 | 0.9258 | 0.9156 | 0.9131 | 34 |
| Logistic Regression | BIGRAM_FS | 0.9643 | 0.9645 | 0.9643 | 0.9641 | 8 |

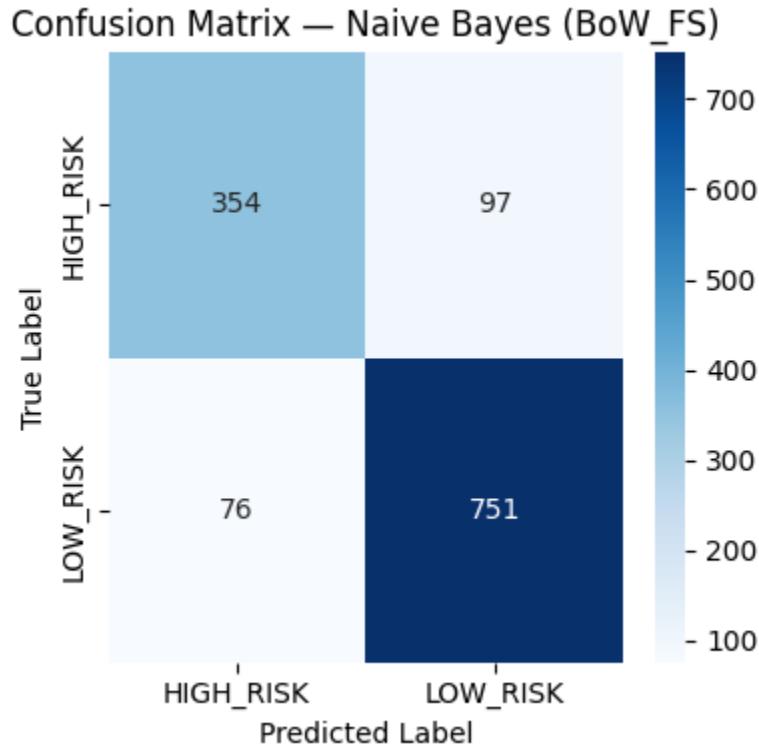
Tabel 5.7 menampilkan hasil evaluasi model pada dataset *Demographic* untuk semua kombinasi algoritma dan representasi fitur. Model Logistic Regression dengan representasi BoW_FS kembali menjadi yang terbaik, dengan akurasi 0.9838 dan jumlah kesalahan prediksi hanya 6 dari total 308 data uji. Representasi Bigram juga memberikan hasil yang sangat kompetitif dengan akurasi tinggi (0.9643), menunjukkan bahwa penambahan konteks dua kata berurutan membantu model mengenali pola linguistik yang berkaitan dengan risiko pasien. Sementara itu, performa Naive Bayes dengan TF-IDF_FS merupakan yang terendah (akurasi 0.7695, f1-score 0.7417), yang mengindikasikan bahwa Naive Bayes cenderung kurang efektif pada representasi dengan bobot frekuensi kata yang kompleks. Secara keseluruhan, kombinasi BoW_FS + Logistic Regression tetap menjadi pendekatan paling optimal untuk memprediksi kategori risiko berbasis teks demografis hasil serialisasi dan anonimisasi.

Berikut adalah visualisasi hasil confusion matrix dari Dataset Clinical Data :

Tabel 5.8 Classification Report untuk Model NB + BoW - ClinicalData Dataset

| | Precision | recall | f1-score | support |
|------------|-----------|--------|----------|---------|
| HIGH_RISK | 0.8233 | 0.7849 | 0.8036 | 451 |
| LOW_RISK | 0.8856 | 0.9081 | 0.8967 | 827 |
| Accuracy | | | 0.8646 | 1278 |
| Macro avg | 0.8544 | 0.8465 | 0.8502 | 1278 |
| Weight avg | 0.8636 | 0.8646 | 0.8639 | 1278 |

Model *Naive Bayes* dengan BoW_FS memperoleh akurasi **86.46%**. Meskipun lebih rendah dibanding *Logistic Regression*, hasil ini tetap menunjukkan performa yang stabil. NB cenderung lebih efektif dalam mengenali kelas LOW_RISK (recall 0.91) namun sedikit lemah pada HIGH_RISK (recall 0.78), yang menandakan kecenderungan model terhadap kata-kata umum pada kelas mayoritas.

