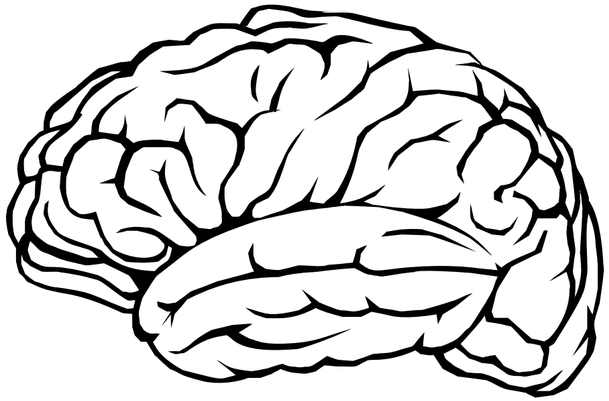


Stroke Prediction Through Multi-Model Machine Learning Approach

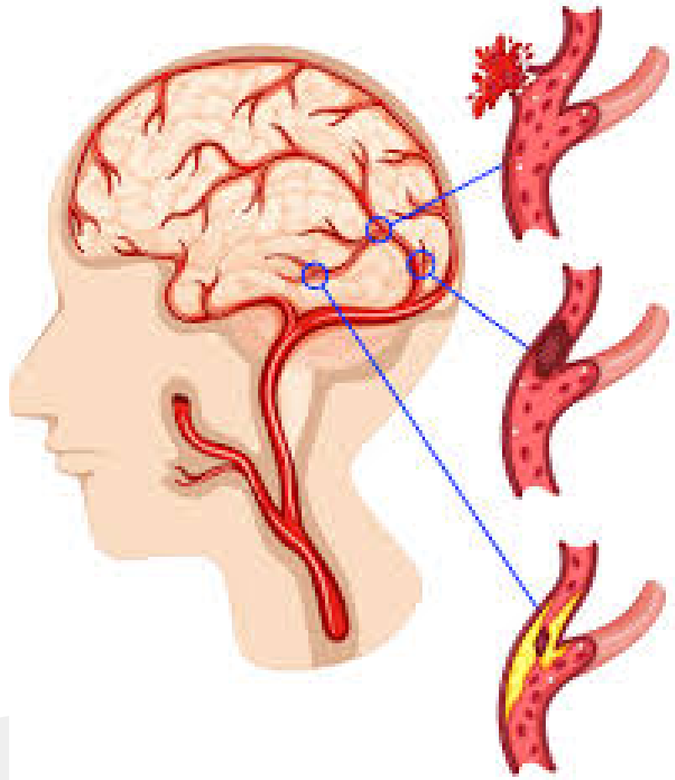
FINAL PROJECT DS 27 B

By : Aliifah Qurrotaayun

About the disease

Stroke, a medical emergency that occurs due to the interruption of flow of blood to a part of brain because of bleeding or blood clots. Worldwide, it is the second major reason for deaths with an annual mortality rate of 5.5 million. Every year, more than 15 million people worldwide have a stroke, and in every 4 minutes, someone dies due to stroke.

By analyzing medical data, we will train Four machine learning models to identify patterns and risk factors associated with stroke.



