

Classification of Stroke Disease Using the Learning Vector Quantization Algorithm

Andriarmi, Chairina Wirdiastuti*, Syafriandi

Departemen Statistika, Universitas Negeri Padang, Padang, Indonesia

*Corresponding author: chairinawirdiastuti01@email.com

Submitted : 30 April 2026

Revised : 20 Mei 2026

Accepted : 28 Mei 2026

ABSTRACT

Stroke is one of the leading causes of death and disability worldwide, thereby making early detection crucial for timely and appropriate medical treatment. In clinical practice, stroke diagnosis is generally carried out through medical examinations and patient history analysis, but this process is time-consuming and depends on the subjective judgment of medical personnel. Therefore, machine learning approaches can be utilized to support disease classification more quickly and objectively. This study aims to analyze the performance of the Learning Vector Quantization (LVQ) method in classifying stroke disease using a dataset obtained from Kaggle. The dataset used in this study is imbalanced; therefore, the SMOTE (Synthetic Minority Over-sampling Technique) method was applied to handle class imbalance. The research stages included data preprocessing, splitting data into training and testing sets, LVQ model training, parameter optimization using learning rate and maximum epoch, and model evaluation using accuracy and sensitivity. The results show that the LVQ model trained on the original dataset achieved an accuracy of 95,72%, but failed to detect stroke cases with a sensitivity of 0%. After applying SMOTE, the best model achieved a stroke sensitivity of 90%, although the accuracy decreased to 49,49% due to the high number of false positives. These findings indicate that LVQ is highly sensitive to data distribution and model parameters, making its performance on this dataset less optimal for stroke classification and more suitable as an initial screening tool.

Keywords: Imbalanced Dataset, Learning Vector Quantization, Machine Learning, SMOTE, Stroke Classification.



This is an open access article under the Creative Commons 4.0 Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. ©2022 by author and Universitas Negeri Padang.

I. PENDAHULUAN

Stroke merupakan salah satu penyebab utama kematian dan kecacatan secara global yang terjadi akibat terganggunya aliran darah menuju otak, sehingga menimbulkan kerusakan permanen pada jaringan otak (World Health Organization, 2025). Selain itu, stroke juga menimbulkan beban ekonomi yang besar, baik bagi individu maupun sistem kesehatan, karena memerlukan perawatan jangka panjang serta rehabilitasi yang intensif (Feigin dkk, 2014). Tingginya insiden stroke serta dampaknya yang signifikan terhadap kualitas hidup penderita menjadikan upaya deteksi dini sebagai hal yang sangat krusial dalam penanganan penyakit ini (Johnson dkk, 2019).

Dalam praktik klinis, diagnosis stroke umumnya dilakukan melalui serangkaian pemeriksaan medis serta analisis riwayat kesehatan pasien. Namun proses tersebut seringkali memakan waktu dan sangat bergantung pada penilaian subjektif tenaga medis (Katan & Luft, 2018). Oleh karena itu, pendekatan berbasis *machine learning* mulai banyak digunakan untuk mendukung proses klasifikasi penyakit agar lebih cepat, konsisten, dan objektif (Obermeyer & Emanuel, 2016).

Klasifikasi merupakan salah satu teknik dalam *machine learning* yang bertujuan untuk mengelompokkan data ke dalam kelas atau kategori tertentu berdasarkan karakteristik yang dimilikinya. Dalam bidang medis, teknik klasifikasi dimanfaatkan untuk mendukung proses pengambilan keputusan klinis, seperti mengidentifikasi kondisi pasien atau menentukan tingkat risiko terhadap suatu penyakit. Dengan memanfaatkan pola yang diperoleh dari data historis, metode klasifikasi mampu menghasilkan prediksi yang lebih sistematis, konsisten, dan berbasis data (Han dkk, 2012)

Salah satu metode yang dapat digunakan adalah *Learning Vector Quantization* (LVQ), yaitu algoritma *supervised learning* berbasis *prototype* yang mengklasifikasikan data berdasarkan kedekatan jarak terhadap vektor representatif dari setiap kelas (Kohonen, 2001). Metode ini mampu menemukan pola dalam data dan mengelompokkan data baru berdasarkan pola tersebut. LVQ memiliki keunggulan berupa struktur model yang sederhana serta mudah diinterpretasikan, sehingga memiliki potensi untuk diterapkan pada data medis (Kaden dkk, 2015; Villmann dkk, 2017).