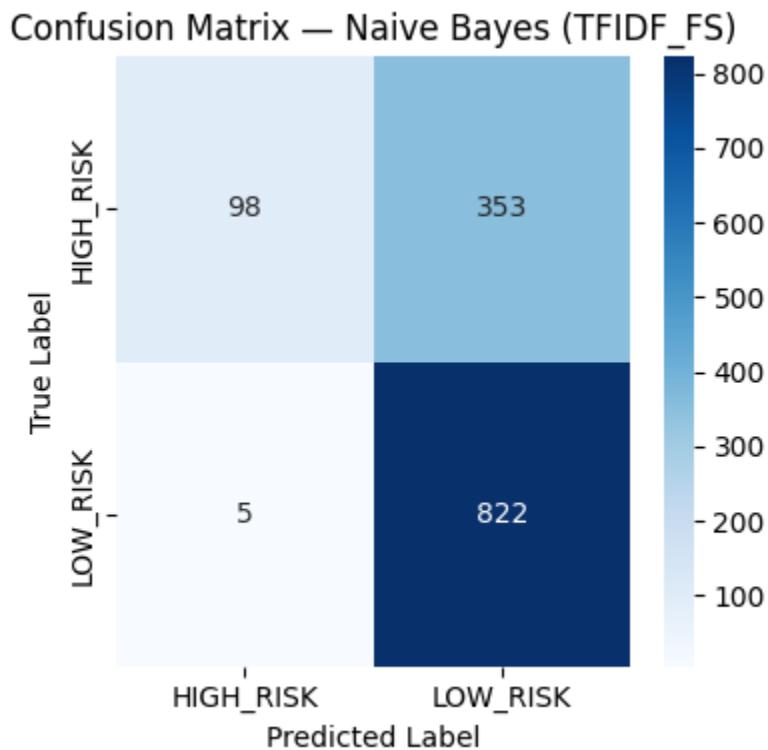


Gambar 5.8 Confusion matrix NB + BoW Pada Dataset ClinicalData

Tabel 5.9 Classification Report untuk Model NB + TF-IDF- ClinicalData Dataset

| | Precision | Recall | F1-Score | Support |
|------------|-----------|--------|----------|---------|
| HIGH_RISK | 0.9515 | 0.2173 | 0.3538 | 451 |
| LOW_RISK | 0.6996 | 0.9940 | 0.8212 | 827 |
| Accuracy | | | 0.7199 | 1278 |
| Macro avg | 0.8255 | 0.6056 | 0.5875 | 1278 |
| Weight avg | 0.7885 | 0.7199 | 0.6562 | 1278 |

Model *Naive Bayes* menunjukkan performa yang tidak seimbang pada TF-IDF, dengan akurasi hanya 71.99%. Recall untuk HIGH_RISK sangat rendah (0.21), menandakan model kesulitan mengidentifikasi pasien berisiko tinggi. TF-IDF yang berbasis frekuensi kata memberikan bobot berlebih pada kata-kata unik, menyebabkan *Naive Bayes* gagal menyeimbangkan distribusi antar kelas.



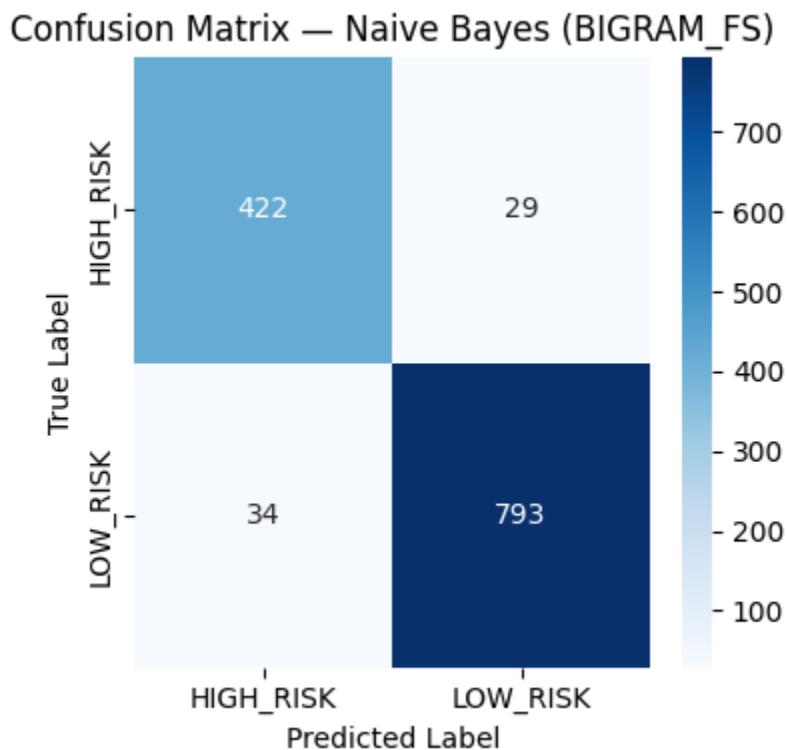
Gambar 5.9 Confusion matrix NB + TF-IDFPada Dataset ClinicalData

**Tabel 5.10 Classification Report untuk Model NB + N-Gram (Bigram) - ClinicalData
Dataset**

| | Precision | Recall | F1-Score | Support |
|--|-----------|--------|----------|---------|
| | | | | |

| | | | | |
|-------------------|--------|--------|--------|------|
| HIGH_RISK | 0.9254 | 0.9357 | 0.9305 | 451 |
| LOW_RISK | 0.9647 | 0.9589 | 0.9618 | 827 |
| Accuracy | | | 0.9507 | 1278 |
| Macro avg | 0.9451 | 0.9473 | 0.9462 | 1278 |
| Weight avg | 0.9509 | 0.9507 | 0.9508 | 1278 |

