**Predictive Modeling and Visualization of Stroke Risk with Random Forests**

**Big Data Processing - COMP6579001**

**Oleh:**

2602142363 - Arif Jiwo Hakimi

2602145421 - Axelle Farand Maestra

2602139886 - David Yovi Wardana

2602163866 - Ervan Vincentius

2602145213 - Rafael Djefry Alvindja

# BAB I: LATAR BELAKANG DAN PENGENALAN DATASET

Stroke merupakan penyakit serius yang terjadi saat pasokan darah ke bagian otak terputus atau berkurang secara signifikan, menyebabkan jaringan otak kekurangan oksigen dan nutrisi. Akibatnya, sel-sel otak mulai mati dalam hitungan menit. Stoke dapat menyebabkan kerusakan permanen pada otak dan berbagai komplikasi kesehatan serius lainnya. Efek dari stroke termasuk kelemahan atau kelumpuhan pada satu sisi tubuh, kesulitan berbicara atau memahami pembicaraan, dan masalah penglihatan. Komplikasi ini merupakan penyebab utama kematian dan kecacatan di seluruh dunia. Pada tahun 2019, stroke merupakan penyebab utama sekitar 6,2 juta kematian, dan sebagian besar terjadi pada pasien di atas 70 tahun. Di Indonesia, stroke merupakan salah satu penyebab utama kematian dan kecacatan, dengan prevalensi yang terus meningkat seiring bertambahnya usia populasi. Oleh karena itu, masyarakat harus memahami dan mengetahui bahaya dari penyakit stroke serta langkah-langkah yang bisa dilakukan untuk mencegahnya, seperti menjaga tekanan darah, menghindari merokok, dan menjalani gaya hidup sehat.

Dataset yang kita gunakan berjudul ‘Stroke Prediction Dataset’, dataset ini dipublikasikan di website *Kaggle.com*. Menurut Organisasi Kesehatan Dunia (WHO), stroke merupakan penyebab kematian terbesar kedua secara global, bertanggung jawab atas sekitar 11% dari total kematian. Dataset ini digunakan untuk memprediksi apakah seseorang pasien kemungkinan besar mengalami stroke berdasarkan parameter input seperti jenis kelamin, usia, berbagai jenis penyakit, dan status merokok. Setiap baris dalam data menyediakan informasi yang relevan tentang pasien. Data ini diolah untuk membuat model yang diharapkan dapat memprediksi penyakit stroke pada pasien berdasarkan Riwayat medis dan pola hidup mereka. Selain itu, informasi tentang atribut medis pasien dan hubungannya dengan stroke dapat diperdalam dengan menggunakan dataset ini. Hasilnya akan menjelaskan ciri mana yang paling relevan untuk menentukan apakah pasien memiliki stroke atau tidak.

# BAB II: METODOLOGI

Berikut adalah langkah-langkah yang kami gunakan untuk mengolah data penyakit stroke menjadi model prediksi dan mengumpulkan informasi tambahan tentang hubungan antara stroke dan elemen medis lainnya.

1. **Preparation**

A screenshot of a computer

Description automatically generated

* Import Module
  + Pandas untuk manipulasi data dan analisis.
  + Numpy untuk komputasi numerik.
  + sklearn.model\_selection.train\_test\_split untuk bagi dataset menjadi data pelatihan dan pengujian.
  + sklearn.preprocessing.StandardScaler untuk standarisasi fitur dengan menghilangkan mean dan menskalakan ke unit variance.
  + sklearn.ensemble.RandomForestClassifier adalah algoritma yang digunakan untuk klasifikasi.
  + sklearn.metrics.classification\_report, confusion\_matrix, accuracy\_score untuk mengevaluasi performa model klasifikasi.
  + matplotlib.pyplot untuk membuat plot dan grafik.
  + Seaborn untuk visualisasi data yang lebih mudah.
* Baca Dataset
  + Menggunakan pandas dan menyimpannya dengan variabel “data”
* Menampilkan Dataset
  + data.head() untuk menampilkan 5 baris pertama dari dataset
  + data.columns untuk menampilkan semua nama kolom dari dataset

## Pre-Processing dan Feature Engineering



* Pilih fitur penting
* Hapus nilai null
* Konversi data non-numerik ke numerik
* Bagi data menjadi 2 subset training dan testing

## 3. Generate Model & Evaluation

A computer screen shot of a computer code

Description automatically generated

* Evaluasi akurasi model dengan zero\_division diatur ke 1
* Menampilkan Classification Report
* Menghitung dan Menampilkan Nilai Akurasi

Hasil klasifikasi dan matrix :

A screenshot of a computer screen

Description automatically generated

Berikut persentase dari akurasi model:



## Visualization

* Menggunakan matplotlib untuk menampilkan grafik atau chart
* Heatmap dan grafik feature importances ditampilkan untuk mengetahui seberapa penting fitur-fitur yang berkorelasi dengan strokeA graph of blue rectangular bars with white text

  Description automatically generatedA diagram of a number of numbers

  Description automatically generated with medium confidence

1. Distribusi gender dengan jumlah pasien stroke

A screen shot of a computer screen

Description automatically generatedA graph of a number of men and women

Description automatically generated

1. Distribusi usia dengan jumlah pasien strokeA screen shot of a computer screen

   Description automatically generated

A graph of blue and orange bars

Description automatically generated

1. Perbandingan BMI pada Pasien Stroke dan Tidak StrokeA screen shot of a computer screen

   Description automatically generatedA graph of a number of people

   Description automatically generated with medium confidence
2. Relasi antara Tingkat Gula Darah pada Pengidap Stroke dan Tidak StrokeA screen shot of a computer program