Outline

01
Business Undersanding

02
Exploratory Data Analysis

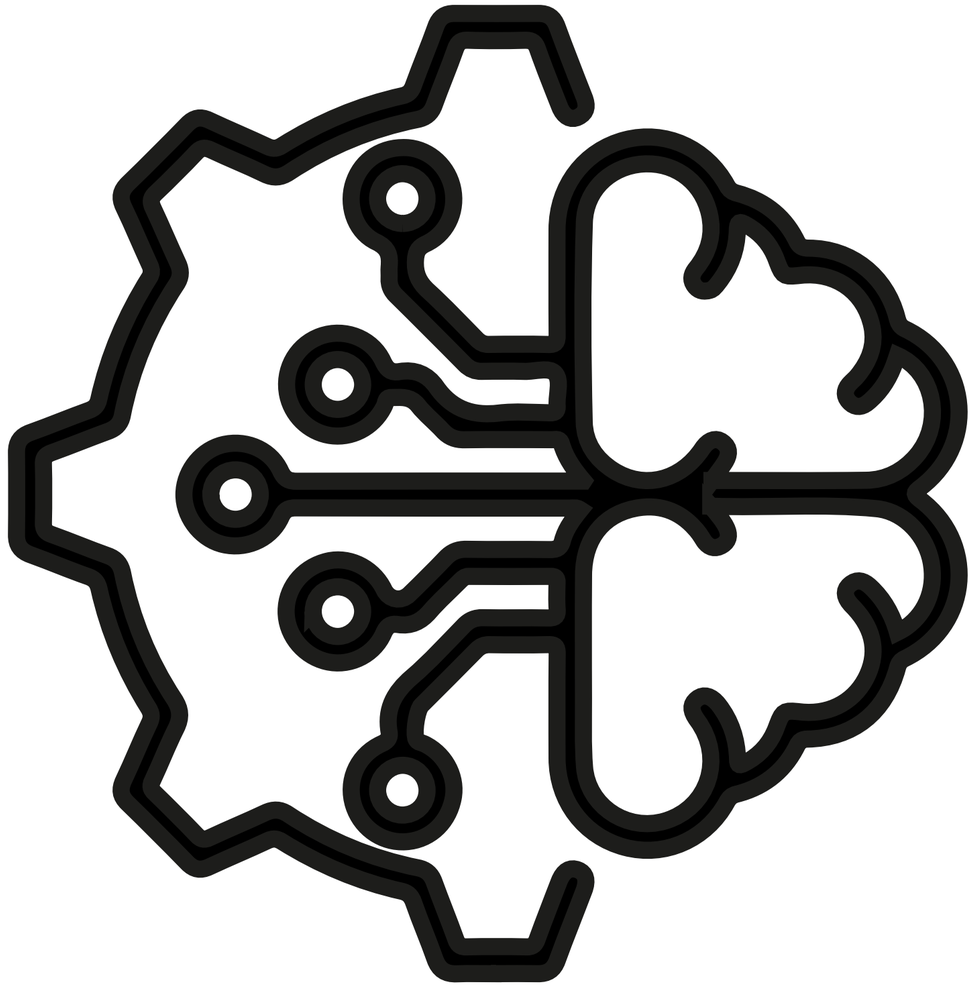
03
Data Prepocessing

04
Modelling & Evaluation

05
Business Recommendation

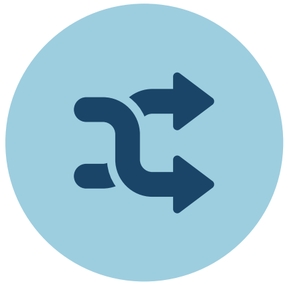
Business Understanding

- Problem Statement
- Goals, objective & Metrics



Problem statement

Increased Incident of Stroke

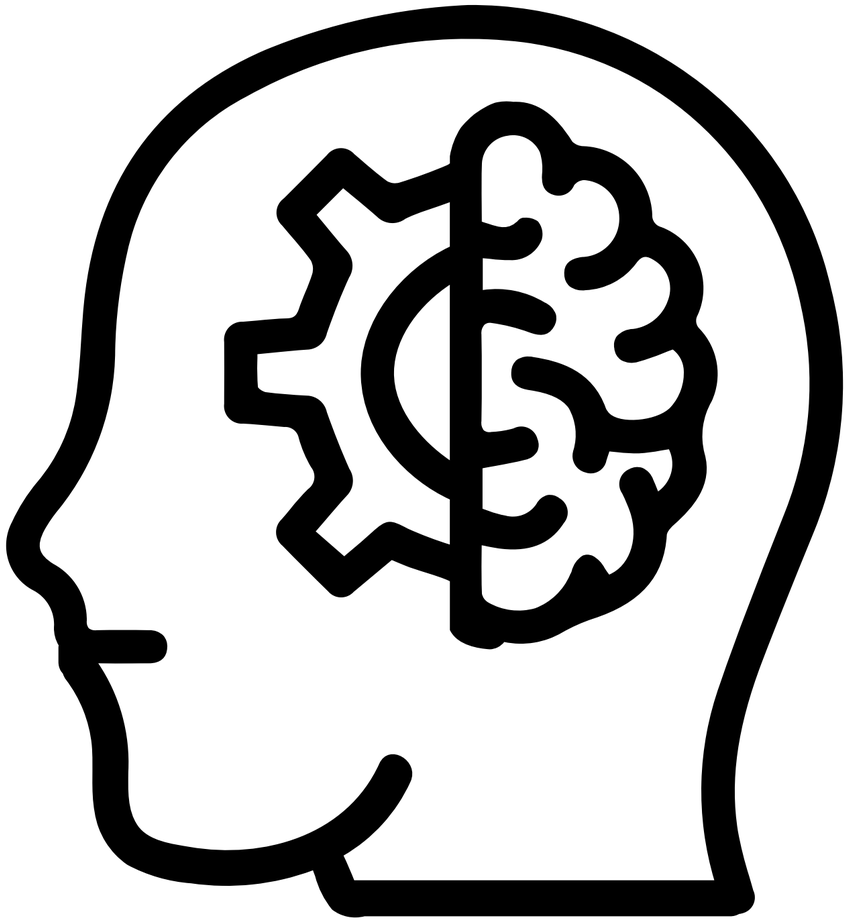


High Healthcare Costs

Lack of Reliable Prediction Tools

Goals, Objective & Metrics

Goals	Objective	Metrics
The main goal of the stroke prediction project is to **save lives** by detecting stroke risks early and providing timely medical intervention recommendations. It aims to **reduce healthcare costs** for patients and medical institutions through effective prevention while **improving patients' quality of life** by mitigating the risk of stroke.	**Building an accurate stroke prediction model**	**Recall**
	Identifying key risk factors	**Precision**
	Providing actionable insights	**F1-Score**



EXPLORATORY DATA ANALYSIS

DATASET

Dataset memiliki 11 kolom dan 5110 baris

Kolom BMI memiliki 201 nilai null (4,9% data null)

Tidak terdapat data Duplicate

DATASET

Id	Unique identification number for each patient
Age	The age of the patient
Hypertension	Whether the patient has hypertension
Heart Disease	Whether the patient has history of heart disease
Ever married	Whether the patient has been married
Work Type	The type of work the patient does
Residence Type	Whether the patient resides in an urban or rural area
Glucose Level	The patient's glucose level
BMI (body mass index)	A measure of body fat based on height and weight
Smoking Status	The Smoking habit of the patient
Stroke	The target variable indicating whether the patient had a stroke

Multivariate Analysis

25%

AGE

There is a moderate positive relationship between age and stroke (the older the age, the greater the likelihood of stroke).

