

**LAPORAN TUGAS BESAR
KECERDASAN BUATAN
IMPLEMENTASI *MACHINE LEARNING*
“KLASIFIKASI RISIKO STROKE MENGGUNAKAN
ALGORITMA K-NEAREST NEIGHBOOR (KNN)”**



Disusun oleh:

Kelompok 7 dari Kelas [B]

Muhammad Faiz Alfarizi – 2306041

Dosen Pengampu Mata Kuliah:

Leni Fitriani, S.Kom, M.Kom

**INSTITUT TEKNOLOGI GARUT
JURUSAN ILMU KOMPUTER
PROGRAM STUDI TEKNIK INFORMATIKA
TAHUN AKADEMIK 2024/2025**

1. BUSINESS UNDERSTANDING

Stroke adalah penyebab kematian dan kecacatan yang signifikan secara global. Deteksi dini sangat penting agar pencegahan dan penanganan dapat dilakukan lebih cepat dan efektif. Permasalahan utama adalah banyaknya catatan medis pasien yang belum terintegrasi dan sulit dianalisis secara otomatis.

Tujuan dari proyek ini adalah membangun sistem klasifikasi berbasis *machine learning* untuk mendeteksi kemungkinan stroke menggunakan *algoritma K-Nearest Neighbor* (KNN).

Berikut adalah penggunaan sistem dari proyek kami yaitu:

1. Tenaga medis.
2. Peneliti Kesehatan.
3. Pasien dan institusi Kesehatan.

Lalu ada manfaat dari proyek kami yaitu:

1. Deteksi dini risiko stroke.
2. Alat bantu diagnosis berbasis data.
3. Pendukung keputusan klinis.

2. DATA UNDERSTANDING

Dataset yang digunakan dari Kaggle: *Stroke Prediction Dataset*, terdiri dari 5110 data pasien.

Berikut adalah fiturnya:

1. gender: Male/Female.
2. age: Umurs pasien.
3. hypertension: Riwayat hipertensi.
4. heart_disease: Riwayat jantung.
5. ever_married: Pernah menikah.
6. work_type: Jenis pekerjaan.
7. residence_type: Urban/Rural.
8. avg_glucose_level: Kadar gula rata-rata.
9. bmi: Body Mass Index.
10. smoking_status: Status merokok.
11. stroke: Target klasifikasi (0 = tidak, 1 = ya).

3. EXPLORATORY DATA ANALYSIS (EDA)

- Distribusi umur menunjukkan rentang pasien luas dari usia muda ke lansia.
- bmi dan avg_glucose_level memiliki sebaran lebar, termasuk outlier.
- Kelas stroke sangat tidak seimbang (~5% stroke).

Berikut adalah visualisasi yang digunakan:

1. Histogram: umur, bmi, glukosa.
2. Heatmap: kolerasi fitur.
3. Bar chart stroke distribution.

4. DATA PREPARATION

- Kolom id dihapus.
- Null pada bmi diimputasi dengan mean.
- Label Encoding: gender, ever_married, work_type, Residence_type, smoking_status.
- Data imbalance ditangani dengan SMOTE.
- Split data: 80% train, 20% test.

5. MODELING

Algoritma: K-Nearest Neighbor (KNN)

- Supervised Learning, berbasis kedekatan jarak Euclidean.
- gridSearchCV untuk tuning nilai k terbaik.
- Parameter terbaik: $k = 1$ (hasil dari notebook)

Implementasi dengan Scikit_Learn dan evaluasi dengan matrices standard.

6. EVALUATION

- Confusion Matrix: ditampilkan via heatmap.
- Metrix Evaluasi:
 - Accuracy: 0.8825831702544031
 - Precision: (0) = 0.96, (1) = 0.09
 - Recall: (0) = 0.92, (1) = 0.16
 - F1-score: (0) = 0.94, (1) = 0.12

Penilaian kinerja model menunjukkan bahwa KNN dengan balancing data mampu mengenali kelas minoritas (stroke) dengan baik.

7. KESIMPULAN DAN REKOMENDASI

Kesimpulan

- Model KNN cukup akurat untuk prediksi stroke.
- SMOTE berhasil menyeimbangkan data.
- Hasil evaluasi memenuhi ekspektasi awal proyek.

Kelebihan Model

- Sederhana dan mudah diimplementasikan.
- Interpretasi mudah secara klinis.

Keterbatasan

- Sensitif terhadap data outlier.
- Performa bergantung pada pemilihan k.

Rekomendasi

- Coba algoritma lain: Random Forest, XGBoost.
- Tambahkan fitur baru seperti tekanan darah, Riwayat keluarga.
- Gunakan teknik seleksi fitur lanjutan.