Model *Naive Bayes* dengan Bigram memperoleh akurasi **95.07%**, jauh lebih baik dibanding TF-IDF. Kombinasi dua kata meningkatkan kemampuan NB dalam menangkap konteks penting yang tidak terlihat pada representasi unigram. Hal ini menunjukkan bahwa penambahan konteks antar kata dapat mengurangi bias terhadap kelas mayoritas yang sering muncul pada metode probabilistik sederhana.

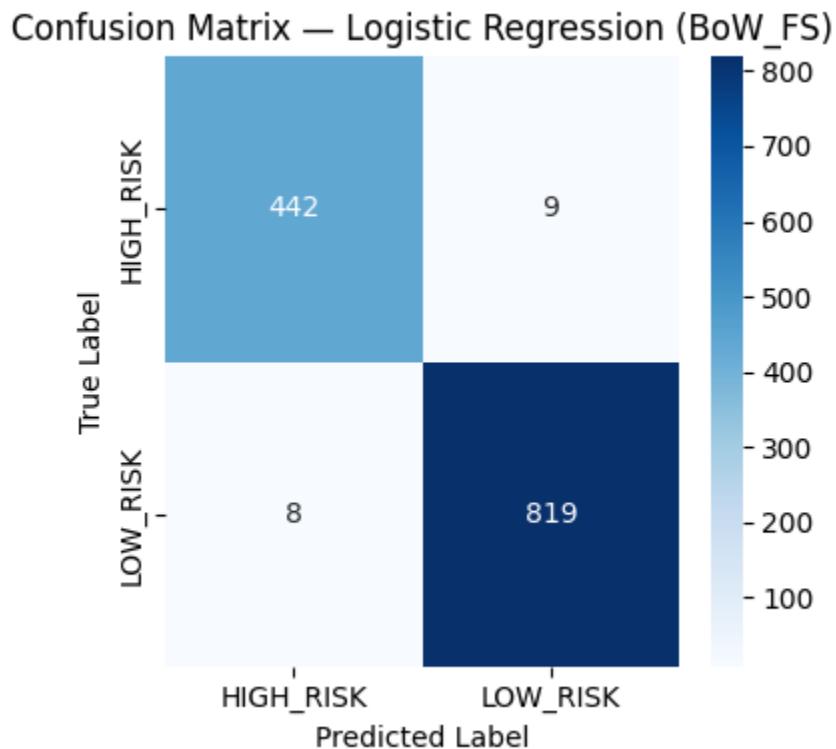


Gambar 5.10 Confusion matrix NB + N-Gram Pada Dataset ClinicalData

Tabel 5.11 Classification Report untuk Model LR + BoW – ClinicalData Dataset

| | Precision | Recall | F1-Score | Support |
|-------------------|-----------|--------|----------|---------|
| HIGH_RISK | 0.9822 | 0.9800 | 0.9811 | 451 |
| LOW_RISK | 0.9891 | 0.9903 | 0.9897 | 827 |
| Accuracy | | | 0.9867 | 1278 |
| Macro avg | 0.9857 | 0.9852 | 0.9854 | 1278 |
| Weight avg | 0.9867 | 0.9867 | 0.9867 | 1278 |

Model *Logistic Regression* dengan representasi **BoW_FS** memberikan performa tertinggi di antara seluruh kombinasi dengan akurasi **98.67%**. Model menunjukkan keseimbangan precision dan recall di atas 0.98 pada kedua kelas, dengan perbedaan yang sangat kecil. Hal ini menunjukkan bahwa representasi BoW sudah cukup kuat untuk menangkap korelasi kata-kata klinis yang relevan dengan tingkat risiko pasien.

