BINUS UNIVERSITY
SCHOOL OF COMPUTER SCIENCE

GROUP 4

PROJECT BIG DATA PROCESSING

Ferlie Hernata - 2702231262
Giovincent Ricel's Tanoto - 2702226786
Josh Nicholas Sutanto - 2702234825
Nikolaus Marvin Liayasa - 2702233702
Rendy Riady - 2702234421
Matthew Ethan Laurent - 2702231496

BINUS UNIVERSITY
SCHOOL OF COMPUTER SCIENCE

GROUP 4

PERBANDINGAN KINERJA MODEL RANDOM FOREST DAN NAIVE BAYES DALAM PREDIKSI STROKE

LATAR BELAKANG

STROKE

Stroke merupakan salah satu penyakit tidak menular yang terjadi akibat adanya gangguan aliran darah ke otak, baik karena penyumbatan (iskemik) maupun pecahnya pembuluh darah (hemoragik). Kondisi ini mengakibatkan terganggunya suplai oksigen dan nutrisi ke jaringan otak, sehingga sel-sel otak mulai mati dalam hitungan menit.

INDONESIA

Di Indonesia, prevalensi stroke terus meningkat setiap tahunnya. Berdasarkan Riset Riskesdas 2018, prevalensi stroke nasional tercatat sebesar 10,9 per mil dan menunjukkan tren kenaikan dibandingkan tahun-tahun sebelumnya. Pencegahan dan deteksi dini terhadap risiko stroke menjadi penting dalam menekan angka kejadian dan kematian.

PERAN AI

Dengan berkembangnya teknologi, khususnya dalam bidang kecerdasan buatan (Artificial Intelligence) dan big data, analisis prediktif kini dimanfaatkan untuk membantu deteksi awal penyakit seperti stroke. Dalam proses prediksi, pemilihan model klasifikasi yang tepat sangat menentukan akurasi dan keandalan hasil.

