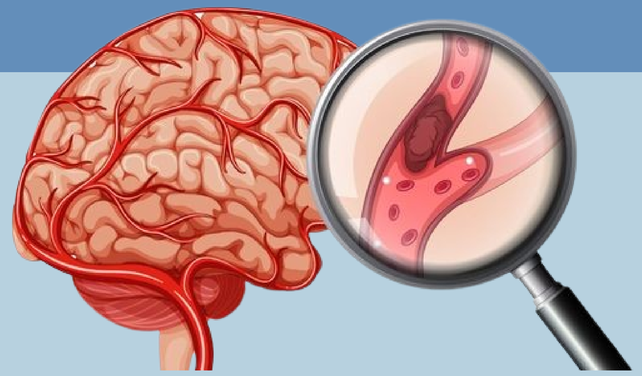
Prediction Model for Stroke Risk Identification Based on Health and Lifestyle Factors

Jasmin Putri Jelita, Nazilullaily Nur Aisyah, Valencia Sefiana Putri

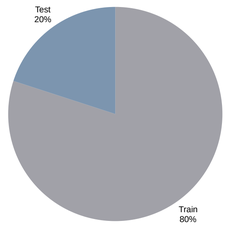


Business Objective

Membangun sebuah model prediktif yang dapat digunakan oleh para profesional kesehatan, rumah sakit, atau organisasi terkait dalam mengidentifikasi pasien yang berisiko tinggi terkena stroke. Model ini akan mempertimbangkan berbagai faktor seperti usia, jenis kelamin, riwayat penyakit, dan gaya hidup, termasuk status merokok. Dengan adanya model prediktif ini, diharapkan dapat membantu dalam deteksi dini dan pencegahan stroke.

Features and Data

- **Dataset:** Stroke Prediction Dataset
- **Data Description:** Data tersebut terdiri dari 5110 observasi dengan 12 atribut. Informasi ini mencakup data pasien dan riwayat kesehatannya, yang memungkinkan identifikasi individu yang berpotensi tinggi mengalami stroke berdasarkan parameter input seperti jenis kelamin, usia, berbagai penyakit, dan status merokok.
- **Data Mining Task:** Classification
- 80% Training / 20% Testing



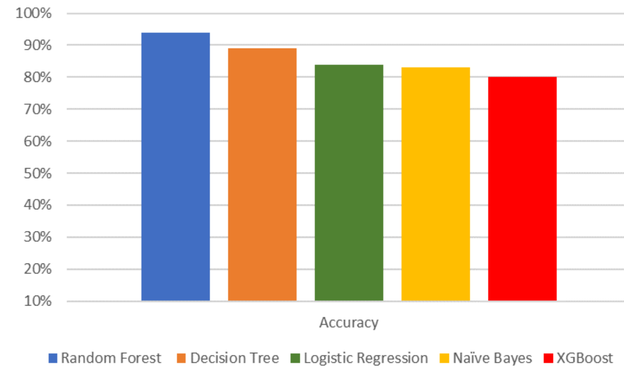
Dataset Feature

- id
- gender
- age
- hypertension
- heart_desease
- ever_married
- work_type
- residence_type
- avg_glucose
- bmi
- smoking_status
- stroke

Algorithms

- **Decision Tree Classifier:** Mengidentifikasi fitur signifikan dan hubungan antar fitur untuk klasifikasi stroke.
- **Random Forest:** Membentuk ansambel pohon keputusan untuk stabilitas dan akurasi tinggi.
- **Logistic Regression:** Memperkirakan probabilitas kejadian biner menggunakan fungsi logit.
- **Naive Bayes:** Menggunakan Teorema Bayes untuk klasifikasi dengan asumsi independensi fitur.
- **XGBoost:** Menggunakan boosting untuk meningkatkan akurasi prediksi

Accuracy Comparison



Berdasarkan analisis data, **Random Forest** menunjukkan performa terbaik dengan akurasi prediksi mencapai **94%**, mengungguli algoritma lain seperti Decision Tree Classifier (91%), Logistic Regression (84%), Naive Bayes (83%), dan XGBoost (80%).

Result

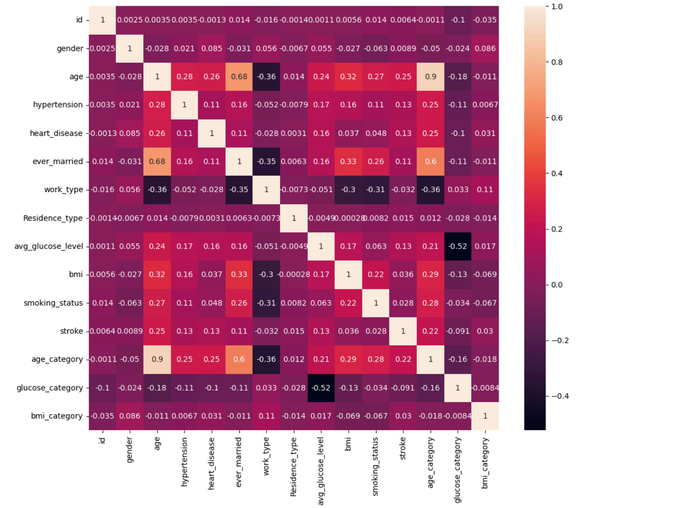


Diagram diatas menunjukkan hubungan antara berbagai faktor kesehatan dan demografi.

- Usia dan jenis kelamin berperan dalam risiko penyakit jantung dan tekanan darah tinggi.
- Status pernikahan dan pekerjaan memiliki hubungan dengan kadar glukosa darah dan status merokok.
- Tempat tinggal dapat memengaruhi tingkat obesitas.
- Gaya hidup seperti merokok dan indeks massa tubuh (BMI) berhubungan dengan berbagai kondisi kesehatan.

Methods

Data Cleaning:

- Isu Ditemukan: 201 data hilang pada fitur "bmi"
- Solusi: Data yang hilang diganti menggunakan nilai median dari fitur “bmi” untuk menjaga integritas dataset.

Diskretisasi Data:

- Mengubah data kategorikal menjadi numerik ke dalam rentang nilai tertentu.
- Contoh: Anak-anak: 0-18 years, Dewasa: 19-60 years, Lansia: >60 years.

Transformasi Data:

- Mengubah data hasil diskretisasi menjadi numerik.
- Contoh: "age_category" (anak-anak = 0, dewasa = 1, lansia = 2)

Smote

- Mengatasi masalah ketidakseimbangan kelas dalam dataset menggunakan teknik oversampling.
- Before: tidak stroke = 0.951272, stroke = 0.048728
- After: tidak stroke = 0.5, stroke = 0.5

References

- Kaggle Stroke Prediction Dataset, Sumber: Kaggle Dataset
- Wikipedia: Machine Learning, Sumber: Wikipedia - Machine Learning