STROKE PREDICTION

SUBMITTED BY
AMRUTHA R

DATA OVERVIEW

The dataset used is a healthcare dataset. Repository located at the following URL: https://www.kaggle.com/fedesoriano/stroke-prediction-dataset . There are 5110 observations and 11 Variables in the Data Set. There are 6 continuous measure variables and 5 categorical variables. The target response (y) is a binary response indicating whether a person had a stroke or not .

GOAL

1. Build a model that predicts whether a person had a stroke or not .

SPECIFICATIONS

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_job, 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

APPROACH

Preprocessing :

- **Handling missing values:**

  Bmi contains missing values and replacing the null values of bmi with mean of that feature.

- **Outliers handling**

  Outliers can be quickly identified by using boxplot for visualizing boxplot use seaborn . All columns that contains outliers are handled by using upper threshold and lower threshold.

- **Exploratory data analysis**

  Firstly considered categorical variables and then handle the categorical variables using label encoder.

- **Remove unnecessary columns**

  Remove the unwanted features like ID it have no relevance in ML modeling. So will not consider that features for ML model training.

- **Creating X and y Datasets**

  Need to split the data into independent variable and dependent variable (target variable)

  X - Set of independent variable

  y - series of dependent variable

  In this usecase, target variable is 'SalePrice'

Model building :

- Smoting used because of high imbalance in output class.

- **Algorithms used**
    1. K Nearest Neighbor
    2. Random Forest Classifier
    3. XGBoost

RESULT

All models performed with accuracy around 80% except K Nearest Neighbor it showed accuracy of 78 %. Grid Search CV used for parameter tuning. The most suitable to model among these models is XGBoost algorithm with parameter tuning . It give accuracy of 87%.