LOGISTIC REGRESSION

Logistic regression aims to solve classification problems.It is used to calculate or predict the probability of a binary (yes/no) event occurring.

AIMS AND OBJECTIVES OF THE PROJECT

- The prime objective of this project is to construct a prediction model for predicting stroke using machine learning algorithms.
- The prediction of long-term outcomes in brain stroke patients may be useful in treatment decisions.
- Machine learning techniques are being increasingly adapted for use in the medical field because of their high accuracy.
- The purpose of this study is to see whether machine learning techniques might be used to predict long-term outcomes for patients who had suffered a brain stroke.
- The aim of this study is to apply computational methods using machine learning techniques to predict stroke

METHODOLOGY

DATA COLLECTION AND DESCRIPTION OF THE DATASET

The dataset for stroke prediction is from Kaggle [3]. There are 11 columns and 4981 rows in this particular dataset. As the primary attributes, the columns have the following information: "gender," "age," "hypertension," "heart disease," "ever-married," "work type," "Residence type," "avg glucose level," "bmi," "smoking status," and "stroke". The output column 'stroke' has the value as either '1' or '0'. The value '0' indicates no stroke risk detected, whereas the value '1' indicates a possible risk of stroke. This dataset is imbalanced as the possibility of '0' in the output column ('stroke') outweighs that of '1' in the same column. Only 248 rows have the value '1' whereas 4733 rows with the value '0' in the stroke column. For better accuracy, data pre-processing is performed to balance the data. The dataset discussed above is summarized in Table 1.

sl.no	Attributes	Description	Range of Values
1	gender	Gender of the person [Male,Female or other]	0,1

2	age	Age of the patient in years	0.08-82
3	hypertension	0 if the patient doesn't have hypertension, 1 if the patient has hypertension	0,1
4	Heart_disease	0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease	0,1
5	ever_married	Marital status of the patient[No,Yes]	0,1
6	work_type	Categories of work type of patient [children, Govt_job,Never_worked,Private or Self-employed]	0,1,2,3,4
7	Residence_type	Patient's residence type [Rural,Urban]	0,1
8	Avg_glucose_level	average glucose level in blood	55-291.0
9	bmi	value of the patient's Body Mass Index	10.1- 97.6
10	Smoking_status	Smoking status of patient [formerly smoked, never smoked, smokes or Unknown]	0,1,2,3
11	stroke	1 if the patient had a stroke or 0 otherwise	0,1

CLEANING DATASET

Checking for null values in the dataset. In this dataset there is no null value found.

DATA ANALYSIS

DATA VISUALIZATION

The process of finding trends and correlations in our data by representing it pictorially is called Data Visualization. To perform data visualization in python, we can use various python data visualization modules such as Matplotlib, Seaborn etc. Data Visualization is used to communicate information clearly and efficiently to users by the usage of information graphics such as tables and charts. It helps users in analyzing a large amount of data in a simpler way.

DATA PREPROCESSING

Data Preprocessing is required before model building to remove the unwanted noise and outliers from the dataset, resulting in a deviation from proper training. Anything that interrupts the model from performing with less efficiency is taken care of in this stage. After collecting the appropriate dataset, the next step lies in cleaning the data and making sure that it is ready for model building.

LABEL ENCODING

The following step is label encoding, which comes after removing the null values from the dataset. Label encoding encodes the string literals in the dataset into integer values for the machine to understand them. Since the computer is often trained on numbers, the strings must be converted into integers.There are five columns in the collected dataset that have strings as their data type. When label encoding is applied, all the strings are encoded, turning the entire dataset into a collection of numbers.

SPLITTING THE DATA

After completing data preprocessing, the next step is building the model. The data is split into training and testing data for better accuracy and efficiency for this task keeping the ratio as 80% training data and 20% testing data. After splitting, logistic regression is used to train the model.

LOGISTIC REGRESSION

Logistic Regression is a supervised learning algorithm used for predicting the probability of the output variable. This algorithm is the best fit when the output variable has binary values (0 or 1). As the output attribute in the dataset has only two possible values, Logistic Regression is opted. After performing this algorithm on the dataset, the accuracy obtained is 94%. Efficiency of this algorithm can also be found by using various other accuracy metrics like precision score and recall score. Precision score obtained is 0.5 whereas Recall score is 0.02. From this you can see that the model is very biased towards 0 value samples

HANDLING IMBALANCED DATA

The dataset chosen for the task of stroke prediction is highly imbalanced. The entire dataset has 4981 rows, of which 249 rows are suggesting the occurrence of a stroke and 4733 rows

having the possibility of no stroke. Training a machine-level model with such data might give accuracy, but other accuracy metrics like precision and recall are shallow. If such imbalanced data is not handled, the results are not accurate, and the prediction is inefficient. Therefore, to get an efficient model, this imbalanced data is to be first handled. For this purpose, the method of undersampling is used. Undersampling balances the data wherein the majority class is undersampled to match the minority class. In this case, the class with a value as '0' is undersampled for the class with the value' 1'. So after undersampling the resulting dataset will have 249 rows with value ‘0’ and 249 rows with value ‘1’.

UNDER SAMPLING

Correcting imbalance data to reduce risk of their analysis skewing toward the majority.The aim of undersampling is to reduce the number of samples in your majority class so that they match upto the total number of samples in your minority class.

Again we will follow the same process we will split X train, X test, Y train and Y test, build the logistic Regression.
Now while generating the model report, Precision, Recall and F1 score will be better

CONFUSION MATRIX

Confusion Matrix is the easiest way to determine the performance of a classification model by comparing how many positive instances were correctly/incorrectly classified and how many negative instances were correctly/incorrectly classified. In a Confusion Matrix, the rows represent the actual labels and the columns represent the predicted labels

True Positives (TP): True positives are the instances where both the predicted class and actual class is True (1), i.e., when patient actually has complications and is also classified by the model to have complications.

True Negatives (TN): True negatives are the instances where both the predicted class and actual class is False (0), i.e., when a patient does not have complications and is also classified by the model as not having complications.

False Negatives (FN): False negatives are the instances where the predicted class is False (0) but actual class is True (1), i.e., when a patient is classified by the model as not having complications even though in reality, they do.

False Positives (FP): False positives are the instances where the predicted class is True (1) while actual class is False (0), i.e., when a patient is classified by the model as having complications even though in reality, they do not.

ACCURACY OF THE MODEL

ACCURACY

Accuracy determines the number of correct predictions over the total number of predictions made by the model.The formula for Accuracy is:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

PRECISION

Precision is a measure of the proportion of patients that actually had complications among those classified to have complications by the system. The formula for Precision is:

$$Precision = \frac{TP}{TP + FP}$$

RECALL

Recall or sensitivity is a measure of the proportion of patients that were predicted to have complications among those patients that actually had the complications. The formula is:

$$Recall = \frac{TP}{TP + FN}$$

F1 SCORE

F1 Score is the harmonic mean of the Recall and Precision that is used to test for Accuracy. The formula is:

$$F1\ 2 * \frac{Precision * Recall}{Precision + Recall}$$

FUTURE ENHANCEMENT

We can use a variety of machine learning techniques such as Naive Bayes,Random Forest,K-Nearest Neighbors, Decision tree etc. to predict brain stroke and then We can determine which algorithm is better at predicting strokes.

In the future we will extend our work with various deep learning mechanisms using big data to predict stroke risk and analyze the performance.