Beberapa penelitian sebelumnya menunjukkan bahwa LVQ telah banyak diaplikasikan dalam bidang kesehatan. Studi mengenai klasifikasi penyakit ginjal kronis menunjukkan bahwa LVQ mampu mencapai tingkat akurasi sebesar 93,98% dalam mengklasifikasikan kondisi pasien (Zabihullah dkk, 2025). Selain itu, metode ini juga digunakan dalam klasifikasi pasien diabetes untuk mengidentifikasi pola klinis pasien secara efektif (Sianipar & Yasin, 2025). Pada bidang kardiovaskular, LVQ diterapkan untuk mengklasifikasikan aritmia jantung berbasis data elektrokardiogram (ECG) dan menghasilkan kinerja klasifikasi yang baik (Melin dkk, 2014).

Salah satu tantangan utama dalam penerapan klasifikasi pada data medis adalah adanya ketidakseimbangan data (*imbalanced dataset*), di mana jumlah data pada suatu kelas jauh lebih dominan dibandingkan kelas lainnya. Kondisi ini umum terjadi pada kasus stroke, di mana jumlah pasien yang tidak mengalami stroke biasanya lebih banyak dibandingkan yang mengalami stroke. Ketidakseimbangan ini dapat menyebabkan model cenderung bias terhadap kelas mayoritas (tidak stroke) dan menurunkan kemampuan dalam mendeteksi kelas minoritas (stroke) yang justru lebih penting (He & Garcia, 2009). Oleh karena itu, diperlukan penanganan data yang tepat, seperti teknik *oversampling* atau *undersampling* untuk meningkatkan kinerja model klasifikasi.

Sejalan dengan permasalahan tersebut, penelitian ini tidak hanya menggunakan data asli yang bersifat tidak seimbang, tetapi juga menerapkan teknik penanganan ketidakseimbangan data untuk menghasilkan distribusi kelas yang lebih proporsional. Selanjutnya, dilakukan perbandingan kinerja model LVQ dalam mengklasifikasikan data pada kedua kondisi tersebut, yaitu sebelum dan sesudah penanganan *imbalanced dataset*. Pendekatan ini bertujuan menganalisis sejauh mana pengaruh penyeimbangan data terhadap performa model, khususnya dalam meningkatkan kemampuan deteksi terhadap kelas minoritas yang memiliki tingkat kepentingan lebih tinggi dalam konteks medis.

Berdasarkan latar belakang tersebut, penelitian ini bertujuan untuk mengimplementasikan algoritma *Learning Vector Quantization* dalam melakukan klasifikasi penyakit stroke menggunakan dataset yang diperoleh dari *Kaggle*. Selain itu, penelitian ini juga bertujuan untuk membandingkan kinerja model LVQ pada data yang tidak seimbang dan data yang telah dilakukan penanganan *imbalanced dataset*, serta mengevaluasi performa model dengan menggunakan metrik evaluasi yang relevan.

II. METODE PENELITIAN

A. Jenis dan Sumber Data

Data yang digunakan merupakan data sekunder yang diperoleh dari platform *Kaggle* dengan nama *Stroke Prediction Dataset* yang diunggah oleh Fedesoriano. Sumber asli pengumpulan data tidak dijelaskan secara spesifik karena bersifat *confidential source*, sehingga lokasi pengumpulan data tidak diketahui secara pasti. *Dataset* ini dapat diakses melalui (<https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset>). *Dataset* ini digunakan untuk memprediksi apakah seorang pasien berkemungkinan terkena stroke berdasarkan parameter input seperti jenis kelamin, usia, berbagai penyakit, status menikah, jenis tempat tinggal, kadar gula dalam darah, indeks massa tubuh serta jenis pekerjaan. Jumlah keseluruhan data yang digunakan sebanyak 4908 data. Penelitian ini melibatkan 9 variabel prediktor dan 1 variabel target dengan rincian variabel dapat dilihat pada Tabel 1.