13%

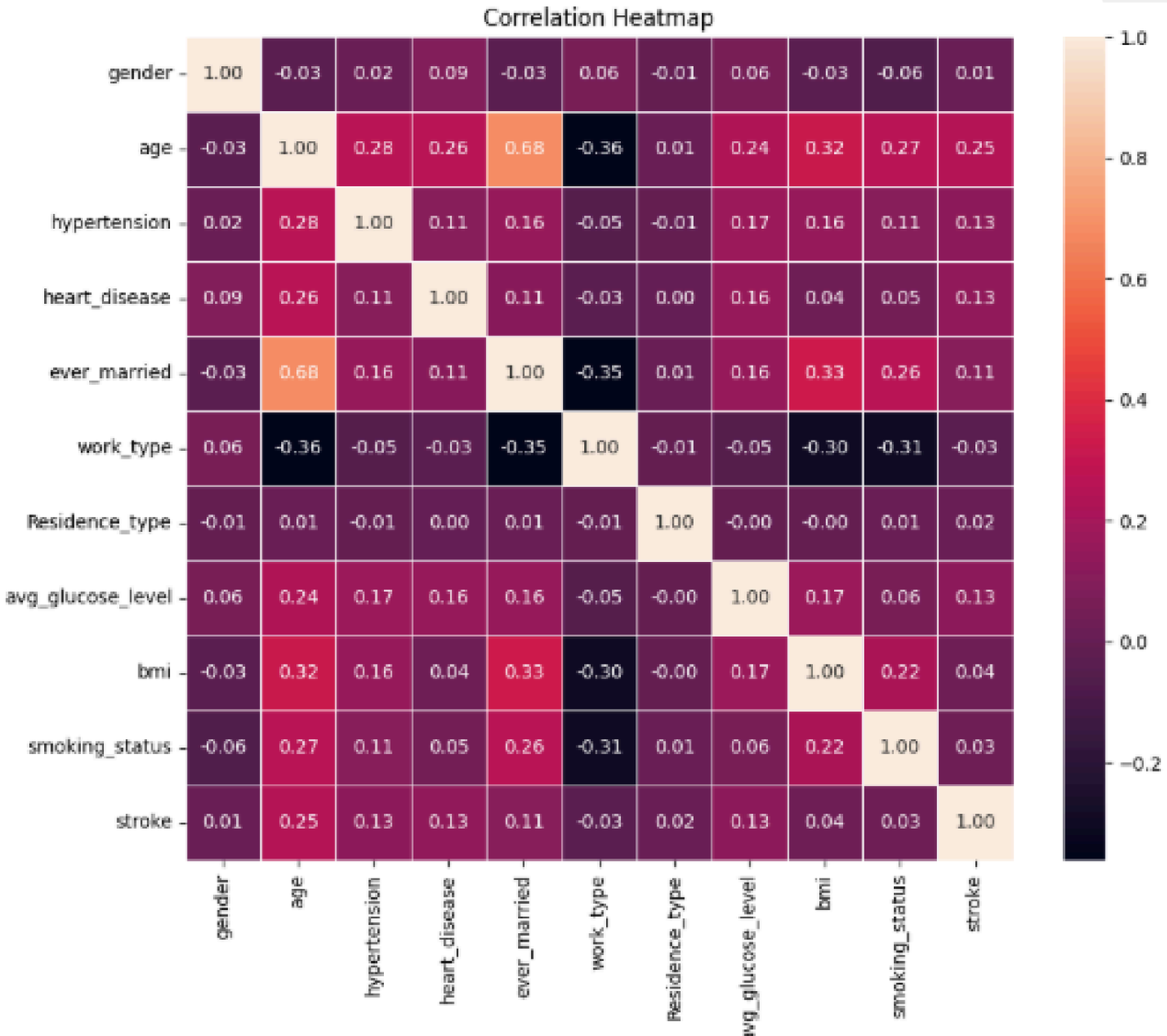
HYPERTENSION

There is a weak positive association between hypertension and stroke.