   Description automatically generated

A graph of blue and orange bars

Description automatically generated

1. Perbandingan pasien Stroke dan non-Stroke yang memiliki HipertensiA screen shot of a computer program

   Description automatically generated

A graph with blue and orange bars

Description automatically generated

1. Perbandingan pasien Stroke dan non-Stroke yang memiliki penyakit Jantung

A screen shot of a computer program

Description automatically generated A graph with blue and orange bars

Description automatically generated

1. Perbandingan pasien Stroke dan non-Stroke yang memiliki Hipertensi dan penyakit Jantung

A screen shot of a computer code

Description automatically generated

A graph with a bar and numbers

Description automatically generated with medium confidence

1. Perbandingan pasien Stroke dan non-Stroke berdasarkan smoking statusA screenshot of a computer program

   Description automatically generated

A comparison of pie charts

Description automatically generated

# BAB III: HASIL DAN PENEMUAN ANALISIS

A screenshot of a computer

Description automatically generated

Dengan akurasi 94.60%, hasil model prediksi algoritma Random Forrest menunjukkan bahwa pasien dengan penyakit stroke dapat didiagnosis jika ada informasi tentang tingkat gula darah, indeks massa tubuh (BMI), umur, jenis kelamin, penyakit jantung, dan hipertensi.

Berdasarkan hasil dan penemuan kami, tingkat gula darah berkontribusi 29.53% pada diagnosis stroke, sedangkan BMI adalah 23.26% dan umur adalah 22.62%.

Dengan melihat data yang telah dilakukan, juga dapat dilihat gambaran statistik penderita stroke dari dataset, yaitu:

1. Distribusi gender pada pasien stroke

- Dari 4699 penderita stroke, 1922 merupakan laki-laki dan 2777 merupakan perempuan.

- Artinya, gender berpengaruh besar terhadap stroke.

1. Distribusi usia pada pasien stroke

- Seperti grafik yang ditampilkan, pasien stoke terbanyak terjadi pada umur 55-60 tahun.

- Artinya semakin tua, maka beresiko terkena penyakit stoke walaupun ada factor lainnya.

1. Perbandingan BMI pada pasien stroke dan non-stroke

- BMI pada pasien stroke (biru) terjadi dikisaran angka yang cukup tinggi yaitu dari 20-40 sedangkan pada pasien non-stroke (orange) cenderung lebih rendah.

1. Perbandingan tingkat gula darah pada pasien stroke dan non-stroke

- Penderita stroke cenderung memiliki Tingkat gula darah yang rendah yaitu dibawah 90 dimana Tingkat penderita berbeda drastis.

1. Pasien stroke yang memiliki hipertensi

- Dari 4699 penderita stroke, 4308 pasien merupakan penderita hipertensi dan 391 pasien sisanya tidak mengalami hipertensi.

- Pasien yang memiliki hipertensi berpengaruh besar terhadap stroke.

1. Pasien stroke yang memiliki penyakit jantung

- Dari 4699 penderita stroke, 4496 pasien memiliki penyakit jantung dan 203 pasien sisanya tidak.

- Pasien yang memiliki penyakit jantung berpengaruh besar terhadap stroke.

1. Pasien stroke yang memiliki hipertensi dan penyakit jantung

- Dari 4699 penderita stroke, 4652 pasien memiliki hipertensi dan penyakit jantung dan 47 pasien sisanya tidak.

- Pasien yang memiliki hipertensi dan penyakit jantung berpengaruh besar terhadap stroke.

1. Perbandingan kasus stroke berdasarkan status merokok

- Penderita stroke dan non-stroke yang sedang aktif merokok dan yang tidak pernah merokok memiliki selisih 1%.

- Penderita stroke dan nonstroke yang merupakan mantan perokok memiliki selisih 12% Dimana penderita stroke 12% lebih banyak yang merupakan mantan perokok.

- Artinya, persentase terkena stroke karena merokok hanya berpengaruh pada mantan perokok

**BAB IV: KESIMPULAN**

Stroke adalah kondisi medis serius yang terjadi secara tiba-tiba ketika aliran darah ke otak terganggu atau terhenti. Hal ini dapat terjadi karena pembuluh darah di otak pecah atau terjadi penyumbatan. Untuk menentukan resiko stroke, sejumlah faktor penting dipertimbangkan, termasuk usia, jenis kelamin, indeks massa tubuh (BMI), hipertensi, penyakit jantung, status merokok, dan kadar glukosa darah. Data yang disajikan di atas menunjukkan hubungan langsung antara diabetes dan kadar glukosa darah, seperti yang ditunjukkan oleh model prediksi yang menggunakan algoritma Klasifikasi Random Forest. Oleh karena itu, menjaga pola hidup sehat sangat penting untuk mengurangi resiko stroke, terutama pada orang yang lebih tua.

**REFERENSI**

Dataset : Diabetes Prediction Dataset – Kaggle

<https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset/data>

Mitra Keluarga. (2024, January 03). *Penyakit Stroke: Gejala, Penyebab, dan Pengobatan yang Benar.* <https://www.mitrakeluarga.com/artikel/penyakit-stroke>

IHC Telemed. (2021). *Stroke.* <https://telemed.ihc.id/artikel-detail-419-Stroke.html>