

**DATA MINING & MACHINE LEARNING
(Stroke Risk Prediction)**



OLEH

Kelompok 10

- | | |
|------------------------|------------|
| 1. A. Afif Alhaq | H071221064 |
| 2. Adrian Hidayat | H071221091 |
| 3. Evan Pandu Nata | H071221057 |
| 4. Muhammad Rifky Aqid | H071211077 |

**PROGRAM STUDI SISTEM INFORMASI
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS HASANUDDIN
MAKASSAR
2024**

Latar Belakang

Stroke merupakan suatu kondisi kesehatan serius yang terjadi ketika pasokan darah ke bagian otak terganggu atau terputus, menyebabkan sel-sel otak tidak mendapatkan cukup oksigen dan nutrisi. Kondisi ini dapat berdampak fatal dan memiliki dampak yang signifikan pada kesehatan seseorang. Menurut Data pada artikel, stroke menjadi penyebab utama kematian di Indonesia. Bahkan prevalensi pasien stroke meningkat dari 7% per mil menjadi 10,9% per mil. Statistik ini mencerminkan urgensi untuk mengembangkan model machine learning yang dapat memprediksi risiko stroke pada individu, sebagai langkah proaktif dalam mengurangi beban kesehatan global yang disebabkan oleh stroke.

Pentingnya memprediksi risiko stroke juga dapat dilihat dari data-data konkret. Misalnya, dalam konteks internasional, Stroke menyerang sekitar 15 juta orang per tahun, dan dampaknya tidak hanya terbatas pada tingkat kematian. Banyak yang selamat dari stroke mengalami kecacatan jangka panjang, memerlukan perawatan intensif, dan mengalami penurunan kualitas hidup yang signifikan. Oleh karena itu, upaya preventif menjadi krusial untuk mengurangi beban sosial dan ekonomi yang ditimbulkan oleh stroke.

Dalam upaya meningkatkan pemahaman dan pencegahan terhadap stroke, sebuah proyek penelitian akan dilakukan untuk mengembangkan model machine learning yang dapat memprediksi risiko stroke pada individu. Proyek ini akan melibatkan analisis berbagai faktor kesehatan, seperti tekanan darah, kadar gula darah, indeks massa tubuh (BMI), riwayat merokok, dan faktor risiko lainnya yang dapat mempengaruhi kemungkinan seseorang mengalami stroke. Dimana variabel-variabel tersebut didukung oleh data, yang mengatakan bahwa penyebab stroke dapat diakibatkan oleh tekanan darah, kurangnya aktivitas fisik, konsumsi makanan, kebiasaan merokok, dan faktor-faktor lainnya.

Dengan memanfaatkan data kesehatan dan riwayat medis, model ini diharapkan dapat memberikan prediksi akurat tentang risiko stroke pada seseorang. Implementasi proyek ini dapat membantu para profesional medis dalam memberikan peringatan dini kepada pasien dengan risiko tinggi, sehingga tindakan pencegahan dapat diambil lebih awal. Selain itu, model ini dapat menjadi alat yang berguna bagi masyarakat dalam memahami dan mengelola faktor risiko kesehatan mereka, serta mendukung kebijakan kesehatan masyarakat yang lebih proaktif dalam pencegahan stroke.

Proyek ini dibangun untuk membantu para profesional medis dalam memberikan peringatan dini kepada pasien dengan risiko tinggi, sehingga tindakan pencegahan dapat diambil lebih awal. Selain itu, model ini dapat menjadi alat yang berguna bagi masyarakat dalam memahami dan mengelola faktor risiko kesehatan mereka, serta mendukung kebijakan kesehatan masyarakat yang lebih proaktif dalam pencegahan stroke.

Identifikasi Masalah

1. Bagaimana mengidentifikasi faktor-faktor risiko yang berkontribusi pada kemungkinan seseorang terkena stroke?
2. Bagaimana membangun model prediktif untuk memprediksi risiko seseorang terkena stroke berdasarkan faktor-faktor tertentu?
3. Bagaimana menentukan dampak variabel-variabel tertentu terhadap prediksi risiko stroke?

Tujuan

1. Menganalisis faktor-faktor risiko yang berkontribusi pada kemungkinan seseorang terkena stroke.
2. Mengembangkan model machine learning untuk memprediksi risiko stroke berdasarkan variabel-variabel tertentu.
3. Menyajikan hasil analisis dan model secara jelas dan terukur untuk mendukung pengambilan keputusan.