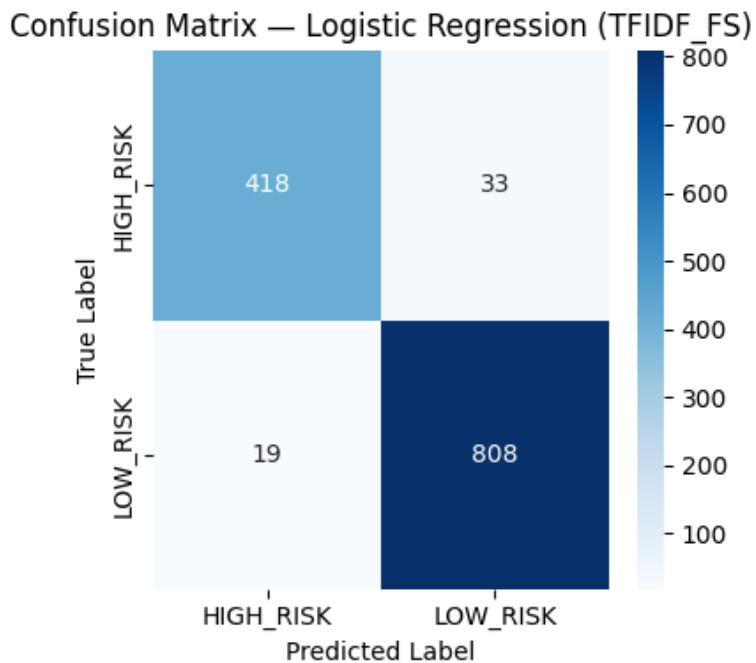


Gambar 5.11 Confusion Matrix LR + BoW – ClinicalData Dataset

Tabel 5.12 Classification Report untuk Model LR + TF-IDF – ClinicalData Dataset

| | Precision | Recall | F1-Score | Support |
|-------------------|-----------|--------|----------|---------|
| HIGH_RISK | 0.9565 | 0.9268 | 0.9414 | 451 |
| LOW_RISK | 0.9608 | 0.9770 | 0.9688 | 827 |
| Accuracy | | | 0.9593 | 1278 |
| Macro avg | 0.9586 | 0.9519 | 0.9551 | 1278 |
| Weight avg | 0.9593 | 0.9593 | 0.9592 | 1278 |

Dengan akurasi **95.93%**, model *Logistic Regression* dengan TF-IDF menunjukkan keseimbangan performa yang baik antara recall dan precision. Nilai recall tinggi untuk **LOW_RISK** (0.98) menandakan bahwa model ini sangat jarang gagal mengenali pasien berisiko rendah. Kombinasi TF-IDF dengan LR juga memberikan stabilitas yang baik pada hasil cross-validation, memperlihatkan kemampuan generalisasi yang kuat.

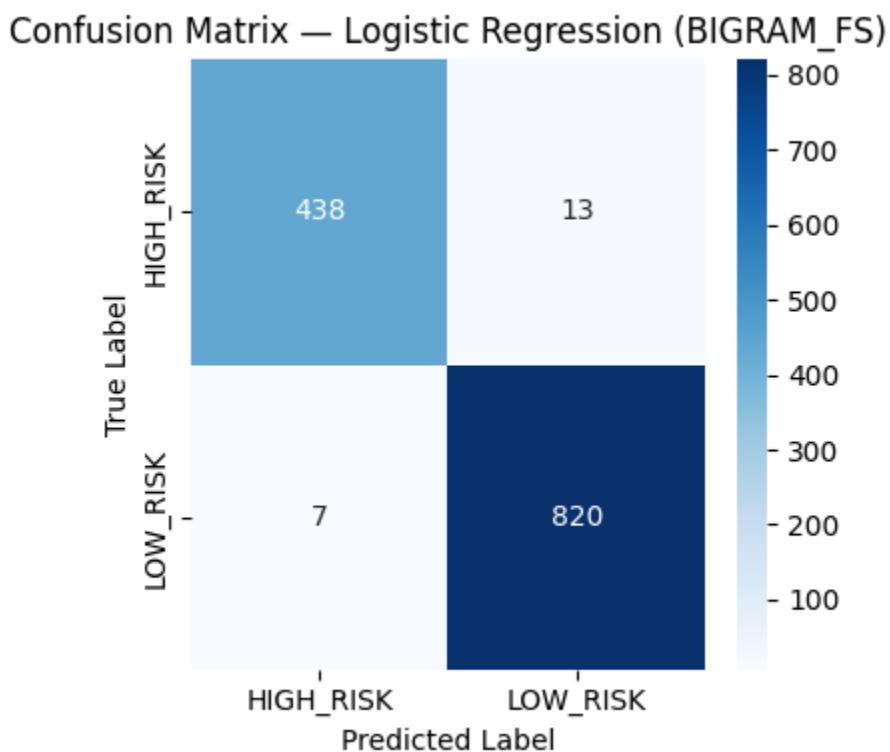


Gambar 5.12 Confusion Matrix LR + LR + TF-IDF – ClinicalData Dataset

Tabel 5.13 Classification Report untuk Model LR + N-Gram (Bigram) - ClinicalData Dataset

| | Precision | Recall | F1-Score | Support |
|-------------------|-----------|--------|----------|---------|
| HIGH_RISK | 0.9843 | 0.9712 | 0.9777 | 451 |
| LOW_RISK | 0.9844 | 0.9915 | 0.9880 | 827 |
| Accuracy | | | 0.9844 | 1278 |
| Macro avg | 0.9843 | 0.9814 | 0.9828 | 1278 |
| Weight avg | 0.9843 | 0.9844 | 0.9843 | 1278 |

Model *Logistic Regression* dengan representasi Bigram menghasilkan akurasi **98.44%**, sedikit di bawah BoW_FS namun tetap unggul secara konsisten. Bigram membantu model mengenali konteks dua kata seperti “gagal jantung”, “penyakit kronis”, dan “komplikasi operasi”, yang sangat relevan dengan klasifikasi risiko medis.