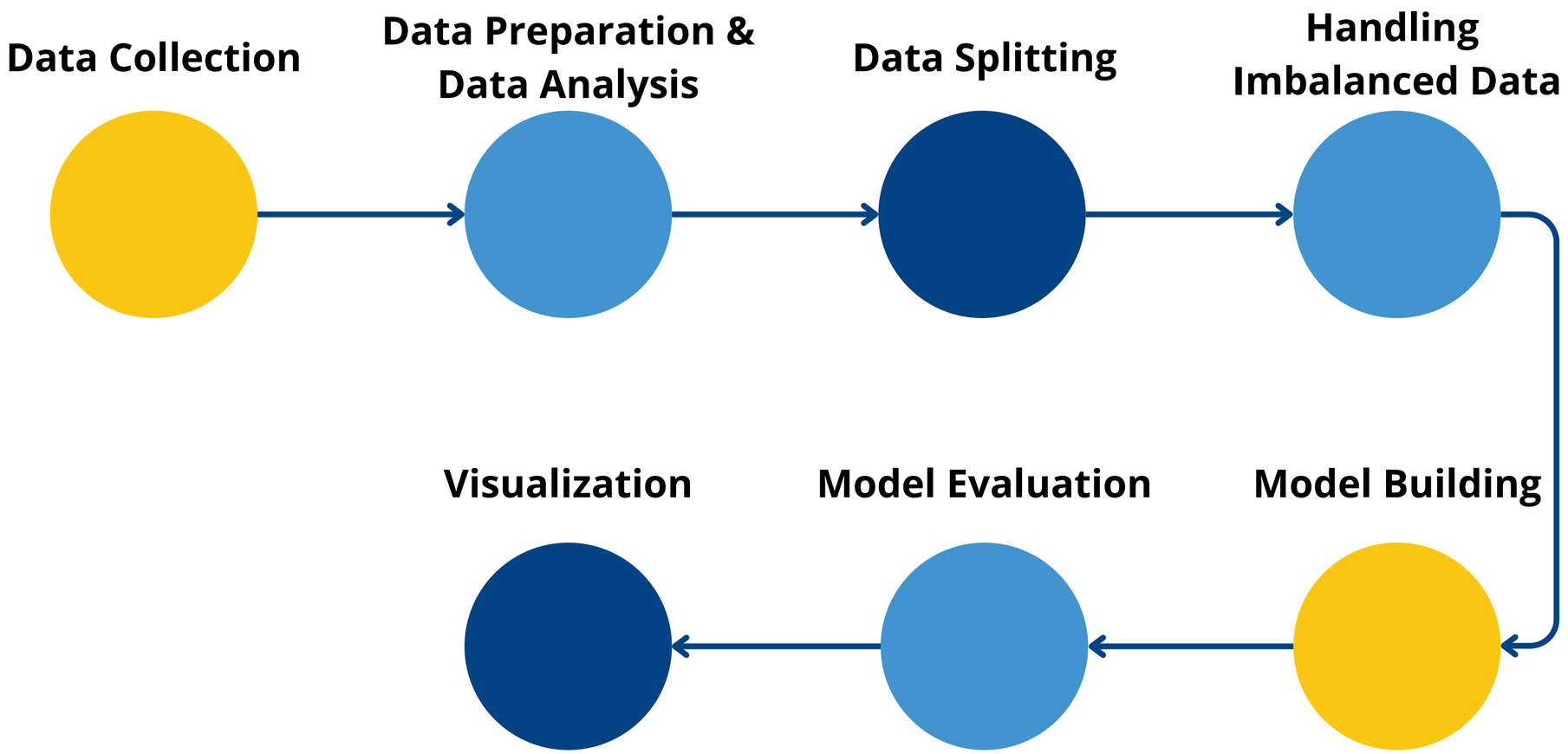
DATASET

Brain Stroke Dataset

By: Jillani SofTech

URL :
https://www.kaggle.com/datasets/jillanisofttech/brain-stroke-dataset

ALUR PEKERJAAN



METODOLOGI

DATA COLLECTION

Dataset diambil dari sumber eksternal agar dapat diproses lebih lanjut. Dataset diperoleh dari platform Kaggle menggunakan Kaggle API melalui library kagglehub.

DATA PREPARATION DAN EXPLORATORY DATA ANALYSIS

1. Data Inspection
2. Exploratory Data Analysis (EDA)
3. Beberapa aktivitas lain seperti Univariat, Bivariat, Korelasi, Distribusi Target

DATA SPLITTING

Proses ini dilakukan menggunakan fungsi train_test_split dari library scikit-learn dengan rasio 80:20, di mana 80% data digunakan untuk melatih model dan 20% sisanya digunakan untuk menguji performa model.

HANDLING IMBALANCED DATA

Kita menerapkan metode Synthetic Minority Oversampling Technique (SMOTE) pada data latih. SMOTE bekerja dengan cara membuat data sintetis dari kelas minoritas berdasarkan nilai fitur dari data yang ada.

MODEL BUILDING

Tahap modeling dilakukan dengan membangun dua algoritma klasifikasi yaitu Random Forest Classifier dan Naive Bayes Classifier. Kedua model dilatih menggunakan data latih yang telah diproses dan diseimbangkan dengan metode SMOTE.