Pernyataan Solusi

- Menyiapkan data agar dapat digunakan untuk proses analisis dan pemodelan machine learning.
- Melakukan eksplorasi data untuk memahami hubungan antar variabel dan melihat tren dalam dataset.
- Melakukan feature engineering (Standarisasi, One-hot-encoding) untuk meningkatkan kinerja model.
- Membangun model prediktif menggunakan algoritma klasifikasi seperti K-Nearest Neighbors, Random Forest dan Adaptive Boosting.
- Melakukan evaluasi model menggunakan metrik evaluasi yang relevan seperti akurasi, presisi, recall, dan F1-score. Memilih model terbaik berdasarkan performa evaluasi.
- Melakukan hyperparameter tuning menggunakan grid search untuk meningkatkan performa model.
- Melakukan Deployment Model untuk Aplikasi Prediksi Stroke

Kontribusi Model dalam Kesehatan Masyarakat

Penyadaran Risiko Stroke: Model ini dapat memberikan informasi kepada masyarakat tentang faktor risiko stroke yang perlu diwaspadai. Dengan pemahaman ini, individu dapat mengambil langkah-langkah preventif yang lebih baik.

Pengelolaan Faktor Risiko Kesehatan: Melalui hasil prediksi model, masyarakat dapat mengelola faktor risiko kesehatan mereka secara lebih efektif. Misalnya, dengan menyesuaikan pola makan, meningkatkan aktivitas fisik, atau mengontrol tekanan darah, mereka dapat mengurangi risiko stroke.

Dukungan Kebijakan Kesehatan Masyarakat: Model ini dapat menjadi dasar bagi kebijakan kesehatan masyarakat yang lebih proaktif dalam pencegahan stroke. Dengan menargetkan populasi dengan risiko tinggi, pihak berwenang dapat mengimplementasikan intervensi yang lebih efisien dan efektif.

Peran Edukasi dan Kesadaran: Model ini dapat digunakan sebagai alat edukasi untuk meningkatkan kesadaran masyarakat tentang pentingnya deteksi dini dan pencegahan stroke. Ini dapat menciptakan budaya kesehatan yang lebih baik di masyarakat.

Pembahasan

1. Data Understanding

Dataset yang digunakan adalah dataset yang diambil dari Kaggle. Dataset ini merupakan data kesehatan yang berisi informasi tentang kondisi kesehatan dan riwayat medis dari pasien yang berpotensi mengalami stroke.

Berikut informasi dari dataset :

- Dataset memiliki format CSV (Comma-Separated Values).
- Dataset memiliki 5110 baris dan 12 kolom.
- Dataset memiliki 4 kolom bertipe numerik dan 8 kolom bertipe kategorikal.
- Dataset memiliki 201 baris data yang memiliki nilai null pada kolom BMI.

Variabel-variabel pada dataset adalah sebagai berikut:

- ✓ **id**: ID unik untuk setiap data
- ✓ **gender**: Jenis kelamin pasien
- ✓ **age**: Usia pasien
- ✓ **hypertension**: Pasien menderita hipertensi atau tidak
- ✓ **heart_disease**: Pasien menderita penyakit jantung atau tidak
- ✓ **ever_married**: Pasien pernah menikah atau tidak
- ✓ **work_type**: Tipe pekerjaan pasien
- ✓ **Residence_type**: Tipe tempat tinggal pasien
- ✓ **avg_glucose_level**: Rata-rata kadar glukosa pasien dalam satuan mg/dL
- ✓ **bmi**: Indeks massa tubuh pasien
- ✓ **smoking_status**: Status merokok pasien
- ✓ **stroke**: Pasien mengalami stroke atau tidak

2. Eksplorasi Data

Informasi Data

#	Column	Non-Null Count	Dtype
0	id	5110 non-null	int64
1	gender	5110 non-null	object
2	age	5110 non-null	float64
3	hypertension	5110 non-null	int64
4	heart_disease	5110 non-null	int64
5	ever_married	5110 non-null	object
6	work_type	5110 non-null	object
7	Residence_type	5110 non-null	object
8	avg_glucose_level	5110 non-null	float64
9	bmi	4909 non-null	float64
10	smoking_status	5110 non-null	object
11	stroke	5110 non-null	int64

Deskripsi Data

	id	age	hypertension	heart_disease	avg_glucose_level	bmi	stroke
count	5110.000000	5110.000000	5110.000000	5110.000000	5110.000000	4909.000000	5110.000000
mean	36517.829354	43.226614	0.097456	0.054012	106.147677	28.893237	0.048728
std	21161.721625	22.612647	0.296607	0.226063	45.283560	7.854067	0.215320
min	67.000000	0.080000	0.000000	0.000000	55.120000	10.300000	0.000000
25%	17741.250000	25.000000	0.000000	0.000000	77.245000	23.500000	0.000000
50%	36932.000000	45.000000	0.000000	0.000000	91.885000	28.100000	0.000000
75%	54682.000000	61.000000	0.000000	0.000000	114.090000	33.100000	0.000000
max	72940.000000	82.000000	1.000000	1.000000	271.740000	97.600000	1.000000