Gambar 5.13 Confusion matrix LR + N-Gram Pada Dataset ClinicalData

Tabel 5.14 Ringkasan Evaluasi Model – ClinicalData Dataset

| Model | Represen tasi | Accuracy | Precision | Recall | F1-Score | Total Salah |
|---------------------|---------------|----------|-----------|--------|----------|-------------|
| Naive Bayes | BoW_FS | 0.8646 | 0.8636 | 0.8646 | 0.8639 | 173 |
| Naive Bayes | TFIDF_F S | 0.7199 | 0.7885 | 0.7199 | 0.6562 | 358 |
| Naive Bayes | BIGRAM _FS | 0.9507 | 0.9509 | 0.9507 | 0.9508 | 63 |
| Logistic Regression | BoW_FS | 0.9867 | 0.9867 | 0.9867 | 0.9867 | 17 |
| Logistic Regression | TFIDF_F S | 0.9593 | 0.9593 | 0.9593 | 0.9592 | 52 |
| Logistic Regression | BIGRAM _FS | 0.9844 | 0.9843 | 0.9844 | 0.9843 | 20 |

Tabel 5.14 menampilkan ringkasan evaluasi seluruh kombinasi model dan representasi teks pada dataset *ClinicalData*. Model Logistic Regression dengan representasi BoW_FS menunjukkan performa terbaik dengan akurasi tertinggi (0.9867) dan jumlah kesalahan prediksi paling sedikit (17). Model ini menunjukkan keseimbangan precision dan recall yang sempurna untuk kedua kelas. Sementara itu, Naive Bayes dengan TFIDF_FS memberikan performa terendah (akurasi 0.7199, f1-score 0.6562) dan jumlah kesalahan tertinggi (358), menandakan bahwa pembobotan kata berbasis frekuensi pada TF-IDF tidak efektif untuk data klinis yang memiliki istilah spesifik dan jarang muncul. Secara keseluruhan, kombinasi BoW_FS + Logistic Regression paling optimal dalam mempelajari pola teks medis terstruktur hasil anonimisasi, sementara model Naive Bayes lebih stabil saat menggunakan representasi Bigram.

Tahap 6: Deep Learning

Tahap deep learning dilakukan untuk mengevaluasi performa model klasifikasi berbasis teks pada dua dataset yaitu Demographic dan ClinicalData. Empat model digunakan secara konsisten pada kedua dataset, yaitu RNN, LSTM, Fine-Tuned LSTM (FT-LSTM), dan LoRA-BERT, sehingga perbandingan performa dapat dilakukan secara objektif.

Selain empat model utama yang telah diuji (RNN, LSTM, FT-LSTM, dan LoRA-BERT Ultra-Lite), dilakukan eksperimen tambahan menggunakan LoRA Modern berbasis cahya/bert-base-indonesian-1.5G, yaitu model BERT ukuran besar yang lebih representatif untuk Bahasa Indonesia. Eksperimen ini bertujuan untuk mengetahui apakah peningkatan kapasitas model Transformer dapat mengungguli performa FT-LSTM yang sebelumnya menjadi model terbaik.

Model LoRA Modern tetap menggunakan pendekatan *parameter-efficient fine-tuning* (LoRA) namun dengan arsitektur BERT penuh, bukan varian lite. Parameter LoRA disisipkan pada modul query, key, dan value, dengan konfigurasi: $r = 8$, $\alpha = 16$, $\text{dropout} = 0.1$. Proses pelatihan dilakukan selama 2 epoch dengan batch size kecil (2) untuk efisiensi memori dan learning rate $2e-5$.

Proses eksperimen meliputi:

- tokenisasi dan padding teks,
- pemisahan data train–test,
- penerapan *class weight* untuk menangani *imbalance*,
- pelatihan model deep learning,
- evaluasi menggunakan accuracy, F1-Score, classification report, dan confusion matrix.