Tabel 1. Variabel Penelitian

Variabel	Keterangan
(X_1) Jenis Kelamin	0 : Laki-laki 1 : Perempuan
(X_2) Usia	Kontinu
(X_3) Hipertensi	0 : Tidak 1 : Ya
(X_4) Penyakit Jantung	0 : Tidak 1 : Ya
(X_5) Pernah Menikah	0 : Tidak 1 : Ya
(X_6) Jenis Tempat Tinggal	0 : Pedesaan 1 : Perkotaan
(X_7) Kadar Gula Darah	Kontinu
(X_8) Indeks Massa Tubuh	Kontinu
(X_9) Jenis Pekerjaan	0 : Belum Bekerja 1 : Sektor Pemerintahan 2 : Sektor Swasta

	3 : Tidak Bekerja
	4 : Wiraswasta
(Y) Diagnosa Stroke	0 : Negatif
	1 : Positif

B. Teknik Analisis Data

Pada penelitian ini, metode analisis yang digunakan adalah *Learning Vector Quantization* (LVQ), yaitu algoritma *supervised learning* berbasis *prototype* yang bekerja dengan mengklasifikasikan data berdasarkan kedekatan jarak antara data input dengan vektor representatif dari masing-masing kelas. LVQ melakukan proses pembelajaran dengan memperbarui bobot secara iteratif agar dapat merepresentasikan pola dari setiap kelas dengan lebih baik (Kohonen, 2001). Tahapan analisis data pada penelitian ini adalah sebagai berikut:

1. Melakukan *preprocessing* data, yang meliputi *cleaning* data, dan transformasi data kategorikal menjadi numerik dengan teknik label *encoding*. Pada tahap transformasi, setiap kategori pada variabel kategorik diubah menjadi representasi angka agar dapat diproses oleh algoritma LVQ. Proses ini dilakukan dengan memberikan nilai numerik yang berbeda pada setiap kategori dalam suatu variabel tanpa mengubah informasi yang ada pada data.
2. *Splitting* data. Membagi *dataset* menjadi data *training* dan data *testing* menggunakan metode *holdout validation* dengan proporsi 80% untuk data *training* dan 20% untuk data *testing*. Pembagian data dilakukan menggunakan fungsi *train_test_split* pada *library scikit_learn* dengan metode pengacakan (*random splitting*) untuk menjaga distribusi data pada data *training* dan data *testing* tetap representatif. Proporsi 80:20 dipilih karena mampu memberikan jumlah data pelatihan yang cukup untuk proses pembelajaran model dan menyediakan data pengujian yang memadai untuk evaluasi performa model (Géron, 2019).
3. Normalisasi data. Normalisasi dilakukan menggunakan metode *Min-Max Normalization*, yang mengubah rentang nilai data ke interval [0,1], sehingga dapat menghindari dominasi atribut tertentu dalam perhitungan jarak.

$$X'_i = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (1)$$

Keterangan:

X'_i : nilai hasil normalisasi ke i

X_{min} : nilai minimum dari variabel

X : nilai aktual

X_{max} : nilai maksimum dari variabel

4. Penanganan *imbalanced dataset*. Untuk mengatasi ketidakseimbangan data, dilakukan *oversampling* pada kelas minoritas menggunakan metode *Synthetic Minority Over-sampling Technique* (SMOTE). Penerapan SMOTE hanya dilakukan pada data *training* agar distribusi data pada data *testing* tetap merepresentasikan kondisi data asli, sehingga evaluasi model tetap objektif.
5. Melakukan proses klasifikasi LVQ. Proses ini bertujuan untuk melatih model dalam mempelajari pola dan karakteristik data melalui proses pembaruan bobot berdasarkan kedekatan jarak antara data input dan vektor representatif tiap kelas pada data *training*. Adapun tahapan proses klasifikasi LVQ dilakukan melalui langkah-langkah berikut:
 - a. Inisialisasi nilai-nilai:
 - 1) Bobot awal (w_{ij}), yaitu nilai bobot ke-i pada fitur/variabel ke-j
 - 2) *Learning rate* (α), dengan nilai $0 < \alpha < 1$
 - 3) *MaxEpoch*, jumlah iterasi maksimum
 - 4) *Error minimum* yang diharapkan (*eps*)
 - b. Menetapkan target berupa kelas (T) dan kondisi awal *Epoch* = 0
 - c. Lakukan proses jika (*epoch* < *maxEpoch*) atau ($\alpha > eps$):
 - 1) *Epoch* = *Epoch* + 1
 - 2) Menghitung jarak antara data input dengan setiap bobot menggunakan jarak *Euclidean*