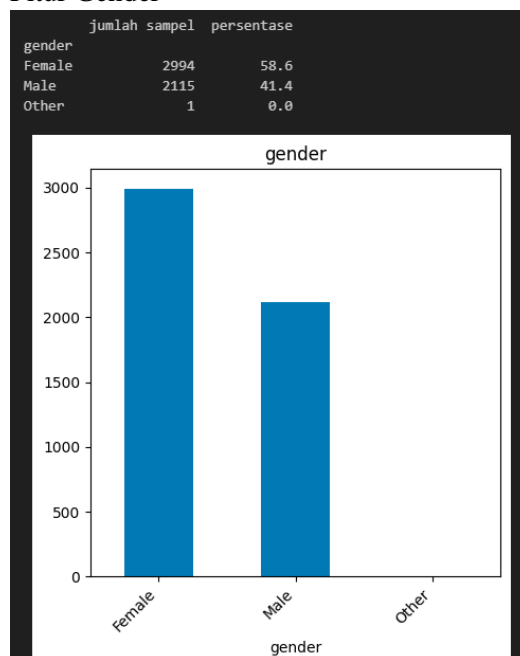
Dari informasi data diatas, dapat dilihat bahwa kolom id tidak mempengaruhi proses analisis dan pemodelan, karena hanya sebagai index saja. Oleh karena itu, kolom id akan dihapus dari dataset.

Univariate Analysis

- **Fitur Kategorikal**

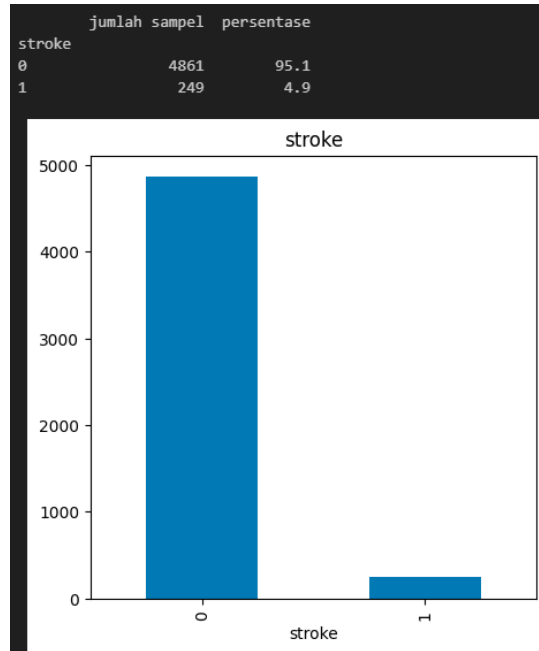
Berikut beberapa grafik yang menunjukkan distribusi data pada fitur kategorikal

- **Fitur Gender**



Dari grafik pada gambar, dapat dilihat bahwa jumlah pasien perempuan lebih banyak daripada pasien laki-laki. Selain itu, terdapat gender yang tidak diketahui, sehingga akan dihapus dari dataset.