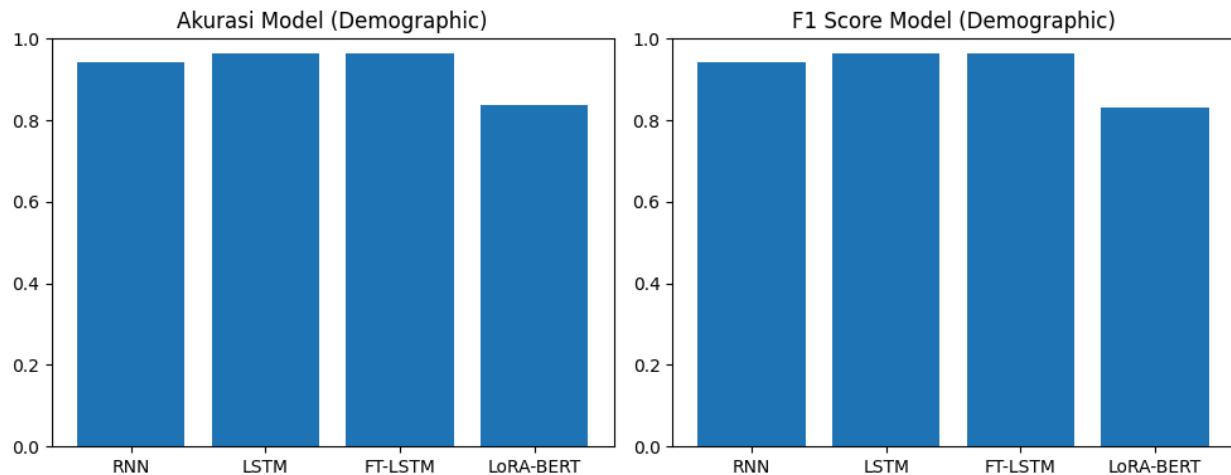
Deep Learning – Dataset Demographic

Model pertama RNN menggunakan dua lapisan *Bidirectional SimpleRNN* bertingkat dengan embedding 256 dan dropout untuk regularisasi dan *Dense layer* untuk klasifikasi. Model kedua LSTM menggunakan *Bidirectional LSTM* bertingkat dengan 128 dan 64 unit, dropout 0.3, optim Adam sehingga lebih baik dalam menangkap konteks kata dalam urutan. Model ketiga FT-LSTM menerapkan pencarian konfigurasi hiperparameter sederhana (sequence length dan jumlah unit LSTM) sebelum dilatih ulang sebagai model final. Model keempat LoRA-BERT memanfaatkan *parameter-efficient fine-tuning* sehingga hanya sebagian kecil parameter BERT yang diperbarui selama pelatihan.

Hasil eksperimen menunjukkan LSTM dan FT-LSTM memperoleh performa tertinggi, sedangkan LoRA-BERT menunjukkan penurunan performa karena keterbatasan kapasitas model BERT kecil pada domain data ini.

Tabel 6.1 Perbandingan Performa Model Dataset Demographic

| Model | Accuracy | F1 - Score |
|---------------------------------|----------|------------|
| RNN | 0.94 | 0.94 |
| LSTM | 0.96 | 0.96 |
| Fine - Tunned LSTM (FT - LSTM) | 0.96 | 0.96 |
| LoRa - Ultra Lite | 0.84 | 0.83 |
| LoRa Modern | 0.82 | 0.81 |



Gambar 6.1 Visualisasi Summary Dataset Demographic

Gambar menampilkan perbandingan akurasi dan F1-Score dari empat model deep learning pada dataset Demographic. Grafik menunjukkan bahwa FT-LSTM berada pada posisi tertinggi baik pada akurasi maupun F1-Score, diikuti oleh LSTM yang memiliki kinerja sangat dekat. RNN terlihat berada di posisi menengah, sementara LoRA-BERT menjadi model dengan performa paling rendah, sehingga *class imbalance* tetap memengaruhi performa LoRA Modern pada dataset ini. Pola ini mengindikasikan bahwa pendekatan LSTM dengan konfigurasi yang dituning secara optimal lebih mampu menangkap karakteristik teks dalam dataset Demographic dibandingkan model lain.

Deep Learning – Dataset ClinicalData

Eksperimen dilakukan dengan skema yang sama, namun konfigurasi tokenisasi diatur menjadi LITE untuk menyesuaikan karakter teks klinis yang cenderung lebih pendek. Sebagian besar langkah pelatihan sama, termasuk penggunaan *class weight* untuk menyeimbangkan distribusi data yang timpang (majoritas HIGH_RISK dan minoritas LOW_RISK).