$$D_j = \sqrt{\sum_{j=1} (x - w_j)^2} \quad (2)$$

- 3) Tentukan nilai minimum dari setiap jarak kelas sehingga menjadi *output* (C_j)
- a. Perbarui bobot w_j dengan rumus:
 - 1) Jika $T = C_j$

$$2) \text{ Jika } T \neq C_j \quad w_j(\text{baru}) = w_j(\text{lama}) + \alpha[x - w_j(\text{lama})] \quad (3)$$

$$\quad \quad \quad w_j(\text{baru}) = w_j(\text{lama}) - \alpha[x - w_j(\text{lama})] \quad (4)$$

6. Mengevaluasi kinerja model LVQ menggunakan *confusion matrix* berdasarkan hasil prediksi pada data *testing*. *Confusion matrix* merupakan representasi matriks yang digunakan untuk mengevaluasi kinerja model klasifikasi dengan cara menghitung jumlah prediksi yang benar dan salah untuk setiap kelas (Tharwat, 2018). *Confusion matrix* ditunjukkan pada Tabel 2.

Tabel 2. *Confusion matrix*

Classification	Actual class	
	1	0
Predicted class	1 True Positive (TP)	0 False Positive (FP)
0	False Negative (FN)	True Negative (TN)

Keterangan:

Actual class : Kelas sebenarnya

Predicted class : Kelas yang diperoleh dari hasil prediksi pada model klasifikasi

(TP) : Jumlah data positif yang berhasil di prediksi sebagai positif

(FN) : Jumlah data positif yang salah diprediksi sebagai negatif

(FP) : Jumlah data negatif yang salah diprediksi sebagai positif

(TN) : jumlah data negatif yang berhasil diprediksi sebagai negatif

Kinerja model diukur menggunakan metrik *accuracy* dan *sensitivity* dengan rumus sebagai berikut:

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \times 100\% \quad (5)$$

$$Sensitivity = \frac{TP}{TP + FN} \times 100\% \quad (6)$$

III. HASIL DAN PEMBAHASAN

A. *Preprocessing Data*

Pada data terdapat variabel kategorik yang memiliki lebih dari 2 kategori. Hal ini dapat menyebabkan model menganggap adanya urutan atau tingkatan antar kategori, padahal sebenarnya tidak ada dan dapat mengakibatkan hasil klasifikasi kurang akurat. Oleh karena itu, digunakan *one-hot encoding*, yaitu mengubah setiap kategori menjadi kolom terpisah berisi “0” dan “1”. Sehingga model hanya mengenali kategori tersebut sebagai kategori yang berbeda, bukan sebagai nilai yang bertingkat. Hasil dari *preprocessing* data dapat dilihat pada Tabel 3.

Tabel 3. Data Awal

No	X ₁	X ₂	X ₃	X ₄	X ₅	X ₆	X ₇	X ₈	X ₉	X ₁₀	X ₁₁	X ₁₂	X ₁₃
1	1	67	0	1	1	1	228,69	36,6	0	0	1	0	0
2	1	40	0	1	1	0	105,92	32,5	0	0	1	0	0
3	0	29	0	0	1	1	171,23	34,4	0	0	1	0	0
...
4908	0	44	0	0	1	1	85,28	26,2	0	0	1	0	0

Tabel 3 menunjukkan dataset yang telah dilakukan *one-hot encoding*, di mana variabel X₉ – X₁₃ merupakan hasil *encoding* dari variabel jenis pekerjaan. Pada proses ini, setiap kategori direpresentasikan dalam bentuk variabel biner dengan nilai “1” yang menunjukkan keberadaan suatu kategori pada data, sedangkan nilai “0” menunjukkan kategori

tersebut tidak dimiliki oleh data terkait. Dengan demikian, nilai “0” dan “1” pada hasil *encoding* tidak merepresentasikan tingkatan atau skala tertentu, melainkan hanya sebagai penanda keberadaan kategori.

