

Aktivitas 7
Mohd Azhima

1. Deskripsikan unit analisis dari data yang akan anda proses
2. Deskripsikan model apa yang kira-kira sesuai dengan penelitian anda
3. Buat skenario evaluasi (dalam bentuk tabel) terkait dengan penelitian anda

Jawab:

1. Unit analisis penelitian saya adalah Dataset pasien stroke yang memiliki **target label stroke dan tidak stroke yang tidak seimbang**. Data ini berisi sekumpulan fitur (variabel independen) seperti usia, tekanan darah, kadar kolesterol, riwayat merokok, gula darah, dan atribut klinis lainnya. Di akhir baris tersebut, terdapat satu target label (variabel dependen) yang Anda teliti, misalnya: "Terkena Stroke" (Ya/Tidak).
2. Pada penelitian sebelumnya, dataset ini saya gunakan untuk klasifikasi stroke menggunakan model bpnn dengan data yg sudah diseimbangkan/ augmentasi menggunakan smote. Maka, di penelitian ini saya akan melakukan Studi Perbandingan Kinerja (Comparative Performance Study) model augmentasi data untuk mengevaluasi dampak dari model- model keseimbangan data.

Model Augmentasi (Variabel Eksperimen): Ini adalah model utama yang sedang dibandingkan.

- Metode 1 (Pembanding/Baseline): SMOTE. Metode *oversampling* berbasis interpolasi yang sudah Anda gunakan sebelumnya.
- Metode 2 (Model Baru): Model Generatif CTGAN (Conditional Tabular GAN): Ini adalah pilihan yang sangat baik. Tidak seperti SMOTE yang hanya membuat data "di antara" titik-titik yang ada, CTGAN adalah *Generative Adversarial Network* yang "belajar" *seluruh distribusi statistik* dari data asli Anda. Model ini berpotensi membuat data sintetis yang jauh lebih bervariasi dan realistis, menangkap korelasi kompleks antar fitur (misalnya, hubungan antara usia dan tekanan darah).

Model Klasifikasi (Alat Ukur/Kontrol): Untuk mengukur keefektifan metode augmentasi secara adil, saya akan tetap menggunakan **model klasifikasi bpnn** seperti penelitian sebelumnya agar hasil tetap konsisten di semua eksperimen.

3. Perbandingan Efektivitas Metode Augmentasi Data Tabular

Skenario Eksperimen	Data Latih (Training Set)	Model Klasifikasi	Metrik Kinerja Kunci (di Test Set)	Tujuan Skenario
1. Baseline (Tanpa Augmentasi)	Data Latih Asli (Tidak Seimbang)	BPNN	<ul style="list-style-type: none"> • F1-Score • Recall (Sensitivity) • PR-AUC (Precision-Recall AUC) 	Menetapkan kinerja dasar. Ini adalah angka "kontrol".
2. Metode Sebelumnya (Pembanding)	Data Latih Asli + Augmentasi SMOTE	BPNN	<ul style="list-style-type: none"> • F1-Score • Recall (Sensitivity) • PR-AUC 	Mereplikasi hasil penelitian sebelumnya. Seberapa baik kinerja metode interpolasi?
3. Metode Baru	Data Latih Asli +	BPNN	<ul style="list-style-type: none"> • F1-Score 	Inti Penelitian:

(Eksperimen)	Augmentasi CTGAN		<ul style="list-style-type: none"> • Recall (Sensitivity) • PR-AUC 	Mengukur kinerja metode generatif. Apakah CTGAN secara signifikan mengungguli SMOTE dan Baseline?
4. (Opsional) Analisis Realisme	Data Sintetis Murni dari SMOTE vs. CTGAN	-	<ul style="list-style-type: none"> • Analisis Distribusi (misal: Visualisasi t-SNE) • Perbandingan Korelasi Fitur 	Memeriksa apakah data buatan CTGAN secara statistik lebih "mirip" dengan data asli dibandingkan data buatan SMOTE.