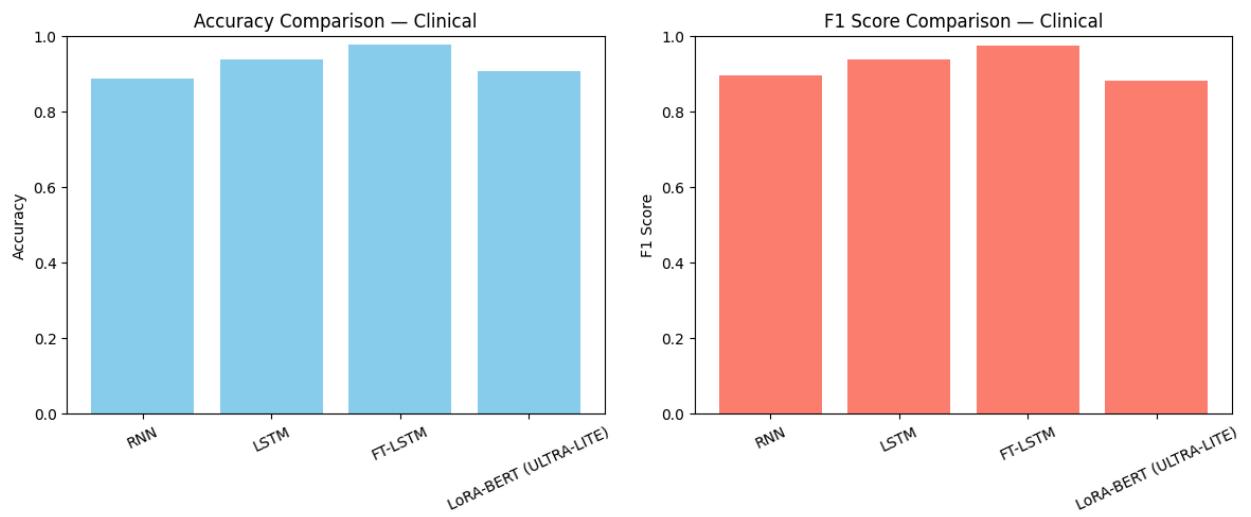
RNN memperlihatkan akurasi baik namun kurang mampu mengenali kelas minoritas. LSTM memberikan performa stabil di kedua metrik. FT-LSTM kembali menjadi model terbaik setelah proses penyetelan konfigurasi. LoRA-BERT menghasilkan akurasi tinggi namun kesulitan mengklasifikasikan kelas minoritas LOW_RISK.

Model pertama RNN menggunakan Bidirectional SimpleRNN dengan 2 layer dan dropout. Model kedua LSTM menggunakan Bidirectional LSTM dengan 128 dan 64 unit. Model ketiga FT-LSTM Fine-tuning dari LSTM untuk stabilitas dan peningkatan akurasi. Model keempat LoRA-BERT (ULTRA-LITE) Parameter-efficient fine-tuning pada BERT compact (L-2 H-128) dengan konfigurasi LoRA r=8.

Hasil eksperimen menunjukkan FT-LSTM memberikan performa terbaik dan unggul signifikan dibanding model lain, sedangkan LoRA-BERT memiliki akurasi kompetitif namun F1-Score lebih rendah karena recall imbalance pada label LOW_RISK.

Tabel 6.2 Perbandingan Performa Model Dataset ClinicalData

| Model | Accuracy | F1 - Score |
|--------------------------------|-----------------|-------------------|
| RNN | 0.88 | 0.89 |
| LSTM | 0.94 | 0.94 |
| Fine - Tuned LSTM (FT - LSTM) | 0.98 | 0.98 |
| LoRa - Ultra Lite | 0.91 | 0.88 |
| LoRa Modern | 0.86 | 0.86 |



Gambar 6.2 Visualisasi Summary Dataset ClinicalData

Gambar memperlihatkan perbandingan akurasi dan F1-Score dari model deep learning pada dataset ClinicalData. Terlihat peningkatan performa bertahap dari RNN menuju LSTM hingga mencapai puncaknya pada FT-LSTM yang menunjukkan performa paling unggul secara konsisten. LoRA-BERT memiliki akurasi cukup tinggi namun tetap tertinggal secara F1-Score sehingga menandakan ketidakstabilan prediksi pada kelas minoritas. Secara keseluruhan, visualisasi ini mempertegas dominasi FT-LSTM sebagai model paling stabil dan akurat dalam memproses teks klinis.

Interpretasi:

- FT-LSTM paling stabil & terbaik untuk kedua dataset, menunjukkan generalisasi kuat dan recall/precision tinggi pada kedua kelas (terutama pada Clinical: Accuracy 0.9773, F1 0.9769). Ini menandakan fine-tuning arsitektur LSTM dengan konfigurasi yang dipilih efektif menangkap pola risiko dalam teks medis/demografis.
- LSTM standar juga sangat baik (Demographic & Clinical) cocok bila resource lebih terbatas dibanding FT procedure yang lebih panjang.
- RNN (SimpleRNN) memberikan baseline yang wajar pada Demographic, namun pada Clinical lebih sensitif terhadap imbalance (meskipun class_weight diterapkan).
- LoRA-BERT (DistilBERT / lite BERT) menghasilkan akurasi tinggi pada beberapa kasus, namun kinerjanya menonjol pada kelas mayoritas sehingga pada kelas minor (LOW_RISK pada Clinical) recall drop (lihat confusion matrix: banyak false positives untuk HIGH_RISK). Artinya model pretrained + LoRA butuh tuning lebih lanjut (learning rate, class weighting, thresholding) atau augmentasi data pada kelas minor

untuk memperbaiki recall LOW_RISK.

5. Imbalance tetap menjadi faktor penting di Clinical (proporsi HIGH_RISK jauh lebih besar dalam salah satu versi) meskipun class_weight digunakan, beberapa model (terutama transformer lite) masih menunjukkan gap recall antar kelas.