B. *Splitting Data*

Pembagian data dilakukan menggunakan metode *holdout*, yaitu membagi dataset menjadi data latih (*training*) dan data uji (*testing*) dengan proporsi 80% untuk data *training* dan 20% untuk data *testing*. Hasil pembagian data dapat dilihat pada Tabel 4.

Tabel 4. Pembagian Data

Kelompok Data	Persentase	Jumlah Sampel
<i>Training</i>	80%	3926
<i>Testing</i>	20%	982

Tabel 4 menunjukkan data *training* yang digunakan sebanyak 3926 sampel dan data *testing* sebanyak 982 sampel. Data *training* akan digunakan selama pelatihan model, sedangkan data *testing* akan digunakan saat pengujian atau evaluasi model.

C. *Normalisasi Data*

Selanjutnya dilakukan normalisasi dengan tujuan untuk meminimalkan potensi *error* pada algoritma dengan menyamakan skala data antar fitur/variabel, hasil dari normalisasi data ditunjukkan pada Tabel 5.

Tabel 5. Normalisasi Data

No	X ₁	X ₂	X ₃	X ₄	X ₅	X ₆	X ₇	X ₈	X ₉	X ₁₀	X ₁₁	X ₁₂	X ₁₃
1	1	0,523926	0	1	1	1	0,419583	0,259486	0	0	1	0	0
2	1	0,609375	0	1	1	0	0,224771	0,220318	0	0	0	0	1
3	1	0,108887	0	0	0	1	0,100452	0,094247	1	0	0	0	0
...
3926	0	0,682617	0	0	1	1	0,263318	0,188494	0	0	0	0	1

Tabel 5 menunjukkan nilai dari setiap variabel prediktor yang telah dinormalisasi kedalam interval [0,1] dengan jumlah sampel sebanyak 3926.

D. *Penanganan Imbalanced Dataset*

Ketidakseimbangan data (*Imbalanced dataset*) akan membuat model menjadi bias, karena model cenderung memberikan prediksi yang lebih akurat pada kelas mayoritas dibandingkan kelas minoritas. Oleh karena itu, dilakukan teknik *oversampling* menggunakan metode SMOTE untuk menghasilkan distribusi kelas yang lebih proporsional. Distribusi kelas sebelum dan sesudah dilakukan metode SMOTE dapat dilihat pada Tabel 6.

Tabel 6. Distribusi Kelas

Kelas	Jumlah Sampel Sebelum SMOTE	Jumlah Sampel Setelah SMOTE
0	3759	3759
1	167	3759

E. *Learning Vector Quantization*

Pada tahap ini dilakukan proses pemodelan klasifikasi menggunakan algoritma *Learning Vector Quantization* (LVQ) terhadap *dataset* stroke. Pemodelan dilakukan pada dua kondisi data, yaitu data asli yang terdiri atas 3926 data dan data hasil penanganan ketidakseimbangan menggunakan SMOTE dengan jumlah data sebanyak 7518 data. LVQ mengklasifikasikan data pasien berdasarkan kedekatan karakteristik pasien terhadap bobot masing-masing kelas, dengan tujuan membantu identifikasi dini kemungkinan terjadinya stroke.

Sebelum melakukan pemodelan, perlu dilakukan optimasi parameter model LVQ untuk memperoleh model dengan performa terbaik. Parameter yang diuji dalam penelitian ini adalah *learning rate* untuk menentukan besarnya perubahan bobot pada setiap iterasi, dan *max epoch* untuk menentukan jumlah iterasi maksimum selama proses pelatihan. Hal ini dikarenakan kedua parameter tersebut berpengaruh terhadap kecepatan konvergensi dan stabilitas model dalam proses pembelajaran.