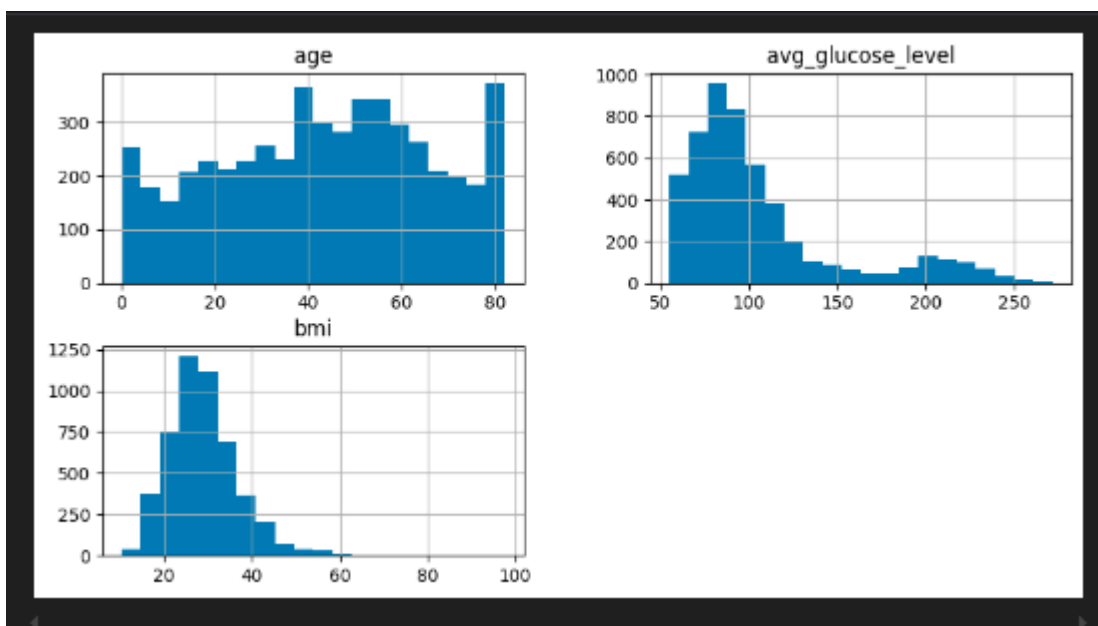
- **Fitur Stroke**



Dari grafik pada gambar diatas, dapat dilihat bahwa jumlah pasien yang tidak mengalami stroke lebih banyak daripada jumlah pasien yang mengalami stroke. Selain itu, dapat dilihat bahwa dataset memiliki distribusi yang tidak seimbang, yang ditunjukkan dengan jumlah pasien yang tidak mengalami stroke lebih banyak daripada jumlah pasien yang mengalami stroke. Sehingga perlu dilakukan oversampling untuk membuat distribusi data pada fitur stroke menjadi seimbang.

- **Fitur Numerikal**

Berikut merupakan grafik yang menunjukkan distribusi data pada fitur numerikal.



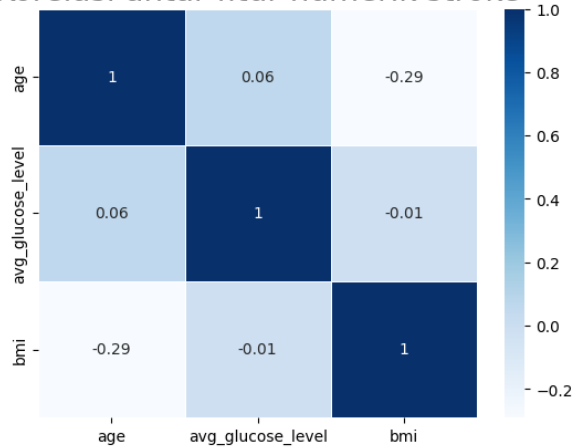
Dari grafik pada gambar diatas, dapat dilihat bahwa:

- Usia pasien memiliki distribusi yang normal.
- Kadar glukosa pasien paling banyak berada pada rentang 50 sampai 125.
- Indeks massa tubuh pasien paling banyak berada pada rentang 10 sampai 30.

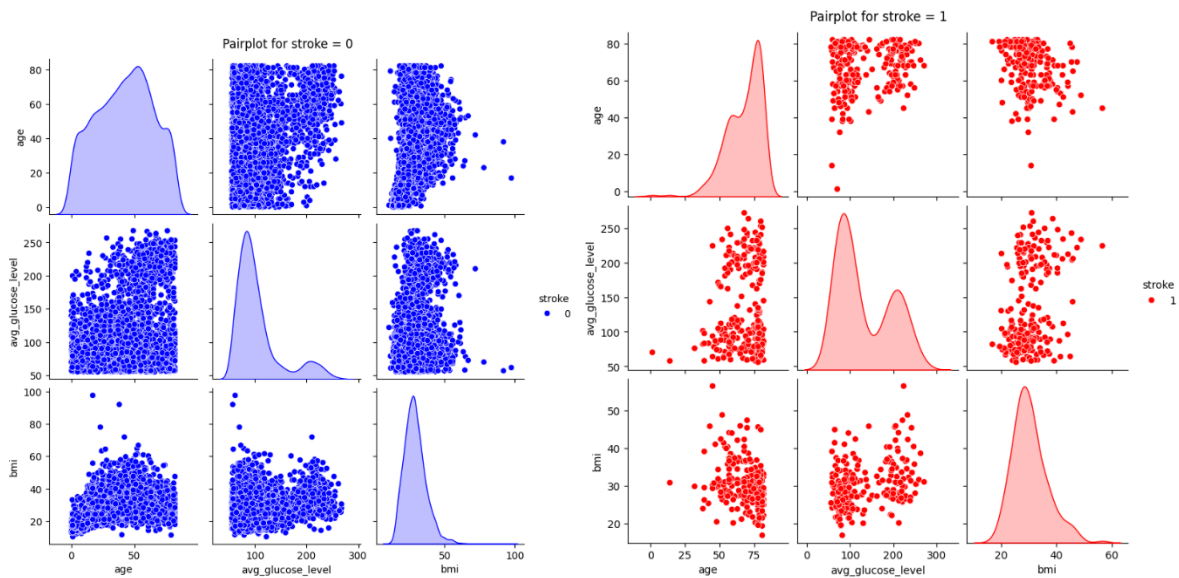
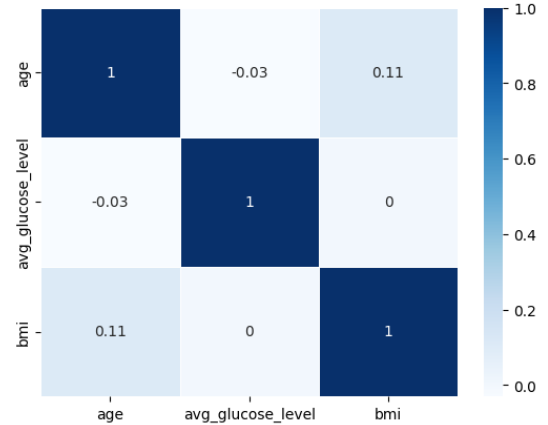
Multivariate Analysis

