Proposal: Predicting Stroke Risk Using Machine Learning

Team Members - Alexandra , Ryan and George

Industry - Healthcare

Problem/Question:

We aim to develop a machine learning model that predicts whether a patient is likely to have a stroke based on various health-related attributes such as age, gender, hypertension, heart disease, marital status, occupation, residence type, average glucose level, body mass index (BMI), and smoking status. Stroke is a significant global health concern, and early identification of at-risk individuals can lead to preventive measures and better healthcare outcomes. Our primary question is: Can we accurately predict the likelihood of a patient having a stroke based on their health-related data?

Data Source:
https://www.kaggle.com/datasets/
We will use a dataset containing the following attributes:

1. **id**: Unique identifier
2. **gender**: Categorical (Male, Female, Other)
3. **age**: Numeric (Age of the patient)
4. **hypertension**: Binary (0 if no hypertension, 1 if hypertension)
5. **heart_disease**: Binary (0 if no heart disease, 1 if heart disease)
6. **ever_married**: Categorical (No, Yes)
7. **work_type**: Categorical (Children, Govt_job, Never_worked, Private, Self-employed)
8. **Residence_type**: Categorical (Rural, Urban)
9. **avg_glucose_level**: Numeric (Average glucose level in blood)
10. **bmi**: Numeric (Body mass index)
11. **smoking_status**: Categorical (Formerly smoked, Never smoked, Smokes, Unknown)
12. **stroke**: Binary (1 if the patient had a stroke, 0 if not)

Machine Learning Algorithms:
We will explore various machine learning algorithms for classification, including:

- Logistic Regression
-Neural Networks

These algorithms will help us build predictive models to classify patients into stroke and non-stroke groups based on the provided attributes.

Exploratory Data Analysis (EDA):
For exploratory data analysis and to gain insights into the dataset, we will employ the following visualizations:
1. **Histograms and Box Plots**: To visualize the distribution of continuous variables like age, avg_glucose_level, and BMI.
2. **Count Plots**: To explore the distribution of categorical variables like gender, work_type, and smoking_status.

Data Preprocessing and Analysis:
- Handling missing data: We will address missing values, especially in the "smoking_status" attribute.
- Encoding categorical variables: Convert categorical variables into numerical format using techniques like one-hot encoding.
- Data splitting: Split the dataset into training and testing subsets.
- Model training and evaluation: Train machine learning models on the training data and evaluate their performance using metrics such as accuracy, precision, recall, and F1-score.
- Hyperparameter tuning: Optimize model hyperparameters to achieve the best possible predictive performance.
- Model interpretation: Examine feature importances to understand which factors are most influential in stroke prediction.

Next Steps:
Once we have a reliable predictive model, we can use it to identify individuals at high risk of stroke. Healthcare providers can then intervene early with preventive measures and personalized care plans. Furthermore, the model can assist in public health initiatives and resource allocation for stroke prevention.

In summary, our project aims to leverage machine learning to predict stroke risk based on patient data, potentially improving stroke prevention and patient outcomes. We are committed to addressing this critical health issue with responsible data handling and model evaluation.

Final Project team members contribution :

1- Find a problem worth solving, analyzing, or visualizing
https://www.kaggle.com/datasets/ in csv format.

2- prepare the data using python pandas. George

3- Machine Learning 75% accuracy & 80% R-squared (Team)

4- visualization Ryan

5- README George

6- Presentation Alexandra