Pengujian dilakukan dengan mencoba beberapa kombinasi nilai *learning rate* dan *max epoch* pada data asli maupun data hasil SMOTE. Setiap kombinasi parameter dievaluasi berdasarkan nilai akurasi yang diperoleh pada data uji. Kombinasi parameter dengan akurasi tertinggi dipilih sebagai parameter terbaik. Parameter terbaik yang diperoleh dari masing-masing *dataset* tersebut dapat dilihat pada Tabel 7.

Tabel 7. Parameter Terbaik

<i>Dataset</i>	<i>Learning Rate</i>	<i>Max Epoch</i>	<i>Accuracy</i>
Data Asli	0,3	200	0,95723
Data SMOTE	0,001	50	0,49491

Tabel 7 menunjukkan parameter terbaik yang diperoleh masing-masing *dataset*, di mana kombinasi parameter terbaik pada data asli berbeda dengan parameter terbaik pada data hasil SMOTE. Perbedaan ini menunjukkan bahwa distribusi data memengaruhi proses pembelajaran model LVQ. Pada data asli yang bersifat *imbalance*, model cenderung lebih mudah mempelajari kelas mayoritas sehingga membutuhkan proses pelatihan yang panjang untuk mencapai konvergensi. Sebaliknya, pada data hasil SMOTE dengan distribusi yang lebih seimbang, model mencapai performa terbaik dengan jumlah iterasi yang lebih sedikit dan *learning rate* yang lebih kecil. Hasil ini menunjukkan bahwa parameter optimal pada suatu kondisi data belum tentu memberikan hasil yang optimal pada kondisi data lainnya.

F. Evaluasi Model

Setelah diperoleh kombinasi parameter terbaik pada data asli dan data hasil SMOTE, dilakukan evaluasi model LVQ pada beberapa pengujian untuk menganalisis pengaruh distribusi data parameter model terhadap hasil klasifikasi stroke. Seluruh pengujian pada tahap evaluasi dilakukan menggunakan data *testing* yang sama agar perbandingan performa model dapat dilakukan secara objektif. Perbedaan pada masing-masing pengujian terletak pada data pelatihan dan kombinasi parameter yang digunakan dalam proses pembentukan model LVQ. Hasil evaluasi tersebut dapat dilihat pada *confusion matrix* berikut.

Tabel 8. Hasil *Confusion Matrix* Model LVQ yang Dilatih Menggunakan Data Asli dan Parameter Terbaik Data Asli

<i>Classification</i>	<i>Actual class</i>	
	1	0
<i>Predicted class</i>	1	0
	0	42
		940

Berdasarkan Tabel 8, model LVQ yang dilatih menggunakan data asli dengan parameter terbaik data asli menghasilkan pasien positif yang diprediksi negatif sebanyak 42 orang, pasien negatif yang diprediksi negatif sebanyak 940 orang, dan tidak ada pasien positif yang di prediksi positif serta pasien negatif yang diprediksi positif. Menggunakan persamaan (5) dan (6), diperoleh hasil akurasi sebesar 95,72% dan *sensitivity* sebesar 0%.

Tabel 9. Hasil *Confusion Matrix* Model LVQ yang Dilatih Menggunakan Data Hasil SMOTE dengan Parameter Terbaik Data Hasil Smote

<i>Classification</i>	<i>Actual class</i>	
	1	0
<i>Predicted class</i>	1	38
	0	4
		492
		448

Berdasarkan Tabel 9, model LVQ yang dilatih menggunakan data hasil SMOTE dengan parameter terbaik data hasil SMOTE menghasilkan pasien positif yang diprediksi negatif sebanyak 4 orang, pasien negatif yang diprediksi negatif sebanyak 448 orang, pasien positif yang di prediksi positif sebanyak 38 orang, dan pasien negatif yang diprediksi positif sebanyak 492 orang. Menggunakan persamaan (5) dan (6), diperoleh hasil akurasi sebesar 49,49% dan *sensitivity* sebesar 90%.