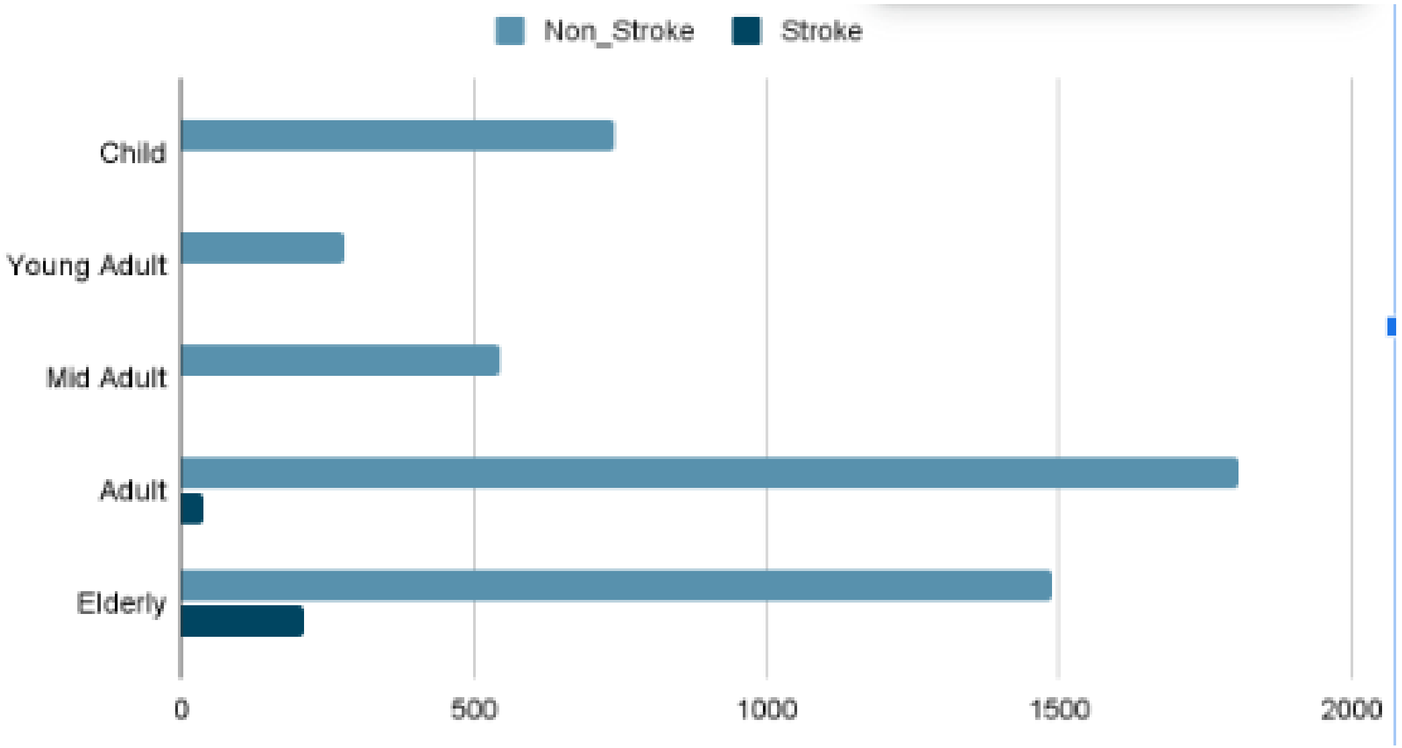
13%

AVG_GLUCOSE

Weak positive association between mean blood sugar levels and stroke.



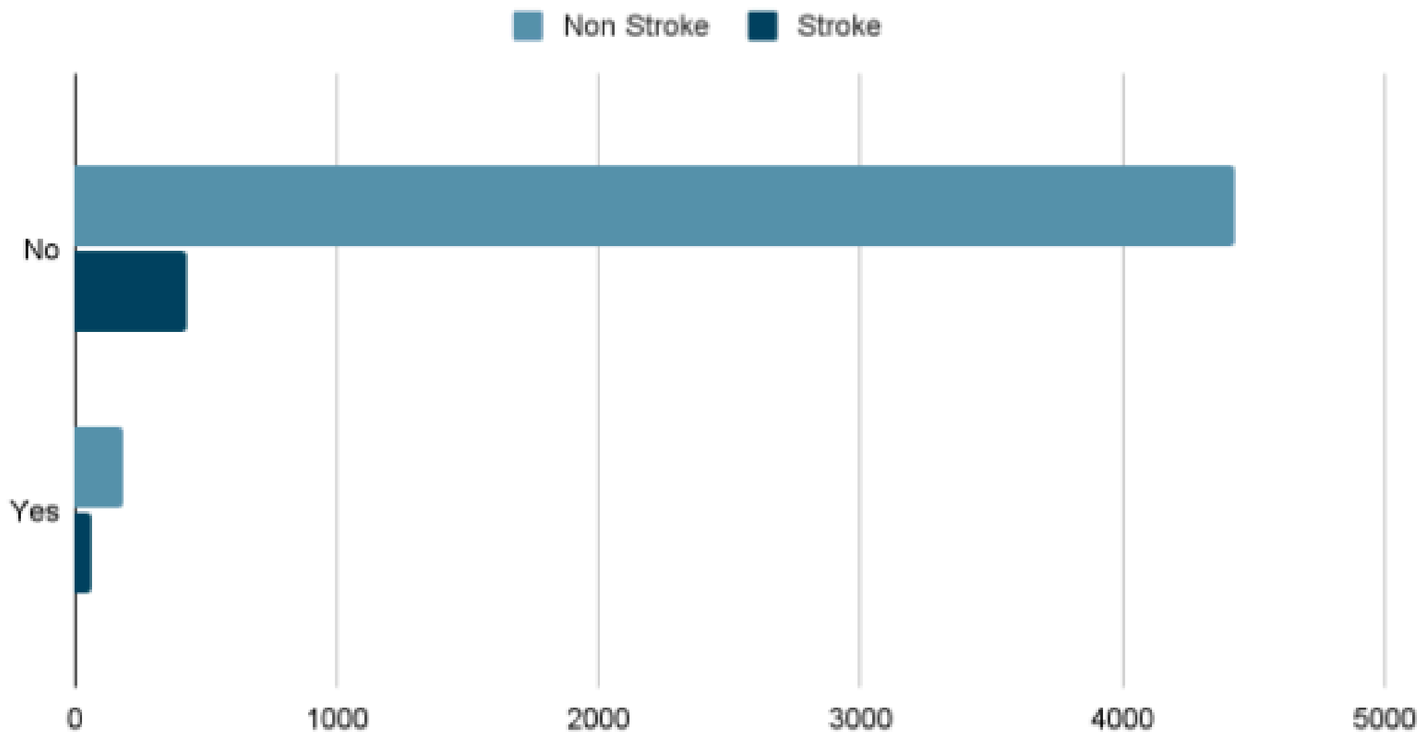
AGE VS STROKE



- Stroke is more common in the Elderly group than in other age groups. This shows that the risk of stroke increases with age.

- The Adult and Mid Adult groups have more cases of "Non-Stroke", indicating a lower risk of stroke in these age groups.