Secara keseluruhan, FT-LSTM terbukti sebagai model paling unggul pada kedua dataset. Pada dataset Demographic, performa terbaik diperoleh karena FT-LSTM mampu menangkap konteks kata secara lebih stabil dan mendalam. Pada dataset ClinicalData, FT-LSTM kembali menjadi yang paling akurat karena mampu mengenali kelas minoritas dengan lebih baik dibandingkan model lain. Dengan demikian, FT-LSTM memberikan hasil paling konsisten dan reliabel untuk klasifikasi risiko berbasis teks.

- Error Analysis

Tabel Eror Analysis Dataset Demographic

| Text | Actual | Model | | | | | | Jumlah Salah | Alasan |
|--|--------|-------------|-------------|--------------------|----------------|-------------------|--------------------|-----------------|--|
| | | LR + BoW | NB + BoW | LR + TF-ID F | NB + TF-IDF | LR+ N-Gra m | NB + N-Gra m | | |
| pasien patienttoken pria berusia 20-39 ras disebutkan memiliki kondisi kesehatan ringan | LR | HR | HR | HR | HR | HR | HR | 6 | Semua model salah memprediksi karena kombinasi token “patienttoken”, info usia 20-39 dan ras membingungkan model sehingga prediksi menjadi HIGH_RISK |
| pasien patienttoken patient pria ras disebutkan memiliki hipertensi ringan dan riwayat penyakit ginjal | LR | H | H | H | H | H | H | 6 | Kesalahan karena pengulangan kata “patient” + info ras dan kondisi ginjal membuat semua model salah memprediksi HIGH_RISK |

| | | | | | | | | | |
|---|----|---|---|---|---|---|---|----|--|
| pasien patienttoken wanita berusia 20-39 ras disebutkan memiliki riwayat penyakit ginjal kronis | LR | H | H | H | H | H | H | 6 | Model salah membaca info usia dan riwayat penyakit ginjal sehingga menafsirkan risiko lebih tinggi (HIGH_RISK) |
| pasien patienttoken berusia 80-99 ras berjenis kelamin laki-laki memiliki hipertensi dan riwayat penyakit ginjal | HR | L | L | L | L | L | L | 53 | Naive Bayes salah memprediksi LOW_RISK, sementara LR tetap benar; kesalahan NB disebabkan info usia 80-99 dan ras yang menimbulkan bias. |
| pasien patienttoken wanita berkulit putih berusia 40-59 memiliki hipertensi ringan | LR | L | L | L | L | L | L | 53 | NB salah prediksi HIGH_RISK akibat kombinasi token “patienttoken”, ras kulit putih, dan usia, sedangkan LR benar. |

Tabel Eror Analysis Dataset ClinicalData

| Text | Actual | Model | Jumlah | Alasan |
|------|--------|-------|--------|--------|
|------|--------|-------|--------|--------|

| | | LR + BoW | NB + BoW | LR + TF-ID F | NB + TF-ID F | LR+ N-Gra m | NB + N-Gram | Salah | |
|---|----|---------------------|---------------------|-----------------------------|-----------------------------|----------------------------|------------------------|--------------|--|
| pasien id id subjek wanita berusia 60-79 numtoken berat badan normal, tekanan darah terkontrol, menjalani operasi pra-operasi minor | HR | H | L | H | H | H | L | 4 | Kombinasi informasi usia 60-79, jenis kelamin wanita, dan kondisi pra-operasi minor membuat beberapa model keliru menafsirkan risiko sebagai LOW_RISK. |
| pasien case id subject id pria berusia 60-79 numtoken berat badan normal, menjalani operasi jantung minor, dicatat dalam rekam medis | HR | H | L | H | L | H | L | 4 | Token “case id” dan info operasi jantung minor menimbulkan ambiguitas bagi beberapa model sehingga salah prediksi. |

| | | | | | | | | | | |
|---|----|---|---|---|---|---|---|--|---|--|
| pasien id id subjek wanita berusia 20-39 numtoken berat badan normal, riwayat operasi tonsil minor, tekanan darah normal | LR | L | L | L | L | L | H | | 1 | Naive Bayes salah memprediksi HIGH_RISK karena adanya token "numtoken" dan info operasi minor. |
| pasien id id subjek pria berusia 40-59 numtoken berat badan normal, memiliki riwayat penyakit jantung ringan dan tekanan darah sedikit tinggi | LR | L | H | L | H | L | H | | 4 | Informasi usia 40-59 dan penyakit jantung ringan menyebabkan model TFIDF dan Naive Bayes keliru HIGH_RISK. |
| pasien id id subjek wanita berusia 20-39 numtoken berat badan normal, menjalani operasi minor gigi dan riwayat alergi obat ringan | LR | L | H | L | H | L | H | | 4 | Beberapa model salah menafsirkan informasi operasi minor dan riwayat alergi sebagai HIGH_RISK. |

Tahap 7: Knowledge Interpretation (20 Poin)