Berdasarkan hasil evaluasi, model LVQ yang dilatih menggunakan data hasil SMOTE dengan parameter terbaik dari data hasil SMOTE menunjukkan performa yang lebih baik dibandingkan model pada data asli dalam konteks klasifikasi medis. Meskipun memiliki akurasi yang lebih rendah, model ini mampu mendeteksi sebagian besar kasus stroke dengan nilai *sensitivity* sebesar 90%, sehingga lebih efektif dalam melakukan *screening* untuk meminimalkan

risiko *false negatif*. Selanjutnya dilakukan pengujian tambahan menggunakan model yang dilatih pada data hasil SMOTE dengan parameter terbaik data asli untuk melihat sejauh mana model LVQ bergantung pada parameter.

Tabel 10. Hasil *Confusion Matrix* Model LVQ yang Dilatih Menggunakan Data Hasil SMOTE dengan Parameter Terbaik Data Asli

Classification	Actual class	
	1	0
Predicted class	1	940
	0	0

Berdasarkan Tabel 10, model LVQ yang dilatih menggunakan data hasil SMOTE dengan parameter terbaik data asli menghasilkan pasien positif yang diprediksi positif sebanyak 42 orang, pasien negatif yang diprediksi positif sebanyak 940 orang, dan tidak ada pasien positif yang di prediksi negatif, serta pasien negatif yang diprediksi negatif. Menggunakan persamaan (5) dan (6), diperoleh hasil akurasi sebesar 4,27% dan *sensitivity* sebesar 100%.

Hasil pengujian menunjukkan bahwa model mengalami penurunan performa yang signifikan, di mana model tersebut memprediksi seluruh data sebagai kelas stroke. Hal ini membuat seluruh data stroke berhasil diprediksi, namun seluruh data non-stroke diprediksi sebagai stroke, sehingga menghasilkan *false positif* yang sangat tinggi dan menurunkan akurasi secara drastis. Kondisi ini menunjukkan bahwa model LVQ memiliki ketergantungan yang cukup tinggi terhadap parameter dan distribusi data yang digunakan dalam proses pelatihan.

IV. KESIMPULAN

Berdasarkan hasil penelitian, algoritma LVQ menunjukkan performa kurang optimal dalam melakukan klasifikasi penyakit stroke pada dataset yang digunakan. Pada data asli yang bersifat *imbalanced*, model menghasilkan akurasi yang tinggi sebesar 95,72%, namun tidak mampu mendeteksi kelas minoritas (stroke) sehingga menunjukkan bias terhadap kelas mayoritas (tidak stroke) yang mengakibatkan tingginya jumlah *false negative*.

Setelah dilakukan penanganan data *imbalanced* menggunakan metode SMOTE, kemampuan model dalam mendeteksi kelas stroke meningkat dengan nilai *sensitivity* sebesar 90%. Namun, peningkatan tersebut diikuti oleh penurunan akurasi menjadi 49,49% dan tingginya jumlah *false positive*. Hasil penelitian juga menunjukkan bahwa parameter model sangat dipengaruhi oleh distribusi data, sehingga parameter terbaik pada data asli tidak dapat diterapkan secara langsung pada data hasil SMOTE. Dengan demikian, algoritma LVQ dinilai kurang cocok untuk klasifikasi stroke pada dataset ini karena belum mampu memberikan performa yang seimbang antara sensitivitas dan ketepatan klasifikasi (akurasi).

Untuk penelitian selanjutnya, disarankan untuk menggunakan algoritma klasifikasi lain yang lebih kuat terhadap data tidak seimbang. Selain itu, metode penanganan lain dapat diterapkan untuk meningkatkan performa model. Penggunaan *dataset* yang lebih besar serta optimasi parameter yang lebih luas juga perlu dilakukan agar hasil klasifikasi menjadi lebih optimal.