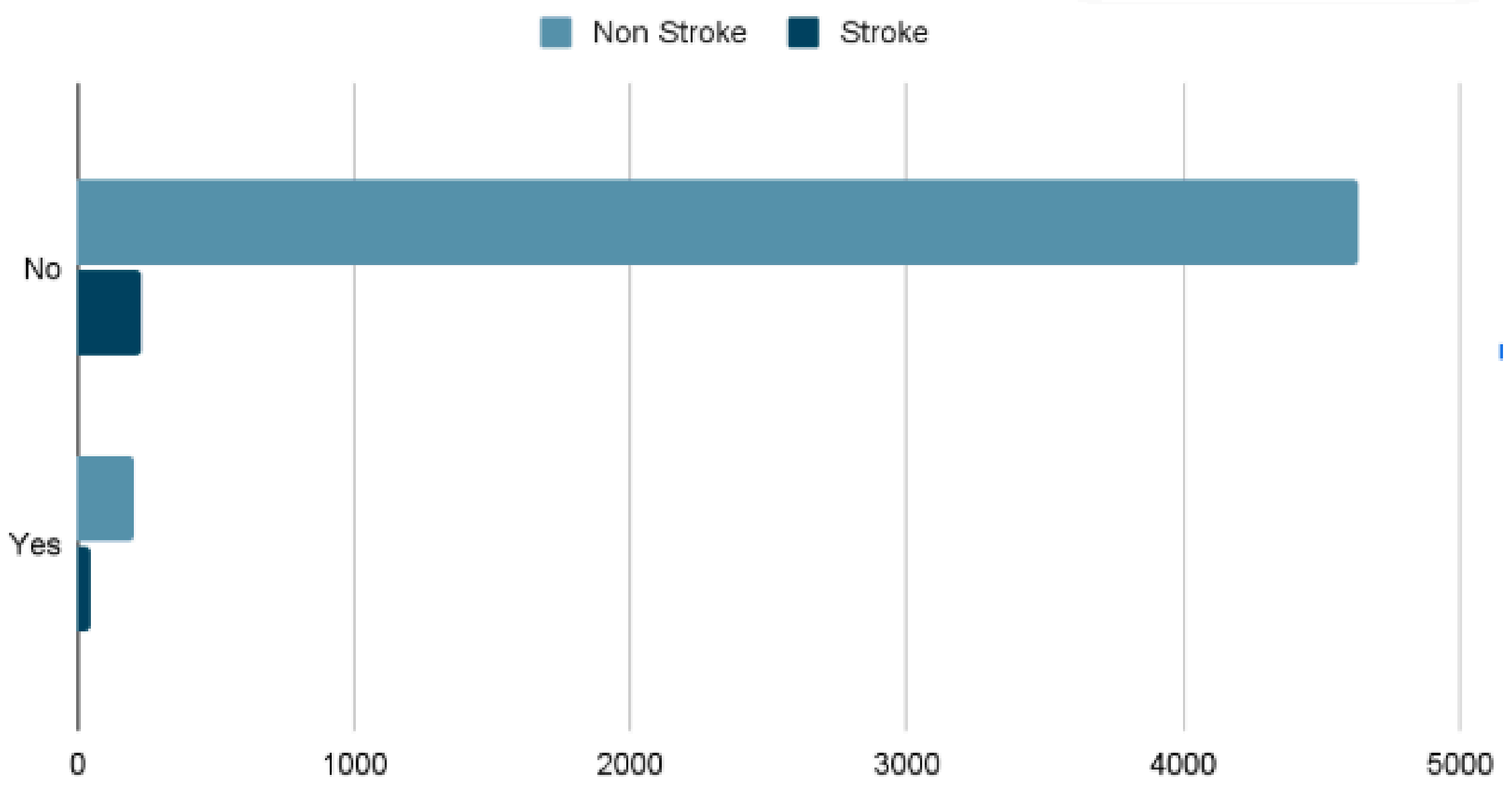
Hypertension VS Stroke



This suggests that hypertension is one of the main risk factors for stroke

The proportion of individuals without hypertension ("No") is much higher overall. This may suggest that hypertension does not always lead to stroke, but increases its probability.