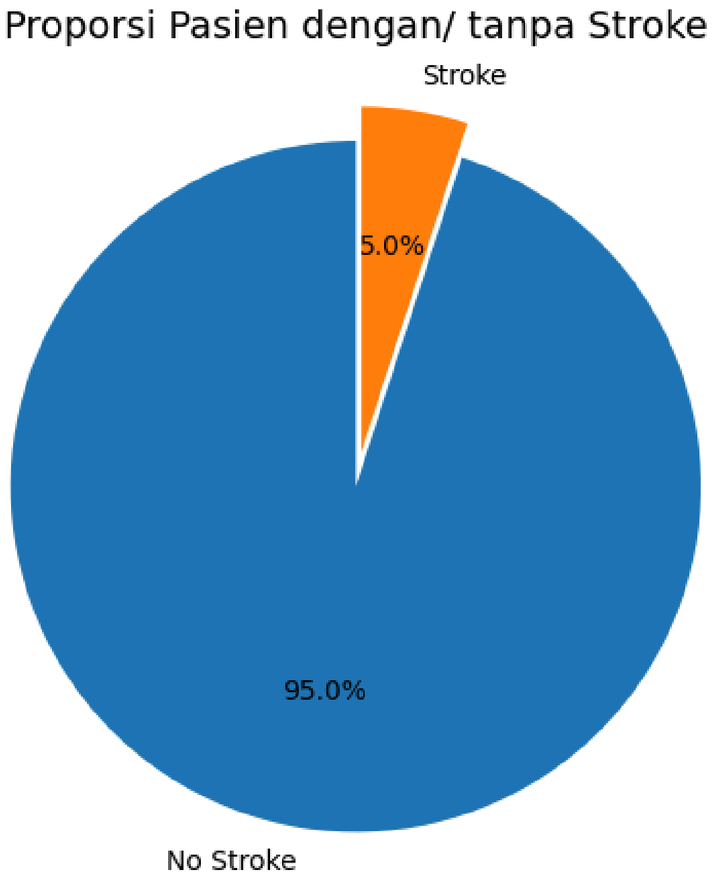
MODEL EVALUATION

Tahap akhir evaluasi dilakukan dengan membandingkan performa kedua model menggunakan beberapa metrik evaluasi, yaitu Accuracy, Precision, Recall, dan F1-Score. Nilai metrik dihitung berdasarkan hasil prediksi terhadap data uji.

VISUALIZATION

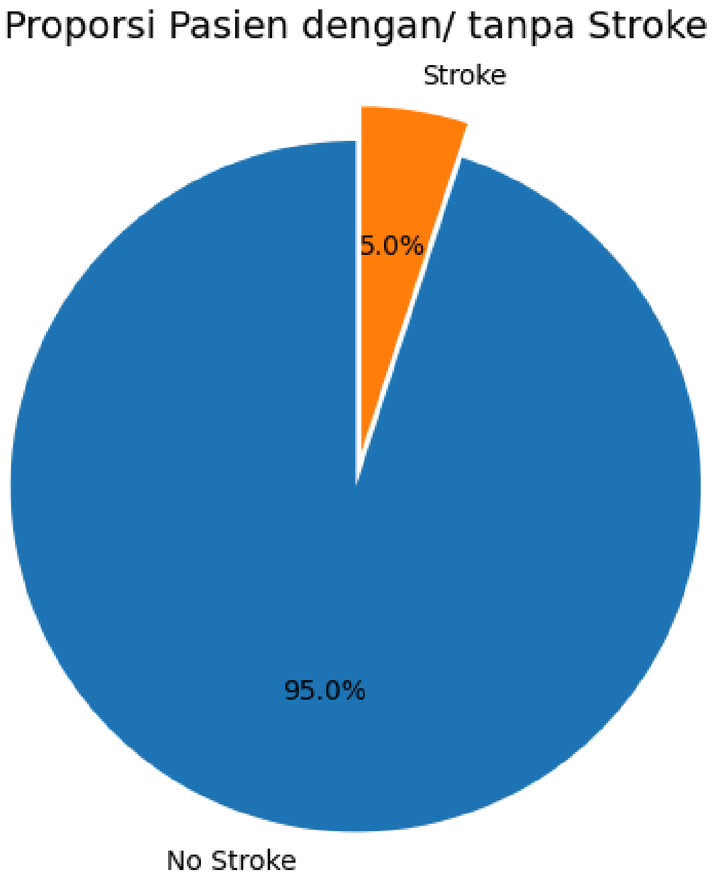
Untuk mempermudah interpretasi hasil evaluasi model, dilakukan visualisasi perbandingan nilai Accuracy, Precision, Recall, dan F1-Score menggunakan grafik batang. Visualisasi ini menampilkan performa masing-masing model dalam satu plot, sehingga perbedaan kinerja antar model dapat terlihat dengan jelas.

