Stroke is a medical emergency. A stroke occurs when blood flow to a part of your brain is interrupted or reduced, preventing brain tissue from getting oxygen and nutrients. Brain cells begin to die within minutes

Risk factors for having a stroke include:

Age: People aged 55 years and over

Hypertension: if the systolic pressure is 140 mm Hg or more, or the diastolic pressure is 90 mm Hg or more

Hypercholesterolemia: If the cholesterol level in the blood is 200 milligrams per deciliter

Smoking

Diabetes

Obesity: if the body mass index (BMI) is 30 or more

We have good distribution for age

I think we have outliers in bmi

Avg glucose distribution is reasonable because the normal avg of blood in sugar is less than 140 , that may be not good this feature will not be helpful to know if diabetes have correlation between diabetes and strokes

Questions to ask:

1) Male/Female who has more strokes.

2) People of which age group are more likely to get a stroke.

3) Is hypertension a cause?

4) A person with heart disease is more likely to get a stroke (need confirmation).

5) Marriage may be a cause of strokes.

6) People working in private jobs may be the majority of people with strokes(mostly cause of stress).

7) People living in urban areas have more chances of getting stroke? (need to confirm)

8) Glucose levels are important and must be observed closely with other things.

9) BMI must be closely observed with age and gender.

10) People who smoke are more likely to get a stroke (need confirmation).

drop id column, check null values if avaible then impute , then convert all object dtatypes columns to dummy variables,💡

In order to make the dataset balanced we will use the package SMOTE for oversampling. Moreover to fix the missing values in the BMI columns we use a imputation technics based on the KNN.

In our dataset, we have both numerical and categorical variables. It is essential to see whether columns are correctly inferred. The most important one to look for is our target variable 'stroke' 'Stroke' is detected as an integer, not as an object. Target variable is coded as 1 for positive cases (has a stroke) and 0 for negative cases (does not have a stroke) Both 'Hypertension' and 'heart disease" are detected as an integer, not as an object. Just remember from the data definition part, they are coded as 1 for the positive cases(has hypertension/heart disease) And 0 for the negative cases (does not have hypertension/heart disease) We don't need to change them, but it is good to see and be aware of it. In addition to them, we have 3 categorical variables, which we have to encode as numerical.

dataset.bmi.replace(to\_replace=np.nan, value=dataset.bmi.mean(), inplace=True)

It is essential to see whether columns are correctly inferred.

The most important one to look for is our target variable 'stroke' 'Stroke' is detected as an integer, not as an object. Target variable is coded as 1 for positive cases (has a stroke) and 0 for negative cases (does not have a stroke) Both 'Hypertension' and 'heart disease" are detected as an integer, not as an object. Just remember from the data definition part, they are coded as 1 for the positive cases(has hypertension/heart disease) And 0 for the negative cases (does not have hypertension/heart disease) We don't need to change them, but it is good to see and be aware of it. In addition to them, we have 3 categorical variables, which we have to encode as numerical.

One of the first steps of exploratory data analysis should always be to look at what the values of y look like.

We see that the target is skewed and thus the best metric for this binary classification problem would be Area Under the ROC Curve (AUC).We can use precision and recall too, but AUC combines these two metrics.

We have already seen the label/target distribution, and we know that it is a binary classification problem with skewed targets. Thus, we will be using StratifiedKFold to split the data

Just for further info, it is not advisable to use accuracy as an evaluation metric, when dealing with higly imbalanced data.

Instances across classes are imbalanced, like in our dataset, we have imbalance data.

The problem is, most of the machine learning algorithm do not work well with the imbalanced data.

Some of the metrics (like accuracy) give us misleading results.

Most of the time in classification problems our interest is to get better predict on the minority class.

In our example: People had a stroke is minority class.

Otherwise our machine learning algorithm falsely predicts majority class.

In our example: No stroke is majority class.

We have three numerical features in our dataset.

All of our numerical features are measured in different scales.

Many machine learning algorithms perform better standard range scaled numerical variables (such as Linear models,artificial neural networks, K-nearest Neighbors,support vector machines, etc.)

Tree models (such as, decision trees,random forest, etc.) work fine with different range numerical features.

decision trees,random forest,neural,nets,svm,logistic Based on the mean & median score differences, we can expect Slightly left skew on the 'age' (mean: 43.22 & median: 45) Slightly right skew on the 'bmi' (mean: 28.89 & median: 28.10)

And right skew distribution on the 'avg\_glucose\_level' (mean: 106.14 & median: 91.88)

Let's see the skewness.



Conclus

One of the first steps of exploratory data analysis should always be to look at what the values of y look like.

Just for further info, it is not advisable to use accuracy as an evaluation metric, when dealing with higly imbalanced data.

Instances across classes are imbalanced, like in our dataset, we have imbalance data.

The problem is, most of the machine learning algorithm do not work well with the imbalanced data.

Age and target variable weak positive relationship (almost .25). Average glucose level's mean scores on the target have differences between a person who has a stroke or not. But this differences are small. BMI does not have any significant relationship with the target variable. A person with hypertension are almost 3.3 time more likely to get stroke than the ones who don't have hypertension. Male compare to female are more likelyto get stroke, but difference between female and male is very small. A person with heart diease are 4.07 times more likely to get stroke than the ones who don't have heart disease. A person is married(or married before) are 5.7 times more likely to get stroke than the ones who don't have marriage history. Self employed person has more probability to get stroke than other work type. Be carefull !!! Person who lives in rural area slightly has more probablity to get sroke than a person who lives in rural area. Difference is small. It is smal difference between who smokes and who does not smoke in regard to probability of getting stroke.

Stroke Prediction Research:

The main question is that we want to understand how the predictor variables can help estimate the probability of sufferign a stroke.

Is there other than age relationship?

Does having a heart disease or high BMI and glucose level related to have a higher change of suffering a stroke?

Some of the metrics (like accuracy) give us misleading results.

Most of the time in classification problems our interest is to get better predict on the minority class.

In our example: People had a stroke is minority class.

Otherwise our machine learning algorithm falsely predicts majority class.

In our example: No stroke is majority class.

Age and target variable weak positive relationship (almost .25). Average glucose level's mean scores on the target have differences between a person who has a stroke or not. But this differences are small. BMI does not have any significant relationship with the target variable. A person with hypertension are almost 3.3 time more likely to get stroke than the ones who don't have hypertension. Male compare to female are more likelyto get stroke, but difference between female and male is very small. A person with heart diease are 4.07 times more likely to get stroke than the ones who don't have heart disease. A person is married(or married before) are 5.7 times more likely to get stroke than the ones who don't have marriage history. Self employed person has more probability to get stroke than other work type. Be carefull !!! Person who lives in rural area slightly has more probablity to get sroke than a person who lives in rural area. Difference is small. It is smal difference between who smokes and who does not smoke in regard to probability of getting stroke.

​

Stroke Prediction Research:

​

The main question is that we want to understand how the predictor variables can help estimate the probability of suffering a stroke.

​

   Is there other than age relationship?

   Does having a heart disease or high BMI and glucose level related to have a higher change of suffering a stroke?