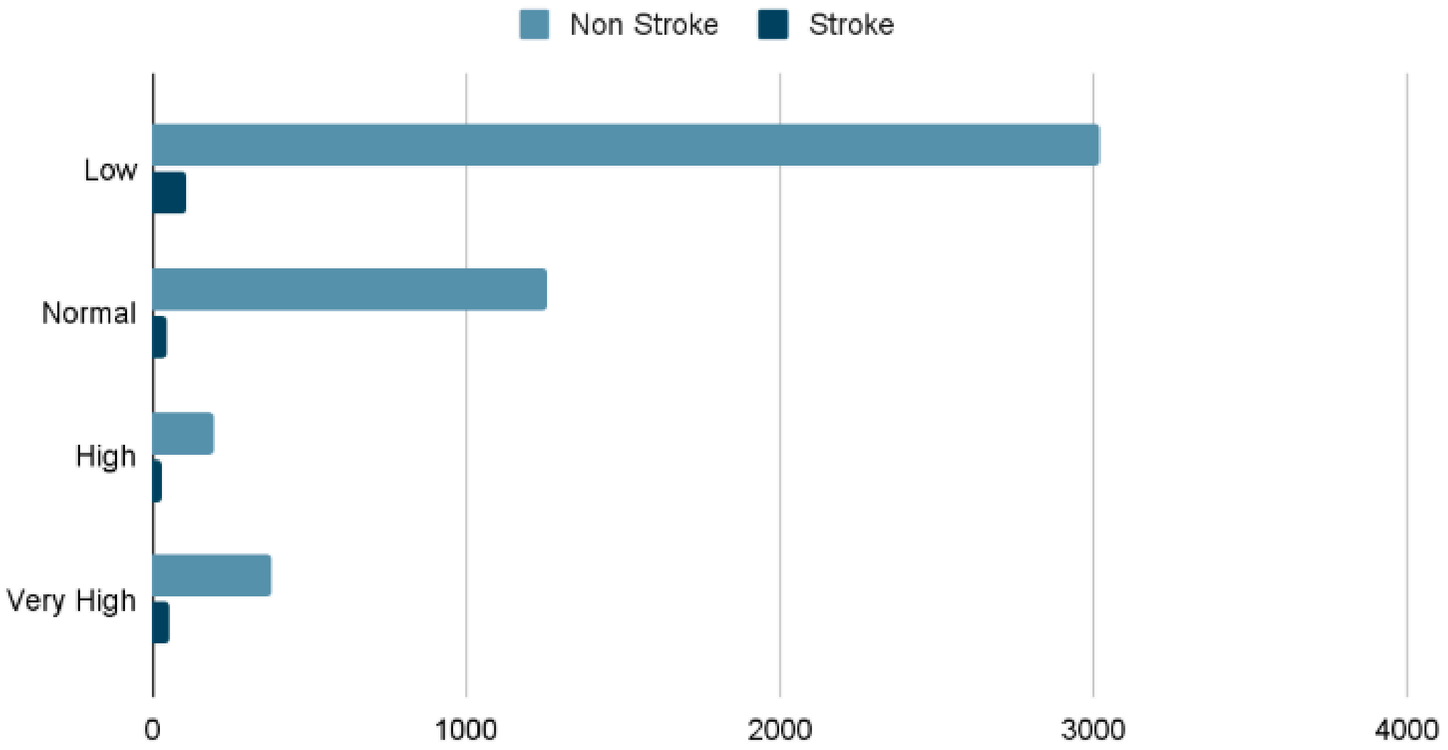
Heart Disease VS Stroke



This suggests that having heart disease increases the risk of stroke

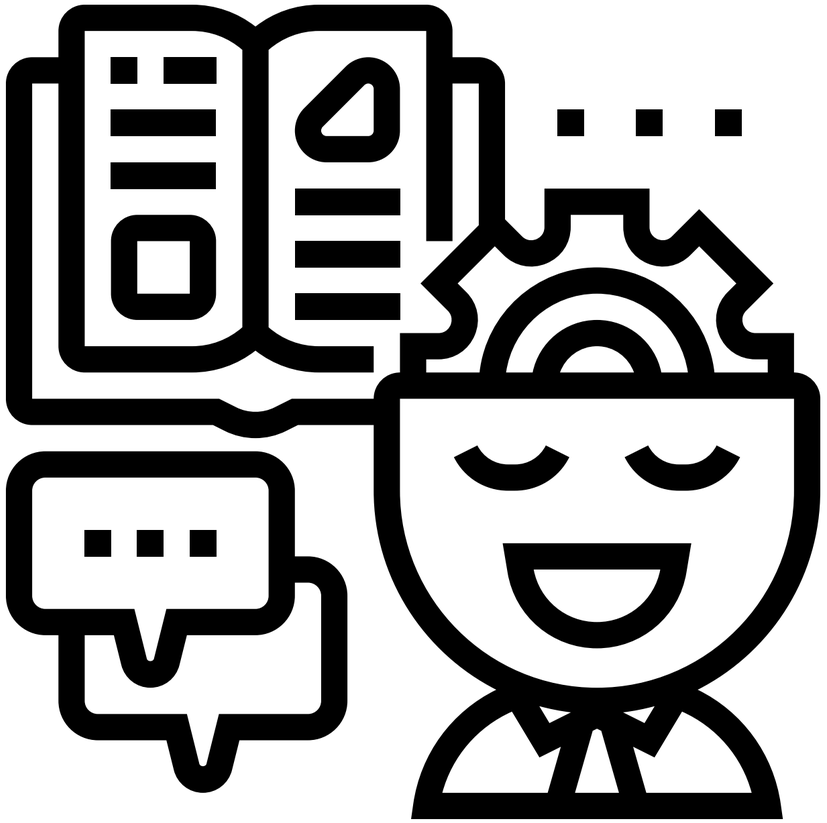
The majority of data come from individuals without heart disease, both in the stroke and non-stroke categories. This means that while heart disease increases the risk of stroke, stroke is less common than in the population without heart disease

Average Glucose Level VS Stroke



- This shows that high glucose levels can increase the risk of stroke. Most individuals in the "Non-Stroke" category have "Low" or "Normal" glucose levels

- there is a strong association between very high glucose levels and stroke risk. This may reflect the negative impact of diabetes or chronic hyperglycemia



DATA PRE-PROCESSING

DATA PRE-PROCESSING

Handling Missing value
BMI (with median)

Drop Unecessary Column
(ID column)

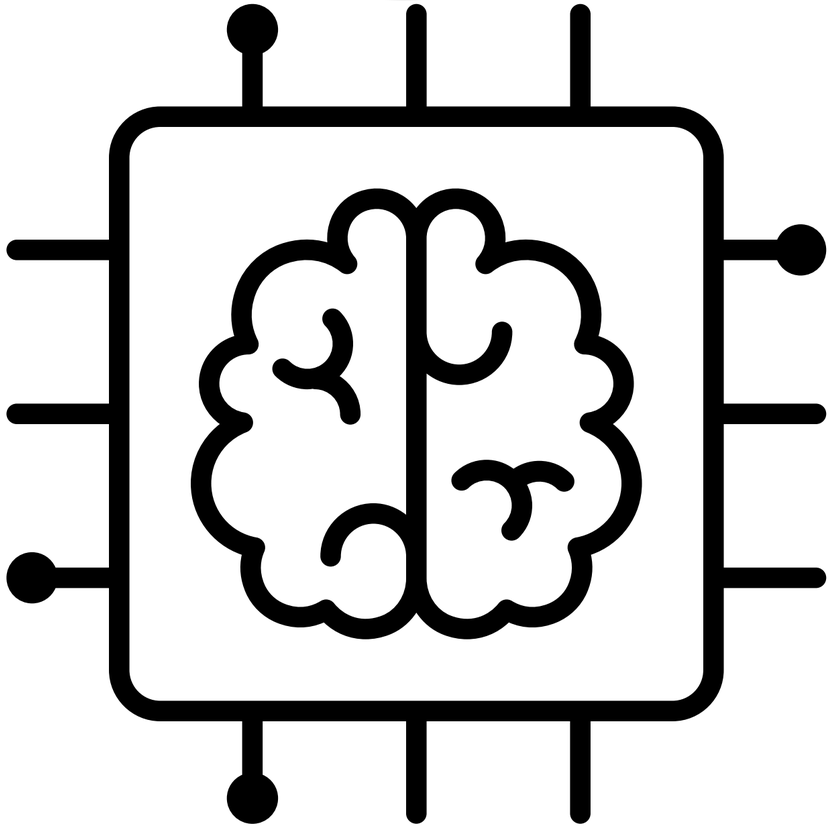
Feature Encoding (Label Encoder)