• Fitur Numerikal

Korelasi antar fitur numerik stroke



Korelasi antar fitur numerik tidak stroke



Pada grafik diatas dapat dilihat bahwa pasien stroke terjadi paling banyak pada:

- Usia 40 keatas
- Kadar glukosa diluar kadar glukosa normal (sekita 160 mg/dL) [5]
- Indeks massa tubuh 20 sampai 50

3. Data Preparation

- **Drop Column**

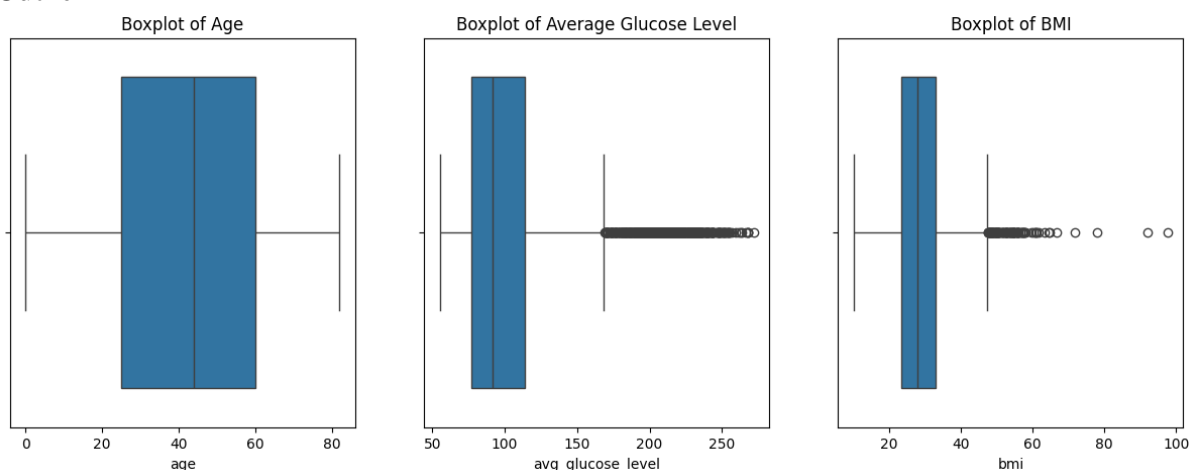
Kolom id akan dihapus dari dataset karena hanya sebagai indeks saja.

- **Missing Value**

kolom	jumlah missing value
gender	0
age	0
hypertension	0
heart_disease	0
ever_married	0
work_type	0
Residence_type	0
avg_glucose_level	0
bmi	201
smoking_status	0
stroke	0

Pada kolom bmi terdapat 201 baris data yang memiliki nilai null. Untuk mengatasi hal ini, nilai null akan dihapus dari dataset.

- **Outlier**



Dari boxplot diatas, dapat dilihat bahwa kolom avg_glucose_level dan bmi memiliki outlier. Untuk mengatasi hal ini, outlier akan dihapus dari dataset.

- **Imputation Beberapa fitur yang diubah**

- Mengubah nilai Other pada kolom gender menjadi Female
- Menghapus data yang bernilai Unknown pada kolom smoking_status
- Mengubah nilai formerly smoked pada kolom smoking_status menjadi smokes

- **One-Hot Encoding**

Pada dataset terdapat beberapa fitur yang memiliki tipe data kategorikal, sehingga perlu dilakukan one-hot encoding untuk mengubah fitur-fitur tersebut menjadi numerikal. One-hot encoding sendiri merupakan proses mengubah fitur kategorikal menjadi vektor biner yang dapat digunakan oleh algoritma machine learning untuk melakukan prediksi. Fitur-fitur yang akan dilakukan one-hot encoding adalah gender, hypertension, heart_disease, ever_married, work_type, residence_type, dan smoking_status.

- **Train Test Split**

Dataset akan dibagi menjadi 2 bagian, yaitu data training dan data testing. Data training akan digunakan untuk melatih model, sedangkan data testing akan digunakan untuk menguji performa model. Data training akan memiliki proporsi 80% dari total data, sedangkan data testing akan memiliki proporsi 20% dari total data.

