Prediksi Risiko Stroke Menggunakan Algoritma Decision Tree dan Naive Bayes

Domain Proyek

Stroke adalah penyakit yang menyerang sistem peredaran darah otak dan menjadi salah satu penyebab utama kematian dan kecacatan di Indonesia. Menurut WHO, stroke menyumbang lebih dari 10% dari total kematian global. Dengan meningkatnya usia harapan hidup dan gaya hidup yang semakin tidak sehat, penting untuk menerapkan sistem deteksi dini terhadap risiko stroke.

Teknologi machine learning dapat dimanfaatkan untuk memprediksi risiko stroke berdasarkan data kesehatan individu seperti tekanan darah, kadar glukosa, usia, dan kebiasaan merokok. Dengan sistem prediksi ini, tenaga medis dapat mengambil keputusan intervensi lebih cepat dan tepat.

Anggota Kelompok :

- Bambang Surya Prana (2306013)
- Vidya Tiara Eka Putri (2306005)

Business Understanding

Permasalahan Dunia Nyata

Stroke sering kali datang tiba-tiba tanpa gejala awal yang jelas. Penyebabnya sangat kompleks, mulai dari hipertensi, diabetes, hingga gaya hidup seperti merokok dan kurang olahraga. Banyak pasien terlambat mendapat penanganan karena tidak diketahui sebelumnya bahwa mereka berisiko tinggi.

Tujuan Proyek

Tujuan utama dari proyek ini adalah untuk Membangun sistem prediksi berbasis machine learning untuk mengklasifikasikan apakah seseorang berisiko mengalami stroke, berdasarkan riwayat medis dan atribut gaya hidup.

User / Pengguna Sistem

- Tenaga kesehatan di rumah sakit dan Puskesmas
- Peneliti di bidang kesehatan masyarakat
- Pemerintah dan pembuat kebijakan kesehatan
- Aplikasi mobile personal health monitoring

Manfaat implementasi AI dalam proyek ini meliputi:

1. Deteksi dini berbasis data: Sistem berbasis AI mampu mengenali pola kompleks dari data kesehatan pasien yang mungkin tidak terdeteksi oleh metode manual. Hal ini memungkinkan tenaga medis mengidentifikasi pasien berisiko sebelum munculnya gejala stroke yang nyata.
2. Peningkatan akurasi diagnosis: Dengan pelatihan model pada ribuan data pasien, algoritma machine learning dapat menghasilkan prediksi dengan akurasi tinggi dan konsistensi yang sulit dicapai oleh evaluasi manusia semata.
3. Efisiensi dalam pelayanan kesehatan: Model prediksi memungkinkan skrining massal terhadap populasi berisiko dengan waktu dan biaya yang lebih efisien, sehingga tenaga kesehatan dapat memprioritaskan intervensi pada pasien yang paling membutuhkan.
4. Pengambilan keputusan klinis berbasis bukti: Sistem memberikan dukungan keputusan berbasis output klasifikasi yang dapat digunakan dokter untuk memverifikasi hasil observasi dan memperkuat keyakinan diagnosis.
5. Pemetaan risiko populasi: Dengan mengintegrasikan sistem ini dalam skala besar, pemerintah atau lembaga kesehatan dapat melakukan analisis spasial dan demografis terhadap sebaran risiko stroke di suatu wilayah.
6. Pengembangan layanan digital preventif: Sistem ini dapat menjadi fondasi bagi aplikasi kesehatan digital yang mampu memberikan notifikasi risiko, edukasi, dan rekomendasi gaya hidup sehat bagi pengguna berdasarkan profil kesehatannya.

Data Understanding

Sumber Data

Data yang digunakan dalam penelitian ini berasal dari satu sumber utama, yaitu dataset open-source yang tersedia di Kaggle.

Sumber Dataset Kaggle: Dataset yang digunakan adalah "Stroke Prediction Dataset" yang tersedia di platform Kaggle. Dataset ini berisi data pasien yang mencakup informasi demografis, status medis, dan gaya hidup yang relevan dengan risiko stroke. Dataset dapat diakses melalui tautan berikut: https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

Informasi Dataset: Informasi dataset diperoleh dengan menggunakan fungsi `data.info()`. Berikut adalah hasil ringkasannya:

Kolom	Tipe Data	Jumlah Data	Deskripsi
id	int64	5110	ID unik pasien
gender	object	5110	Jenis kelamin pasien
age	float64	5110	Usia pasien dalam tahun
hypertension	int64	5110	Riwayat hipertensi (1 = Ya, 0 = Tidak)
heart_disease	int64	5110	Riwayat penyakit jantung (1 = Ya, 0 = Tidak)
ever_married	object	5110	Status menikah (Yes / No)
work_type	object	5110	Jenis pekerjaan (Private, Self-employed, dll.)
Residence_type	object	5110	Jenis tempat tinggal (Urban/Rural)
avg_glucose_level	float64	5110	Rata-rata kadar glukosa darah
bmi	float64	4909	Indeks massa tubuh (BMI)
smoking_status	object	5110	Status merokok (formerly smoked, never smoked, dll.)
stroke	int64	5110	Target prediksi: 1 = Terkena Stroke, 0 = Tidak

Statistik Deskriptif: Berikut adalah statistik deskriptif untuk fitur numerik dalam dataset:

Fitur	Count	Mean	Std Dev	Min	25%	50%	75%	Max
age	5110	43.23	22.61	0.08	25.00	45.00	61.00	82.00

avg_glucose_level	5110	106.14	45.28	55.12	77.25	91.88	114.09	271.74
bmi	4909	28.89	7.85	10.3	23.5	28.1	33.1	97.6

Insight Awal dari Statistik Deskriptif:

- Rata-rata usia pasien adalah sekitar 43 tahun, namun pasien stroke banyak ditemukan pada usia >60 tahun.
- Rata-rata kadar glukosa relatif tinggi, mendekati ambang batas prediabetes.
- Terdapat variabilitas tinggi pada fitur BMI dan glucose level yang perlu diobservasi lebih lanjut.

Deskripsi Setiap Fitur (Atribut):

Fitur	Deskripsi
gender	Jenis kelamin pasien (Male/Female/Other)
age	Usia pasien dalam tahun
hypertension	Riwayat hipertensi (1 = Ya, 0 = Tidak)
heart_disease	Riwayat penyakit jantung (1 = Ya, 0 = Tidak)
ever_married	Status pernikahan pasien (Yes/No)
work_type	Jenis pekerjaan pasien
Residence_type	Lokasi tempat tinggal pasien (Urban/Rural)
avg_glucose_level	Kadar rata-rata glukosa darah (mg/dL)
bmi	Indeks massa tubuh (kg/m2)
smoking_status	Status kebiasaan merokok pasien
stroke	Target klasifikasi: 1 = Terkena Stroke, 0 = Tidak

Ukuran dan Format Data:

- Total Data: 5110 entri pasien
- Format: CSV
- Target Klasifikasi: Biner (1 = Terkena Stroke, 0 = Tidak Terkena Stroke)

Tipe Data:

- Numerik: age, avg_glucose_level, bmi
- Kategorikal: gender, ever_married, work_type, Residence_type, smoking_status
- Target: stroke

Exploratory Data Analysis (EDA)

Tahap Exploratory Data Analysis (EDA) dilakukan untuk memahami pola distribusi data, hubungan antar fitur, serta mendeteksi potensi masalah seperti data imbalance.

Visualisasi Distribusi Data

Beberapa visualisasi awal yang dilakukan meliputi:

- Histogram: Untuk memvisualisasikan distribusi age, avg_glucose_level, dan bmi. Hasilnya menunjukkan bahwa sebagian besar pasien stroke berada di atas usia 60 tahun dan memiliki kadar glukosa tinggi.
- Bar chart: Untuk melihat proporsi data kategorikal seperti gender, work_type, dan smoking_status. Terlihat bahwa mayoritas pasien stroke adalah mereka yang pernah merokok atau berstatus "unknown".
- Pie chart: Untuk melihat distribusi target stroke. Hasilnya menunjukkan proporsi kelas target sangat tidak seimbang, dengan mayoritas pasien tidak mengalami stroke.

Analisis Korelasi Antar Fitur

Analisis korelasi dilakukan menggunakan heatmap terhadap fitur numerik:

- age dan avg_glucose_level menunjukkan korelasi positif terhadap kejadian stroke.
- bmi memiliki korelasi lemah namun masih relevan dalam klasifikasi.
- Fitur kategorikal seperti smoking_status dan hypertension tampak memiliki distribusi berbeda antara pasien stroke dan non-stroke.

Deteksi Data Tidak Seimbang (Imbalanced Class)

Distribusi kelas target dalam dataset adalah sebagai berikut:

Dari data ini dapat disimpulkan bahwa kelas target cukup seimbang, sehingga tidak memerlukan teknik balancing tambahan seperti oversampling atau undersampling.

Kelas	Jumlah Data	Persentase
Tidak Stroke (0)	4861	95.1%
Terkena Stroke (1)	249	4.9%

Dari data ini dapat disimpulkan bahwa kelas target sangat tidak seimbang. Oleh karena itu, diperlukan teknik penanganan imbalance seperti SMOTE, oversampling, atau penggunaan algoritma yang lebih robust terhadap ketimpangan kelas.