Handling Invalid values (Age and Avg_Glucose Level)

Data Splitting
Data Train : 60%
Data Val : 20%
Data Test :20%

Feature Scaling (Standar Scaler)

Handling Imbalance (SMOTE)



Modeling & Evaluation

Modeling and Evaluation

Model Evaluation Parameter

Positive = Stroke Patients Negative : non Stroke Patients	Model Prediction	Actual/Reality	Impact
False Positive Main Target to be reduce	Stroke ✓	Non-Stroke ✗	increasing Medical Costs and wasting medical resource
False Negative Second target to be reduced	Non- Stroke ✗	Stroke ✓	failures in providing appropriate treatment and increasing the risk of serious complications

Modeling and Evaluation

Model Evaluation Parameter

Recall

Mereduksi False Negative

Patients are predicted to Stroke but in reality they don't

Saving Lives
Increasing Customer Trust
Reducing legal Risk

Precision

Mereduksi False Positive

Patient are predicted non Stroke but in reality they having stroke

Optimization of medical resources
Cost efficiency
Enhancing reputation

F1-Score

Keseimbangan precision and recall
ensuring that at-risk patients are not missed (Recall) but also reducing false alarms (Precision)

More comprehensive model performance
Better decision-making
Minimizing risks and losses

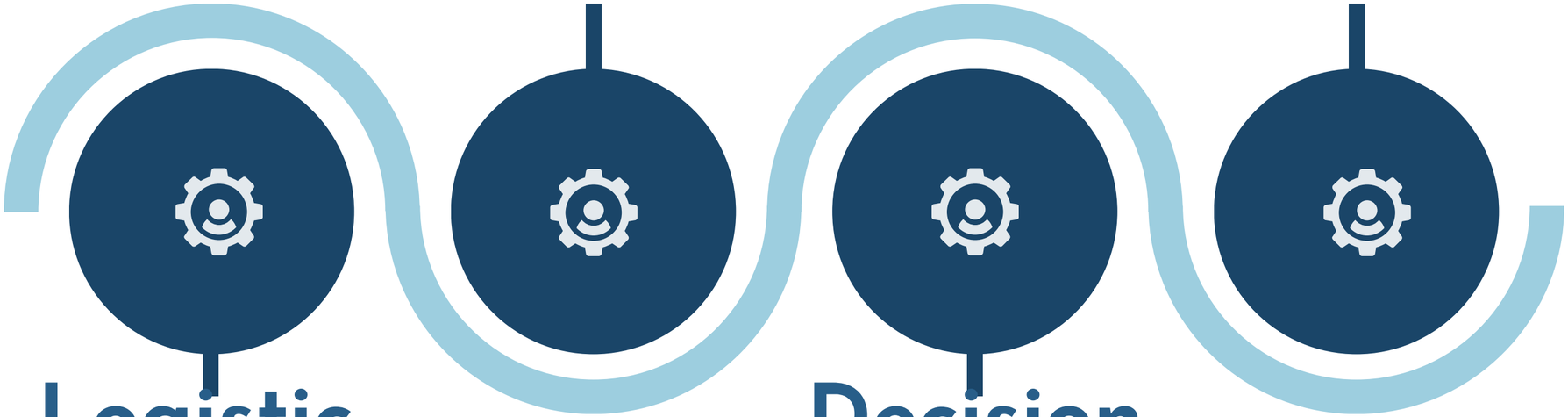
BEST PARAMETER MODELS

SVM

Best Hyperparameters: {'C': 9.588855372533333, 'gamma': 'auto', 'kernel': 'rbf'}

Random Forest

Best Hyperparameters for Random Forest: {'max_depth': None, 'min_samples_split': 2, 'n_estimators': 413}



Logistic Regression

Best Hyperparameters: C=0.01, class_weight='balanced', max_iter=1000, penalty='l1', solver='liblinear')

Decision Tree

Best Hyperparameters: {'criterion': 'gini', 'max_depth': 20, 'min_samples_split': 2}

Modeling and Evaluation

Model Comparison

Train data

Model	F1-Score	Precision	Recall	AUC ROC
Dummy Classifier	0.51	0.51	0.51	0.50
Logistic Regression	0.82	0.78	0.87	0.89
Decision Tree	0.94	0.90	0.98	0.99
SVM	0.82	0.79	0.85	0.90
Random Forest	0.95	0.91	0.99	0.99

Validation data

Model	F1-Score	Precision	Recall	AUC ROC
Dummy Classifier	0.10	0.05	0.56	0.55
Logistic Regression	0.09	0.04	1.0	0.55
Decision Tree	0.11	0.06	0.58	0.57
SVM	0.09	0.04	1.0	0.50
Random Forest	0.02	0.04	0.02	0.53

Machine Learning Techniques

- Logistic Regression
- Decision Tree
- SVM
- Random Forest

Logistic Regression

Best Fit Model

Modeling and Evaluation

Model Selection

Test Data

Model	F1-Score	Precision	Recall	AUC ROC
Logistic Regression	**0.18**	**0.10**	**0.60**	**0.74**
Random Forest	0.02	0.04	0.02	0.49

- High recall (0.60) ensures that more at-risk patients are detected.
- Low precision can be addressed with additional steps such as manual validation of positive predictions.
- A fairly good AUC-ROC (0.74) indicates potential for model improvement.