- **Oversampling**

Pada dataset, distribusi data pada fitur stroke tidak seimbang, yang ditunjukkan dengan jumlah data yang tidak stroke lebih banyak daripada jumlah data yang stroke. Oleh karena itu, perlu dilakukan oversampling untuk membuat distribusi data pada fitur stroke menjadi seimbang. Oversampling sendiri merupakan proses menambahkan data pada kelas minoritas sehingga jumlah data pada kelas minoritas menjadi sama dengan jumlah data pada kelas mayoritas. Pada proyek ini, oversampling akan dilakukan menggunakan metode SMOTE (Synthetic Minority Oversampling Technique).

- **Standarisasi**

Pada dataset terdapat beberapa fitur yang memiliki skala yang berbeda, sehingga perlu dilakukan standarisasi untuk mengubah skala fitur-fitur tersebut menjadi sama. Standarisasi sendiri merupakan proses mengubah fitur numerikal menjadi memiliki nilai rata-rata 0 dan standar deviasi 1. Fitur-fitur yang akan dilakukan standarisasi adalah age, avg_glucose_level, dan bmi.

4. Modeling

- **Algoritma Machine Learning**

Pada proyek ini, akan digunakan 3 algoritma klasifikasi, yaitu K-Nearest Neighbors, Random Forest, dan Adaptive Boosting, dimana algoritma-algoritma ini digunakan karena memiliki performa yang baik dalam melakukan klasifikasi [5].

Ketiga algoritma ini akan dibandingkan performanya untuk menentukan algoritma mana yang paling cocok untuk memprediksi risiko stroke pada individu.

– **Algoritma K-Nearest Neighbors (KNN)** adalah metode sederhana yang menggunakan kesamaan fitur untuk memprediksi nilai data baru. Dalam konteks regresi, KNN memilih sejumlah k tetangga terdekat dan menghitung rata-rata nilai target mereka sebagai prediksi. Parameter yang digunakan pada algoritma ini adalah:

- **n_neighbors** : Jumlah tetangga terdekat yang akan digunakan untuk melakukan prediksi.
- **weights**: Menentukan bobot yang akan digunakan dalam prediksi. Bobot yang tersedia adalah uniform dan distance. Jika menggunakan uniform, maka semua tetangga akan

memiliki bobot yang sama. Jika menggunakan distance, maka tetangga yang lebih dekat akan memiliki bobot yang lebih besar.

- `leaf_size`: Menentukan ukuran leaf tree yang akan digunakan. Semakin besar nilai leaf size, maka semakin cepat proses prediksi, namun semakin besar pula penggunaan memori.
- `p` : Menentukan metrik jarak yang akan digunakan. Metrik jarak yang tersedia adalah 1 (Manhattan Distance) dan 2 (Euclidean Distance).

– **Algoritma Random Forest** adalah model machine learning dalam kategori ensemble learning, yang terdiri dari beberapa model (pohon keputusan) yang bekerja bersama untuk menyelesaikan masalah klasifikasi atau regresi. Model ensemble bekerja dengan menggabungkan prediksi dari setiap model individual untuk menghasilkan prediksi akhir. Beberapa parameter yang digunakan, yaitu:

- `n_estimators` : Jumlah pohon keputusan yang akan digunakan.
- `max_depth` : Maksimal kedalaman pohon keputusan yang akan digunakan.
- `min_samples_split` : Jumlah minimum sampel yang dibutuhkan untuk membagi node. Jika jumlah sampel pada node kurang dari `min_samples_split`, maka node tersebut tidak akan dibagi.
- `min_samples_leaf` : Jumlah minimum sampel yang dibutuhkan untuk menjadi leaf node. Jika jumlah sampel pada node kurang dari `min_samples_leaf`, maka node tersebut tidak akan menjadi leaf node.

– **Algoritma Adaptive Boosting (AdaBoost)** adalah algoritma ensemble learning yang digunakan dalam klasifikasi. Algoritma ini menggabungkan beberapa model machine learning sederhana untuk membuat prediksi yang lebih akurat. Beberapa parameter yang digunakan, yaitu:

- `n_estimators` : Jumlah model yang akan digunakan.
- `learning_rate` : Menentukan seberapa besar bobot yang akan diberikan pada model berikutnya. Semakin besar learning rate, maka semakin besar pula bobot yang diberikan pada model berikutnya.

