# Introduction

Stroke is one of the most serious health problems in the world today. According to the World Health Organization (WHO), it's the second leading cause of death globally, responsible for around 11% of all deaths. Because of how sudden and dangerous strokes can be, being able to predict who might be at risk before it happens can help save lives.
This dataset was chosen because it contains real health-related information for each patient like age, gender, medical conditions, smoking habits, and more. These are the same kinds of factors that doctors consider when evaluating stroke risk.
We picked this dataset because it lets us build a machine learning model that can help predict the risk of stroke (classification task), using simple medical and lifestyle inputs. In real life, this can support healthcare professionals by giving early warnings and helping prioritize patients who are more likely to be at risk. Dataset link (Kaggle) Github link

# Dataset description

Before applying any preprocessing steps, our dataset contained 5110 samples and 12 columns, including the target variable (stroke). One of the features, **bmi**, had 201 missing values, which we handled by replacing them with the mean value of the column.
Additionally, we noticed that the target variable (stroke) is highly imbalanced, with most samples belonging to class 0 (no stroke). This imbalance created a challenge during model training, especially in correctly predicting the class (1 = stroke).
Below is a brief description of each feature in the dataset:
**Id**: A unique number for each patient → Numerical value (Not used for prediction).

**Gender**: Whether the patient is male, female, it is a Categorical type.

**Age**: Patient's age in years, Numerical value.

**Hypertension**: Has high blood pressure, Boolean type (1/0).

**Heart disease**: Has heart disease, Boolean type (1/0).

**Ever married**: Was ever married, Categorical type.

**Work type**: Type of, Categorical type.

**Residence type**: Lives in Urban or Rural area, Categorical type.

**Average glucose level**: Average blood sugar level, Numerical value.

**Bmi**: Body Mass Index, Numerical value.

**Smoking status**: Smoking habit, Categorical type.

**Stroke**: Had a stroke or not, Boolean type (1/0) the target.

# Preprocessing

We started by loading the dataset and removed the id column since it's just a unique identifier that doesn't help with prediction. Then, we converted the text columns like gender, marital status, work type, where the person lives, and smoking status into numbers using label encoding, because machine learning models need numerical data to work properly. After that, we noticed that the bmi column had

201 missing values, so we filled those missing spots with the average value of the column to avoid losing data.
Once the data was clean, we split it into features (everything we'll use to make predictions) and the target (stroke, which is what we're trying to predict). We then split this into training and testing sets, keeping 70% of the data for training and 30% for testing to make sure the model has enough data to learn from, while still keeping a good portion aside to check how well it performs on new, unseen data. Since our dataset has over 5,000 samples, using 70% (around 3,500 records) gives the model plenty of examples to learn patterns. The remaining 30% (around 1,500 records) is enough to fairly test how well the model can predict stroke cases in real life. This balance helps us avoid both overfitting and underfitting, and gives us a more honest view of how the model will work on future data. Finally, we scaled the data so that all the values are on a similar range this helps some models work better and learn more fairly. After all that, we saved the processed data into CSV files so we can use them for training and testing the models.

## Model training phase

We trained a variety of machine learning models to predict whether a person is at risk of having a stroke based on their health and lifestyle data. These models included Decision Tree, Random Forest, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Naive Bayes, Artificial Neural Network (ANN), and Logistic Regression. Each model was evaluated on how well it could detect stroke cases (1) and non-stroke cases (0) using metrics like precision, recall, and F1-score.