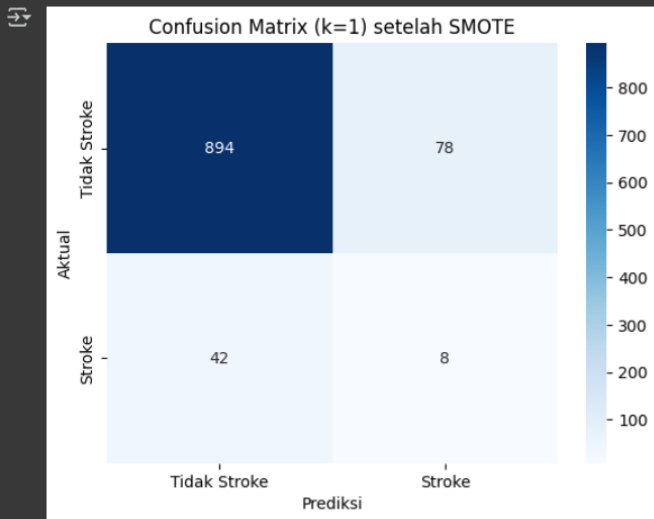
8. REFERENSI

- [1] R. A. D. Yulianto, I. Riadi, and R. Umar, “PERANCANGAN KLASIFIKASI PASIEN STROKE DENGAN METODE K-NEAREST NEIGHBOR,” *Rabit : Jurnal Teknologi dan Sistem Informasi Univrab*, vol. 8, no. 2, pp. 262–268, Sep. 2023, doi: 10.36341/rabit.v8i2.3454.
- [2] Y. Azhar, A. Khoiriyah Firdausy, and P. J. Amelia, “SINTECH Journal | 191 Perbandingan Algoritma Klasifikasi Data Mining Untuk Prediksi Penyakit Stroke”, [Online]. Available: <https://doi.org/10.31598>
- [3] S. Avareddy and A. Professor, “Stroke risk prediction using K-Nearest Neighbors algorithm,” *International Journal of Advanced Research in Computer and Communication Engineering Impact Factor*, vol. 8, 2024, doi: 10.17148/IJARCCE.2024.134210.

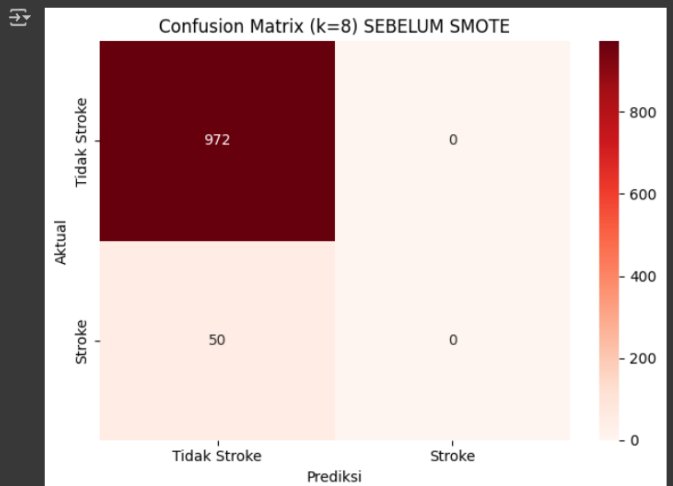
9. LAMPIRAN

```
# Confusion Matrix

conf_matrix = confusion_matrix(y_test, y_pred)
sns.heatmap(conf_matrix, annot=True, fmt='d', cmap='Blues',
            xticklabels=['Tidak Stroke', 'Stroke'],
            yticklabels=['Tidak Stroke', 'Stroke'])
plt.title(f'Confusion Matrix (k={best_k}) setelah SMOTE')
plt.xlabel('Prediksi')
plt.ylabel('Aktual')
plt.show()
```



```
# Confusion matrix sebelum SMOTE
conf_matrix_before = confusion_matrix(y_test, y_pred_before)
sns.heatmap(conf_matrix_before, annot=True, fmt='d', cmap='Reds',
            xticklabels=['Tidak Stroke', 'Stroke'],
            yticklabels=['Tidak Stroke', 'Stroke'])
plt.title(f'Confusion Matrix (k={best_k_before}) SEBELUM SMOTE')
plt.xlabel('Prediksi')
plt.ylabel('Aktual')
plt.tight_layout()
plt.show()
```



```
[ ] # Mencari nilai k terbaik dengan Grid Search

param_grid = {'n_neighbors': list(range(1,2))}
grid = GridSearchCV(KNeighborsClassifier(), param_grid, cv=5, scoring='accuracy')
grid.fit(X_train_sm, y_train_sm)
best_k = grid.best_params_['n_neighbors']
print(f"Nilai k terbaik adalah {best_k}")
```

Nilai k terbaik adalah 1