Context

According to the World Health Organization (WHO) stroke is the 2nd leading cause of death globally, responsible for approximately 11% of total deaths.

This dataset is used to predict whether a patient is likely to get stroke based on the input parameters like gender, age, various diseases, and smoking status. Each row in the data provides relevant information about the patient.

Attribute Information

1) id: unique identifier

2) gender: "Male", "Female" or "Other"

3) age: age of the patient

4) hypertension: 0 if the patient doesn't have hypertension, 1 if the patient has hypertension

5) heart\_disease: 0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease

6) ever\_married: "No" or "Yes"

7) work\_type: "children", "Govt\_jov", "Never\_worked", "Private" or "Self-employed"

8) Residence\_type: "Rural" or "Urban"

9) avg\_glucose\_level: average glucose level in blood

10) bmi: body mass index

11) smoking\_status: "formerly smoked", "never smoked", "smokes" or "Unknown"\*

12) stroke: 1 if the patient had a stroke or 0 if not

\*Note: "Unknown" in smoking status means that the information is unavailable for this patient.

1. ***Stroke Prediction:*** We will be applying Support Vector Machines to solve this problem. SVM is the most extensively used algorithm in the field of healthcare because of some advantages it provides. Therefore, it is necessary to get a hold of this algorithm which will ultimately be very useful when applying it in the healthcare industry. We will also be performing some extensive exploratory data analysis for this project and will hone your EDA skills as well.

**Algorithm:**

This can be implemented using Support Vector Machines. It is advantageous for applications with a small sample size. SVM has demonstrated high performance in solving classification problems in bioinformatics.These are the reasons why it is used so extensively in the healthcare sector.

Fit the data with a linear SVM. Import the library as:

from sklearn.svm import SVC

Now, .fit a Gaussian kernel SVC and see how the decision boundary changes. Use the “rbf” kernel. Apply this using this function:

SVC\_Gaussian = SVC(kernel=’rbf’)

You can also use the Nystroem method. Import the library as:

from sklearn.kernel\_approximation import Nystroem

**Implementation:**

First of all, do some data cleaning. A caveat for using this data set is that it has certain null values and outliers, you can either delete them or replace them with a median value. After that, perform data visualization to understand the underlying relationships and dependencies within the data. Create cat plots, heatmaps, pairplots and boxplots for different features of the data set to look for any relationships between the features and the target variable.

After that, split the data into train and test sets. Train and then predict the random forest model on the data set. In the end get the precision, recall, accuracy scores to check the model performance. From sklearn.metrics, you can import classification\_report, accuracy\_score, precision\_score, recall\_score to check the performance metrics.

**Heart Failure Prediction Dataset**

Cardiovascular diseases (CVDs) are the number 1 cause of death globally, taking an estimated 17.9 million lives each year, which accounts for 31% of all deaths worldwide. Four out of 5CVD deaths are due to heart attacks and strokes, and one-third of these deaths occur prematurely in people under 70 years of age. Heart failure is a common event caused by CVDs and this dataset contains 11 features that can be used to predict a possible heart disease.

People with cardiovascular disease or who are at high cardiovascular risk (due to the presence of one or more risk factors such as hypertension, diabetes, hyperlipidaemia or already established disease) need early detection and management wherein a machine learning model can be of great help.

Attribute Information

1. Age: age of the patient [years]
2. Sex: sex of the patient [M: Male, F: Female]
3. Chest Pain Type: chest pain type [TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic]
4. Resting BP: resting blood pressure [mm Hg]
5. Cholesterol: serum cholesterol [mm/dl]
6. Fasting BS: fasting blood sugar [1: if FastingBS > 120 mg/dl, 0: otherwise]
7. Resting ECG: resting electrocardiogram results [Normal: Normal, ST: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV), LVH: showing probable or definite left ventricular hypertrophy by Estes' criteria]
8. Max HR: maximum heart rate achieved [Numeric value between 60 and 202]
9. Exercise Angina: exercise-induced angina [Y: Yes, N: No]
10. Oldpeak: oldpeak = ST [Numeric value measured in depression]
11. ST\_Slope: the slope of the peak exercise ST segment [Up: upsloping, Flat: flat, Down: downsloping]
12. Heart Disease: output class [1: heart disease, 0: Normal]

Source

This dataset was created by combining different datasets already available independently but not combined before. In this dataset, 5 heart datasets are combined over 11 common features which makes it the largest heart disease dataset available so far for research purposes. The five datasets used for its curation are:

* Cleveland: 303 observations
* Hungarian: 294 observations
* Switzerland: 123 observations
* Long Beach VA: 200 observations
* Stalog (Heart) Data Set: 270 observations

Total: 1190 observations  
Duplicated: 272 observations

Final dataset: 918 observations

Every dataset used can be found under the Index of heart disease datasets from UCI Machine Learning Repository on the following link: <https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/>

Citation

fedesoriano. (September 2021). Heart Failure Prediction Dataset. Retrieved [Date Retrieved] from <https://www.kaggle.com/fedesoriano/heart-failure-prediction>.