Modeling and Evaluation
Model Selection

60% Recall

10% Precision

18% F1-Score

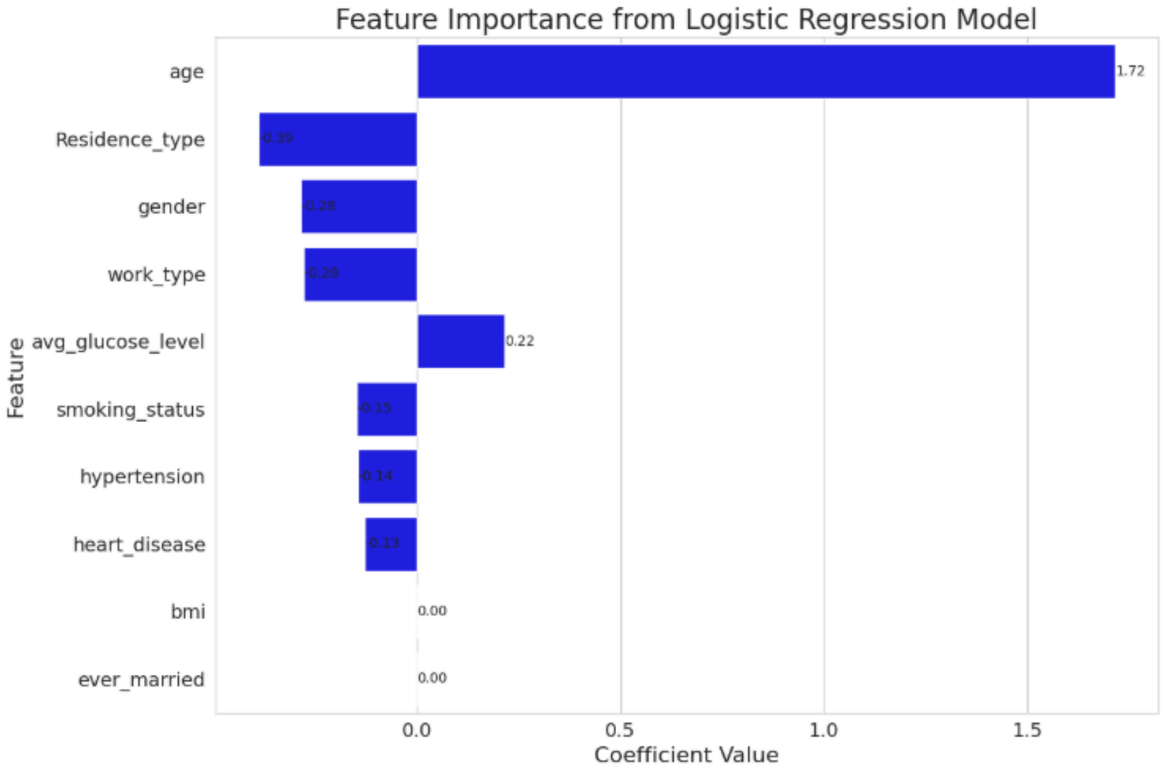
True Positive	True Negative	False Positive	False Negative
Prediction Stroke	Predicted Non Stroke	Predicted Stroke	Predicted Non Stroke
True, Stroke	True, Non Stroke	False, Non Stroke	False Stroke

Modeling & Evaluation

Feature importance

Age

The "age" feature has the highest coefficient value (1.72), indicating that age is the most significant factor in predicting stroke risk



Residence Type

The "Residence_type" feature also shows a substantial impact (0.39). This suggests that living location (urban or rural) may affect stroke risk, possibly due to differences in access to healthcare or lifestyle factors

Avg_Glucose Level

The "avg_glucose_level" feature (0.22) highlights that average blood glucose levels are a significant indicator of stroke risk, potentially linked to diabetes or metabolic disorders

Gender & Work Type

The "gender" and "work_type" features have equal influence (0.28). This indicates that gender and type of occupation also play a role in stroke risk, although to a lesser degree than age.

Business Recommendation

Business Recommendation

Early Risk Assessment Tools

Method: Integrate ML models with Electronic Health Record (EHR) systems for real-time risk prediction.

Goal: Enable early stroke detection, improve patient outcomes, and reduce treatment costs.

Ethical and Regulatory Compliance

Method: Ensure the model complies with HIPAA/GDPR and obtain certifications like FDA approval.

Goal: Build trust and ensure ethical deployment.

Business Recommendation

- **Method**: Offer subscription-based access to the ML model for stroke risk analysis.
- **Goal**: Generate recurring revenue and make the tool accessible for routine check-ups.

Subscription Model for Hospitals

- **Method**: Conduct webinars, publish case studies, and present at medical conferences.
- **Goal**: Raise awareness and boost adoption of the technology.

Educational Campaigns

Business Recommendation

Insurance Partnership for Preventive Care

- **Method**: Collaborate to integrate the ML model into insurance wellness programs, offering discounts for active participation.
- **Goal**: Incentivize preventive healthcare and reduce insurance claims.

Chronic Disease Expansion

- **Method**: Expand the model to include predictions for diabetes, heart disease, and hypertension.
- **Goal**: Diversify offerings and capture a larger market.

Thank You

Do you have any questions?

email : evinamin271@gmail.com
Link Dataset : Link
Link Script : Colab