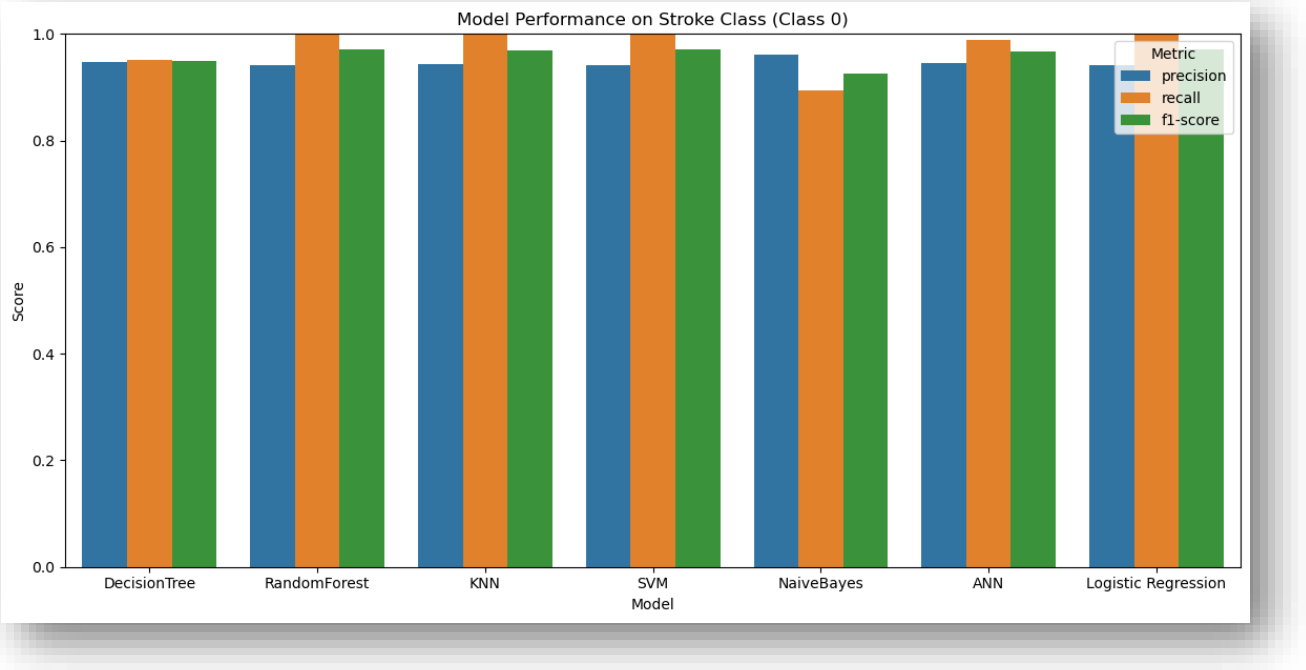
| Model | Accuracy | No Stroke Precision | No Stroke Recall | No Stroke F1-score |
|---|---|---|---|---|
| Decision Tree | 91% | 95% | 95% | 95% |
| Random Forest | 94% | 94% | 100% | 97% |
| KNN | 94% | 94% | 100% | 97% |
| SVM | 94% | 94% | 100% | 97% |
| Naive Bayes | 86% | 96% | 89% | 93% |
| ANN | 93% | 94% | 99% | 97% |
| Logistic Regression | 94% | 94% | 100% | 97% |

When looking at class 0 patients who didn't have a stroke most models performed very well. The highest accuracy came from models like Random Forest, KNN, SVM, and Logistic Regression, all reaching 94% accuracy, meaning they predicted non-stroke cases almost perfectly. This shows they are very good at identifying healthy patients and avoiding false alarms.

Naive Bayes had the lowest accuracy at 86%, but it still gave the best precision for class 0 (96%), meaning when it predicted "no stroke," it was usually right. This kind of result is helpful in real life, where doctors want to trust that someone predicted as low-risk truly is low-risk.

## Visualization

This is a bar plot, a simple and clear way to compare model performance side by side, each group of bars represents one model, and the three bars in each group show the precision, recall, and f1-score for class 0 which refers to patients who did not have a stroke.
This plot helps us quickly see which models are best at detecting non-stroke cases. For example, models like Random Forest, KNN, SVM, and Logistic Regression all reach perfect recall, meaning they correctly identified all non-stroke cases. We can also notice that Naive Bayes, while still strong, has slightly lower recall and f1-score compared to the others. This visual summary makes it easy to compare how well each model performs in identifying safe (non-stroke) patients.




## Conclusion

I chose this dataset because stroke is a major public health concern and one of the leading causes of death worldwide. Being able to predict stroke risk based on health and lifestyle data can support early diagnosis and save lives. The dataset contains key features such as age, hypertension, heart disease, glucose level, BMI, smoking status, and more, which are highly relevant for stroke prediction. In real life, using such a model in hospitals or clinics can help doctors prioritize high-risk patients and allocate resources efficiently. After preprocessing, including filling missing BMI values and encoding categorical features, I split the data into training and testing sets (70/30) and applied standard scaling. I trained seven classification models to evaluate their performance, including Random Forest, Logistic Regression, SVM, KNN, ANN, Decision Tree, and Naive Bayes. Most models performed very well for class 0 (non-stroke cases), with Random Forest, KNN, SVM, ANN, and Logistic Regression

reaching 94% accuracy and perfect recall for detecting non-stroke patients. Naive Bayes, although it had lower accuracy (86%), still provided the highest precision (96%) for class 0, meaning its “no stroke” predictions were usually correct. Visualizing the results with a bar chart made it easy to compare how well each model predicted non-stroke cases. One insight I learned is that precision and recall need to be balanced, especially in medical applications. A model that identifies all non-stroke cases correctly (high recall) is great, but it should also be trustworthy when saying someone is safe (high precision). I also found that class imbalance was a challenge, and focusing only on class 0 can give a misleading impression of overall model quality. Overall, this project helped me understand the importance of preprocessing, model selection, and proper evaluation in healthcare-focused machine learning problems.