- **Hyperparameter Turning (Grid Search)**

Pada proyek ini, akan dilakukan hyperparameter tuning menggunakan grid search untuk meningkatkan performa model. Hyperparameter tuning sendiri merupakan proses mencari kombinasi hyperparameter terbaik untuk meningkatkan performa model. Berikut adalah nilai yang akan digunakan untuk hyperparameter tuning:

Model	Hyperparameter yang akan di-tuning	Nilai yang akan digunakan
K-Nearest Neighbors	n_neighbors	5, 10, 15, 20, 25, 30
	weights	uniform, distance
	leaf_size	10, 20, 30, 40, 50
	p	1, 2
Random Forest	n_estimators	100, 200, 300, 400, 500
	max_depth	5, 10, 15, 20, 25, 30
	min_samples_split	2, 5, 10
	min_samples_leaf	1, 2, 4
Adaptive Boosting	n_estimators	100, 200, 300, 400, 500
	learning_rate	0.01, 0.05, 0.1, 0.3, 1

dimana, nilai yang akan digunakan untuk hyperparameter tuning didapatkan dari nilai-nilai yang biasa digunakan pada algoritma tersebut.

Hyperparameter tuning akan dilakukan pada ketiga algoritma yang digunakan, yaitu K-Nearest Neighbors, Random Forest, dan Adaptive Boosting. Berikut adalah hasil dari hyperparameter tuning yang dilakukan:

Hasil hyperparameter turning

Model	Best Parameters	Best Score
K-Nearest Neighbors	{'leaf_size': 10, 'n_neighbors': 5, 'p': 1, 'weights': 'distance'}	0.9739819004524886
Random Forest	{'max_depth': 30, 'min_samples_leaf': 1, 'min_samples_split': 5, 'n_estimators': 400}	0.978733031674208
Adaptive Boosting	{'learning_rate': 0.3, 'n_estimators': 400}	0.9678733031674207

Dari tabel diatas, dapat dilihat bahwa dari ketiga algoritma yang digunakan, algoritma Radom Forest memiliki skor yang paling baik, dengan nilai 0.978733031674208.

5. Evaluation

Metrik evaluasi yang digunakan pada proyek ini adalah akurasi, precision, recall, dan F1-score. Metrik evaluasi ini dipilih karena dapat digunakan untuk mengevaluasi performa model klasifikasi.

- Akurasi: Akurasi merupakan rasio prediksi yang benar dengan total data yang diprediksi. Akurasi dapat digunakan untuk mengevaluasi performa model klasifikasi jika dataset memiliki distribusi yang sama untuk setiap kelasnya. Namun, jika dataset memiliki distribusi yang tidak sama untuk setiap kelasnya, maka akurasi tidak dapat digunakan untuk mengevaluasi performa model klasifikasi.
- Precision: Precision merupakan rasio prediksi yang benar dengan total data yang diprediksi positif. Precision dapat digunakan untuk mengevaluasi performa model klasifikasi jika terdapat biaya yang tinggi untuk mendeteksi data negatif sebagai positif.
- Recall: Recall merupakan rasio prediksi yang benar dengan total data yang sebenarnya positif. Recall dapat digunakan untuk mengevaluasi performa model klasifikasi jika terdapat biaya yang tinggi untuk mendeteksi data positif sebagai negatif.
- F1-score: F1-score merupakan rata-rata harmonik dari precision dan recall. F1-score dapat digunakan untuk mengevaluasi performa model klasifikasi jika terdapat biaya yang tinggi untuk mendeteksi data positif sebagai negatif dan mendeteksi data negatif sebagai positif.

Karena pada dataset distribusi negatif (tidak stroke) lebih banyak daripada distribusi positif (stroke), dan model diprioritaskan untuk mendeteksi data positif, maka metrik evaluasi yang digunakan adalah metrik recall dengan nilai yang tinggi dan metrik f1-score dengan nilai yang seimbang.

Berikut adalah hasil evaluasi dari ketiga algoritma yang digunakan:

- Hasil Evaluasi Akurasi

	train	test
KNN	1.0	0.942708
RandomForest	1.0	0.961806
Boosting	0.974874	0.949653

- Hasil Evaluasi Precision

	train	test
KNN	1.0	0.125
RandomForest	1.0	0.0
Boosting	0.990539	0.0

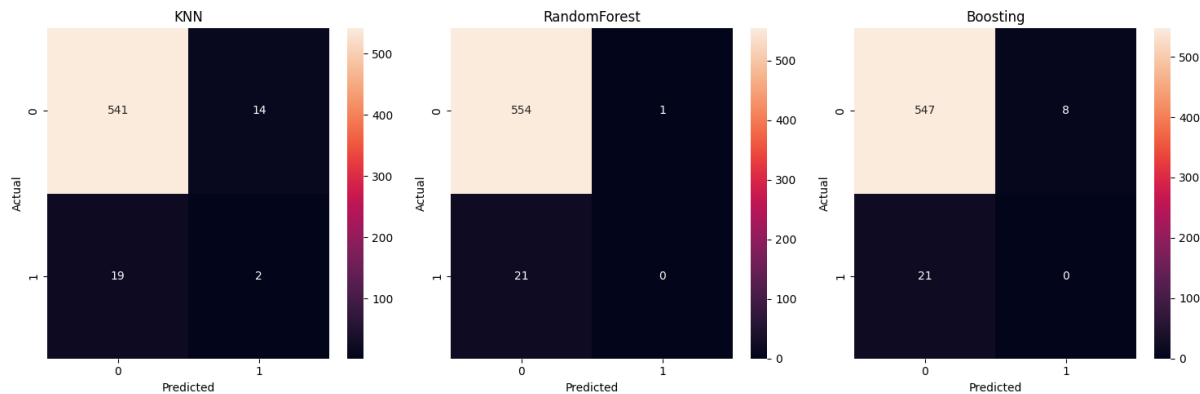
- Hasil Evaluasi Recall

	train	test
KNN	1.0	0.095238
RandomForest	0.986878	0.0
Boosting	0.947511	0.0

- Hasil Evaluasi f1-score

	train	test
KNN	1.0	0.108108
RandomForest	0.993396	0.0
Boosting	0.968548	0.0

- Confusion Matrix



Dari nilai diatas, maka dapat disimpulkan bahwa ketiga model mampu memprediksi dengan baik untuk data negatif (tidak stroke), namun kurang baik untuk data positif (stroke). Namun, karena model diprioritaskan untuk mendeteksi data positif sebagai metode pencegahan risiko stroke, maka model yang dipilih adalah model dengan nilai recall pada data testing paling tinggi, yaitu model K-Nearest Neighbors. Selain itu, model K-Nearest Neighbors juga dapat memprediksi data positif dengan lebih baik daripada model Random Forest dan Adaptive Boosting, yang ditunjukkan dari confusion matrix pada gambar diatas.

Namun, perlu diingat bahwa prediksi yang dilakukan oleh model ini hanya dilakukan untuk keperluan pencegahan risiko stroke, sehingga model ini tidak dapat digunakan untuk diagnosis stroke pada pasien.

6. Tampilan Aplikasi Website

Aplikasi yang kami buat untuk mendeteksi atau memprediksi seseorang dengan beberapa variable yang kami pertimbangkan apakah dia dapat terkena stroke atau tidak. Kami menggunakan tools VScode sebagai wadah dalam membuat aplikasi ini dengan python dan html.

Home

Gender: Male

Age: Enter your Age

Hypertension: Yes

Heart Disease: Yes

Ever Married: Yes

Work Type: Govt Job

Residence Type: Urban

Avg Glucose Level: Avg Glucose Level

BMI: BMI

Smoking Status: Unknown

Submit

You have been diagnosed with Stroke Risk

The algorithm has diagnosed you with Stroke Risk based on your inputs. Please consult a Doctor.

You have been diagnosed with no Stroke Risk. Congratulations

The algorithm has diagnosed you with no Stroke Risk based on your inputs. However it might be better to talk to a doctor regardless.

Penutup

Kesimpulan

Stroke merupakan kondisi kesehatan serius yang terjadi akibat terganggunya pasokan darah ke otak, menyebabkan sel-sel otak kekurangan oksigen dan nutrisi. Kondisi ini dapat berakibat fatal dan menjadi salah satu penyebab utama kematian di Indonesia, dengan prevalensi yang meningkat dari 7% menjadi 10,9% per mil. Situasi ini menggarisbawahi urgensi untuk mengembangkan model machine learning yang mampu memprediksi risiko stroke pada individu sebagai langkah preventif untuk mengurangi beban kesehatan global akibat stroke.

Upaya memprediksi risiko stroke menjadi penting mengingat dampaknya yang luas. Di seluruh dunia, stroke mempengaruhi sekitar 15 juta orang per tahun, dengan banyak yang mengalami kecacatan jangka panjang dan penurunan kualitas hidup. Pendekatan preventif dapat membantu mengurangi beban sosial dan ekonomi yang ditimbulkan oleh stroke. Untuk itu, sebuah proyek penelitian akan dilakukan untuk mengembangkan model machine learning dengan menganalisis berbagai faktor kesehatan seperti tekanan darah, kadar gula darah, indeks massa tubuh, kebiasaan merokok, dan faktor risiko lainnya yang berkaitan dengan stroke.

Berdasarkan analisis yang dilakukan, proyek ini diharapkan dapat memberikan prediksi akurat tentang risiko stroke, membantu para profesional medis dalam memberikan peringatan dini kepada pasien berisiko tinggi, dan memungkinkan tindakan pencegahan lebih awal. Selain itu, model ini akan berguna bagi masyarakat dalam memahami dan mengelola faktor risiko kesehatan mereka serta mendukung kebijakan kesehatan masyarakat yang lebih proaktif dalam pencegahan stroke. Dengan demikian, proyek ini tidak hanya berkontribusi pada penurunan angka kematian akibat stroke tetapi juga pada peningkatan kualitas hidup dan kesehatan masyarakat secara keseluruhan.