DAFTAR PUSTAKA

- Feigin, V. L., Forouzanfar, M. H., Krishnamurthi, R., Mensah, G. A., Connor, M., Bennett, D. A., Moran, A. E., Sacco, R. L., Anderson, L., Truelsen, T., O'Donnell, M., Venketasubramanian, N., Barker-Collo, S., Lawes, C. M. M., Wang, W., Shinohara, Y., Witt, E., Ezzati, M., Naghavi, M., & Murray, C. (2014). Global and regional burden of stroke during 1990-2010: findings from the Global Burden of Disease Study 2010. *Lancet (London, England)*, 383(9913), 245–254. [https://doi.org/10.1016/s0140-6736\(13\)61953-4](https://doi.org/10.1016/s0140-6736(13)61953-4)
- Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow* (2nd Editio). O'Reilly Media.
- Han, J., Kamber, M., & Pei, J. (2012). Data Mining : Concepts and Techniques. In *Data Mining : Concepts and Techniques*. <https://doi.org/https://doi.org/10.1016/C2009-0-61819-5>
- He, H., & Garcia, E. A. (2009). Learning from Imbalanced Data. *Knowledge and Data Engineering, IEEE Transactions On*, 21(9), 1263–1284. <https://doi.org/10.1109/TKDE.2008.239>
- Johnson, C. O., Nguyen, M., Roth, G. A., Nichols, E., & Alam, T. (2019). Global , regional , and national burden of

- stroke , 1990 – 2016 : a systematic analysis for the Global Burden of Disease Study 2016 Bill, F., & Foundation, M. G. (2016). Global , regional , and national burden of stroke , 1990 – 2016 : a systematic analysi. *The Lancet Neurologi*, 18(5), 439–458. <https://doi.org/http://dx.doi.org/10.1016/S1474-4422>
- Kaden, M., Lange, M., Nebel, D., Riedel, M., Geweniger, T., & Villmann, T. (2015). Aspects in Classification Learning - Review of Recent Developments in Learning Vector Quantization. *Foundations of Computing and Decision Sciences*, 39(2), 79–105. <https://doi.org/10.2478/fcds-2014-0006>
- Katan, M., & Luft, A. (2018). Global Burden of Stroke. *Thieme Medical Publishers*, 38(2), 208–211. <https://doi.org/https://doi.org/10.1055/s-0038-1649503>
- Kohonen, T. (2001a). Learning Vector Quantization. In *Self-Organizing Maps* (pp. 245–261). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-56927-2_6
- Kohonen, T. (2001b). *Self-Organizing Maps*. Springer.
- Melin, P., Amezcua, J., Valdez, F., & Castillo, O. (2014). A new neural network model based on the LVQ algorithm for multi-class classification of arrhythmias. *Information Sciences*, 279, 483–497. <https://doi.org/https://doi.org/10.1016/j.ins.2014.04.003>
- Obermeyer, Z., & Emanuel, E. J. (2016). Predicting the Future — Big Data, Machine Learning, and Clinical Medicine. *The New England Journal of Medicine* 375, 375(13), 1216–1219. <https://doi.org/10.1056/NEJMp1606181>
- Sianipar, E. A., & Yasin, M. (2025). Optimasi Klasifikasi Penyakit Diabetes Dengan Algoritma Learning Vector Quantization (LVQ). *Journal of Computer Science and Informatics Engineering*, 04(2), 72–84. <https://doi.org/https://doi.org/10.55537/cosie.v4i2.1121>
- Tharwat, A. (2018). Classification Assessment Methods. *Applied Computing and Informatics*, 17(1), 168–192. <https://doi.org/10.1016/j.aci.2018.08.003>
- Villmann, T., Bohnsack, A., & Kaden, M. (2017). Can Learning Vector Quantization be an Alternative to SVM and Deep Learning? *Journal of Artificial Intelligence and Soft Computing Research*, 7(1), 65–81.
- Zabihullah, F., Budianita, E., Syafria, F., & Afrianty, I. (2025). IMPLEMENTASI LEARNING VECTOR QUANTIZATION 2 DAN INFORMATION GAIN UNTUK KLASIFIKASI PENYAKIT GINJAL KRONIS. *Information System Journal*, 8(1), 40–50. <https://doi.org/https://doi.org/10.24076/infosjournal.2025v8i01.2118>