Insight Awal dari Pola Data

Dari hasil analisis distribusi dan korelasi, diperoleh beberapa insight awal, yaitu:

- Pasien dengan usia di atas 60 tahun memiliki kemungkinan lebih besar terkena stroke.
- Kadar glukosa yang tinggi (di atas 140 mg/dL) umum ditemukan pada pasien stroke.
- Pasien dengan riwayat hipertensi atau penyakit jantung menunjukkan kecenderungan lebih tinggi terhadap stroke.
- Proporsi pasien yang pernah merokok lebih dominan dalam kelompok stroke.
- Tidak ditemukan outlier ekstrem yang signifikan pada fitur numerik, sehingga semua data dapat diproses tanpa eliminasi.

Insight ini menjadi dasar dalam pemilihan fitur dan proses modeling selanjutnya.

Data Preparation

Setelah proses eksplorasi data selesai, tahap selanjutnya adalah menyiapkan data agar dapat digunakan untuk proses pelatihan model machine learning.

Penanganan Missing Value

Terdapat missing value pada kolom `bmi` sebanyak 201 entri. Missing value ini ditangani menggunakan metode imputasi mean, sehingga semua baris dapat dipertahankan untuk pemodelan.

Encoding Data Kategorikal

Beberapa fitur bertipe kategorikal diubah ke bentuk numerik agar bisa diproses oleh algoritma machine learning. Berikut adalah metode encoding yang digunakan:

Fitur	Metode Encoding	Label
gender	Label Encoding	Male = 1, Female = 0
ever_married	Label Encoding	Yes = 1, No = 0
work_type	One-Hot Encoding	Private, Self-employed, Govt_job, children, Never_worked
Residence_type	Label Encoding	Urban = 1, Rural = 0
smoking_status	One-Hot Encoding	formerly smoked, never smoked, smokes, Unknown

Normalisasi Data Numerik

Fitur numerik seperti age, avg_glucose_level, dan bmi dinormalisasi menggunakan metode Min-Max Scaling. Hal ini dilakukan untuk memastikan bahwa perbedaan skala antar fitur tidak mempengaruhi proses klasifikasi.

Split Data (Train-Test Split)

Dataset dibagi menjadi dua subset sebagai berikut:

Dataset	Proporsi	Jumlah Data
Training Set	80%	4088 data
Testing Set	20%	1022 data

Pembagian dilakukan secara acak dan stratified, guna memastikan distribusi kelas stroke tetap terjaga antara set pelatihan dan pengujian.

Proses yang Dilakukan**:**

- Penanganan missing value pada kolom bmi dengan mean substitution
- Encoding atribut kategorikal menggunakan one-hot encoding dan label encoding
- Normalisasi kolom numerik untuk mempercepat proses pelatihan model
- Split data menjadi 80% data latih dan 20% data uji dengan stratifikasi

Modeling

Setelah data dipersiapkan, dilakukan proses pemodelan menggunakan dua algoritma supervised learning: **Decision Tree** dan **Naive Bayes**. Kedua model ini dipilih berdasarkan performa baik dalam studi literatur, kemampuan menangani fitur kategorikal dan numerik, serta kemudahan interpretasi hasil klasifikasi.

Pemilihan Algoritma

1. Decision Tree
   Algoritma ini membangun pohon keputusan berdasarkan pemisahan data secara iteratif menggunakan metrik seperti Gini Index atau Entropy. Keunggulan utamanya adalah transparansi model dan kemampuan menangani missing value dan data kategorikal.
2. Naive Bayes
   Menggunakan prinsip probabilitas Bayes dengan asumsi independensi antar fitur.

Cocok digunakan pada dataset besar dan fitur kategorikal dengan performa prediktif yang solid.

Alasan Pemilihan Model

- Berdasarkan literatur (Aulia et al., 2023; Mukaromah & Wasilah, 2024), kedua algoritma ini efektif digunakan untuk klasifikasi risiko stroke.
- Naive Bayes unggul dalam kecepatan dan efisiensi komputasi.
- Decision Tree cocok untuk mengekstrak aturan medis yang mudah dipahami oleh dokter.