IMBALANCE DATA



- Dataset yang digunakan memiliki distribusi target yang sangat tidak seimbang.
- Mayoritas data adalah pasien tidak mengalami stroke (label 0), sedangkan hanya sebagian kecil mengalami stroke (label 1).
- Rasio kelas mendekati 19:1, yang membuat model lebih sering memprediksi kelas mayoritas untuk mendapatkan akurasi tinggi.
- Akibatnya, model sering gagal mengenali kasus stroke, sehingga nilai recall untuk kelas minoritas menjadi sangat rendah.
- Masalah ini umum terjadi dalam data medis dan bisa menyebabkan sistem prediksi menjadi tidak berguna secara praktis.

IMBALANCE DATA



Untuk mengatasi masalah ketidakseimbangan tersebut, proyek ini menggunakan teknik SMOTE (Synthetic Minority Oversampling Technique). SMOTE berfungsi dengan menghasilkan data sintetis dari kelas minoritas (pasien stroke), dengan cara membentuk sampel baru berdasarkan karakteristik data yang serupa di sekitarnya. Proses ini dilakukan hanya pada data latih, sehingga distribusi kelas menjadi lebih seimbang tanpa memengaruhi data uji dan menghindari data leakage. Dengan distribusi yang lebih proporsional, model memiliki kesempatan yang lebih adil untuk mempelajari pola-pola pada kelas stroke. Hasilnya, performa model—terutama pada metrik recall—meningkat, meskipun nilai F1-score masih menunjukkan bahwa klasifikasi terhadap kelas minoritas belum sepenuhnya optimal.