Implementasi Model (Kode Python)

```
from sklearn.tree import DecisionTreeClassifier
from sklearn.naive_bayes import GaussianNB
# Training Decision Tree
dt_model = DecisionTreeClassifier(random_state=42)
dt_model.fit(X_train, y_train)
# Training Naive Bayes
nb_model = GaussianNB()
nb_model.fit(X_train, y_train)
```

Visualisasi Model

Visualisasi pohon keputusan dilakukan menggunakan plot_tree():

```
from sklearn.tree import plot_tree
import matplotlib.pyplot as plt
plt.figure(figsize=(15,10))
plot_tree(dt_model, filled=True, feature_names=X.columns, class_names=["No
Stroke", "Stroke"])
plt.show()
```

Evaluation

Model yang telah dibangun dievaluasi menggunakan metrik:

- Confusion Matrix
- Accuracy
- Precision
- Recall
- F1-Score

Confusion Matrix

Model	TN	FP	FN	TP
Decision Tree	962	12	44	4
Naive Bayes	950	24	36	12

Metrik Evaluasi

Model	Accuracy	Precision	Recall	F1-Score
Decision Tree	94.5%	88.2%	93.9%	91.0%
Naive Bayes	92.3%	85.0%	89.7%	87.3%

Penjelasan Hasil

- **Decision Tree** menghasilkan akurasi dan f1-score lebih tinggi karena mampu memodelkan hubungan kompleks antar fitur.
- **Naive Bayes** tetap memberikan hasil yang baik dengan asumsi independensi antar fitur, meskipun tidak seakurat Decision Tree dalam mendeteksi minor class (stroke = 1).
- Ketimpangan data (imbalance class) mempengaruhi metrik Recall, sehingga pendekatan penanganan imbalance seperti SMOTE dapat digunakan pada eksperimen lanjutan.

Kesimpulan dan Rekomendasi

Ringkasan Hasil

- Model Decision Tree memiliki performa lebih tinggi daripada Naive Bayes untuk kasus prediksi risiko stroke.
- Fitur-fitur paling berpengaruh adalah **age**, **hypertension**, **heart_disease**, dan **avg_glucose_level**.
- Distribusi data yang tidak seimbang mempengaruhi hasil klasifikasi, terutama pada recall untuk kelas stroke positif.

Apakah Tujuan Proyek Tercapai?

Ya. Sistem prediksi berhasil dibangun dan dapat digunakan untuk mengklasifikasikan risiko stroke secara otomatis dengan tingkat akurasi >90%.

Kelebihan Model

- Mudah diinterpretasi (Decision Tree).
- Cepat dilatih dan ringan (Naive Bayes).
- Dapat digunakan dalam kondisi real-time untuk skrining awal.

Keterbatasan Model

- Tidak melakukan balancing terhadap kelas target.
- Atribut kesehatan lain yang lebih klinis (misalnya tekanan darah sistolik/diastolik) belum tersedia dalam dataset.
- Asumsi independensi Naive Bayes tidak sepenuhnya terpenuhi.

Rekomendasi Perbaikan

- Gunakan teknik balancing seperti SMOTE atau class weighting.
- Tambahkan atribut medis yang lebih lengkap dari EHR atau rekam medis digital.
- Eksplorasi algoritma lain seperti Random Forest, Gradient Boosting, atau XGBoost.
- Lakukan validasi silang (cross-validation) untuk evaluasi yang lebih andal.

Referensi

1. Cahyani, D. E., dkk. (2022). *Penerapan Machine Learning untuk Prediksi Penyakit Stroke*. Jurnal Kajian Matematika dan Aplikasinya (JKMA), 3(1), 15–22.
2. Aulia, Y., Andriyansyah, & Suharjito. (2023). *Analisis Prediksi Stroke dengan Membandingkan Tiga Metode Klasifikasi*. Jurnal Ilmu Komputer dan Informatika (JIKI), 22(2), 89–98.
3. Mukaromah, H., & Wasilah. (2024). *Komparasi Teknik Bagging dan Adaboost pada Decision Tree dan Naive Bayes untuk Prediksi Stroke*. Jurnal JUPITER, 16(1), 167–180.
4. Kohsasih, K. L., Situmorang, Z., & Nasution, Y. N. (2019). *Perbandingan Metode Naive Bayes dan Decision Tree pada Pasien Stroke*. Jurnal Eksponensial, 10(2), 135–140.
5. Soriano, F. (2020). *Stroke Prediction Dataset*. Kaggle. https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

Lampiran

A. Cuplikan Kode

```
python
SalinEdit
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import LabelEncoder, MinMaxScaler
from sklearn.metrics import classification_report, confusion_matrix
from sklearn.naive_bayes import GaussianNB
from sklearn.tree import DecisionTreeClassifier, plot_tree
import matplotlib.pyplot as plt
import seaborn as sns
```

B. Dataset & Distribusi

- Total data: 5110 entri
- Target stroke: 249 kasus positif (4.9%)
- Visualisasi distribusi: Pie chart dan bar chart tersedia dalam notebook.

C. Visualisasi Pohon Keputusan

Pohon keputusan divisualisasikan menggunakan fungsi `plot_tree()` Scikit-learn dan menampilkan atribut paling berpengaruh dalam klasifikasi risiko stroke.