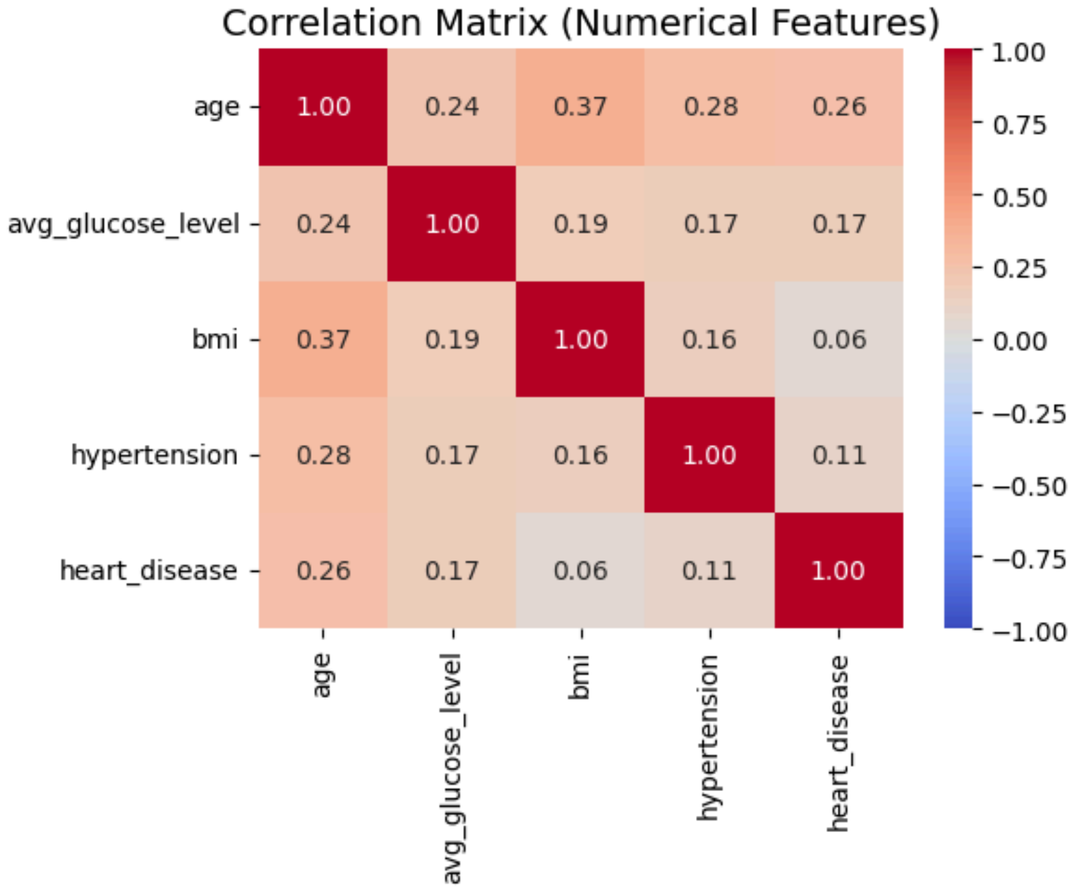
HEATMAP MATRIX KOLERASI

Hasil:

- Fitur age dan avg_glucose_level memiliki korelasi positif terhadap stroke
- Fitur lain seperti bmi memiliki korelasi lemah

Implikasi:

- Tidak ada fitur dominan → model ensemble seperti Random Forest lebih cocok
- Naive Bayes kurang cocok karena mengasumsikan independensi fitur



MODEL EVALUATION

Model	Accuracy	Precision	Recall	F1-Score
Random Forest	0.9147	0.5015	0.501	0.5002
Naive Bayes	0.5707	0.5489	0.7381	0.4487

Berdasarkan tabel di atas, Random Forest menunjukkan nilai accuracy tertinggi (91.47%) namun memiliki recall dan F1-score yang rendah, menandakan bahwa model ini lebih sering benar dalam memprediksi kelas mayoritas (non-stroke) namun kurang mampu mengenali kasus stroke (minoritas). Sementara itu, Naive Bayes memiliki recall tertinggi (73.81%) yang berarti model lebih “sensitif” terhadap kasus stroke, namun akurasi keseluruhannya rendah, menunjukkan banyak prediksi salah di kelas lain.

INTERPRETASI HASIL EVALUASI

- Random Forest adalah model yang secara umum lebih stabil, namun kurang sensitif terhadap kelas minoritas.
- Naive Bayes memiliki performa recall yang baik, namun menghasilkan banyak kesalahan klasifikasi.
- Penerapan SMOTE efektif menyeimbangkan data latih, namun tidak sepenuhnya mengatasi tantangan klasifikasi karena distribusi asli data uji tetap tidak seimbang.
- Perlu pertimbangan lebih lanjut untuk meningkatkan F1-score, seperti:
- Feature engineering tambahan
- Algoritma lain seperti XGBoost atau LightGBM
- Penggunaan threshold tuning atau cost-sensitive learning

KESIMPULAN

- Model Random Forest menunjukkan performa terbaik dari segi akurasi, yaitu sebesar 91.47%.
- Model Naive Bayes memiliki performa recall yang lebih tinggi dibandingkan Random Forest, yaitu sebesar 73.81%.
- Ketidakseimbangan kelas pada dataset stroke menjadi tantangan utama dalam proses klasifikasi. Untuk mengatasi hal ini, diterapkan metode SMOTE (Synthetic Minority Oversampling Technique) pada data latih.
- Visualisasi korelasi antar fitur menunjukkan bahwa tidak ada fitur tunggal yang dominan dalam mempengaruhi prediksi stroke. Oleh karena itu, penggunaan model yang dapat menangani banyak fitur dan interaksi kompleks seperti Random Forest menjadi lebih relevan.
- Secara keseluruhan, meskipun performa model masih belum sempurna, kedua algoritma memiliki potensi untuk digunakan dalam sistem pendukung keputusan medis, dengan catatan dilakukan penyesuaian lebih lanjut.

BINUS UNIVERSITY
SCHOOL OF COMPUTER SCIENCE

GROUP 4

THANK YOU