Class: DSC630 T302 2231-Fall2022
Course: Predictive Analytics
Author: Anjani Bonda
Date: Nov 18, 2022

Term Project Final Report
Stroke Prediction based on risk factors

Introduction:

This is something personal topic to me as I have seen a lot of people in my family and friends experienced a stroke. A stroke, sometimes called a brain attack, occurs when something blocks blood supply to part of the brain or when a blood vessel in the brain bursts. In either case, parts of the brain become damaged or die. A stroke can cause lasting brain damage, long-term disability, or even death. The brain controls our movements, stores our memories, and is the source of our thoughts, emotions, and language. The brain also controls many functions of the body, like breathing and digestion. To work properly, your brain needs oxygen. Your arteries deliver oxygen-rich blood to all parts of your brain. If something happens to block the flow of blood, brain cells start to die within minutes**,** because they can't get oxygen. This causes a stroke.

Common Symptoms:

- Sudden numbness or weakness in the face, arm, or leg, especially on one side of the body.
- Sudden confusion, trouble speaking, or difficulty understanding speech.
- Sudden trouble seeing in one or both eyes.
- Sudden trouble walking, dizziness, loss of balance, or lack of coordination.
- Sudden severe headache with no known cause.

Facts & Stats per CDC:

- In 2020, 1 in 6 deaths from cardiovascular disease was due to stroke.
- Every 40 seconds, someone in the United States has a stroke. Every 3.5 minutes, someone dies of stroke.2
- Every year, more than 795,000 people in the United States have a stroke. About 610,000 of these are first or new strokes.
- About 185,000 strokes—nearly 1 in 4—are in people who have had a previous stroke.2
- About 87% of all strokes are ischemic strokes, in which blood flow to the brain is blocked.

- Stroke-related costs in the United States came to nearly $53 billion between 2017 and 2018. This total includes the cost of health care services, medicines to treat stroke, and missed days of work.
- Stroke is a leading cause of serious long-term disability.2 Stroke reduces mobility in more than half of stroke survivors age 65 and older.
- Stroke is a leading cause of death for Americans, but the risk of having a stroke varies with race and ethnicity.
- Risk of having a first stroke is nearly twice as high for Blacks as for Whites, and Blacks have the highest rate of death due to stroke.
- Though stroke death rates have declined for decades among all race/ethnicities, Hispanics have seen an increase in death rates since 2013.

Some of the risk factors that can possibly increase one's risk of stroke include:

- Age
- Sex
- Race/Ethnicity
- High Blood Pressure
- High Cholesterol
- Heart Disease
- Diabetes
- Unhealthy Diet/Physical Inactivity
- Obesity
- Tobacco/Alcohol Use

Problem Statement/Project Proposal:

How can we predict if a person is likely to experience a stroke?
What's the likeliness of a stroke based on his/her available medical history?

The project proposal therefore is "Stroke Prediction" using common risk factors and applying predictive modeling techniques.

Dataset Used:

The dataset used for this project is sourced from Kaggle.com. This dataset is a popular one from Kaggle with about 12 attributes and 5110 observations. Here is the link for the data source being used.

https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset?resource=download

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

Methodology:

This project will follow CRISP-DM model for data exploration/understanding, modeling, and evaluation. Python programming language will be used for data load, exploration, fitting the models and testing the accuracy along with some of the machine learning libraries that might be required.

Firstly, EDA will be performed to better understand the data and compute descriptive statistics followed by visualizations representing the dataset.

Secondly, various machine learning algorithms will be employed to examine and choose the best possible prediction model. Some of them include Logistic Regression, K Nearest Neighbor (KNN), Decision Tree, Naïve Bayes, Support Vector Machine (SVM) and Random Forest Classifier. I would prefer to apply and check the accuracy of these models as these are mostly standard and widely used ones for building predictive models like this one.

Data Analysis:

Initially, exploratory data analysis was performed to understand the data better and compute some major descriptive statistics. I produced some illustrative visualizations representing the dataset as well. Then, various machine learning algorithms were employed to see the best- resulting model for the prediction model. Mainly, SVM, Logistic Regression, Decision Tree, Random Forest, and KNN are the potential algorithms used.

The collected dataset is analyzed using Python programming language, and Jupyter Notebook is used as a scripting interface. Several python libraries are used for the various purposes of exploratory data analysis, visualization, and machine learning model creation. Some of the notable ones are:
Numpy
Pandas
Matplotlib
Seaborn

Sklearn

After loading the dataset into the pandas data frame, different measures about the dataset were generated, such as the shape of the dataset, data types, missing values, duplicated observations, etc. Then Exploratory data analysis was performed using visualization tools.
Then the critical step of the project, modeling, was performed. Finally, data scaling, numerical and categorical data, and encoding were accomplished.

Data was trained with several classification algorithms as the data processing was complete. Trained models were attempted to improve by hyper-parameter tuning. Data was first trained with Logistic Regression, SVM, GaussianNB, BernoulliNB, Decision tree, Random Forest classifier- Nearest neighbor. Fine-tuning of each of those models was performed and evaluated in each step with testing data. The model evaluation mainly used accuracy, F1 score, and whole classification report. In the case of the KNN classifier, various k values were tested, and the best one was selected to train the model.

Results:

Exploratory Data Analysis

Dataset was imported in Python's Jupyter notebook. Initially, the dataset was imported as a panda's data frame. Here's what the dataset looked like.

	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
0	9046	Male	67.0	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
1	51676	Female	61.0	0	0	Yes	Self-employed	Rural	202.21	NaN	never smoked	1
2	31112	Male	80.0	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1
3	60182	Female	49.0	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
4	1665	Female	79.0	1	0	Yes	Self-employed	Rural	174.12	24.0	never smoked	1

Figure 1 : Dataset preview

Dataset has 12 columns and 5110 rows of observations. As soon as the data was loaded, missing and duplicated values were checked. It was found that there were 201 rows missing values in bmi. I have imputed the missing values with the median value of the column.

There is one row with gender value as "other" (not Male and Female). I have removed that row from the data set as a cleanup.

Here is the pie chart diagram of categorical variables in the data set.

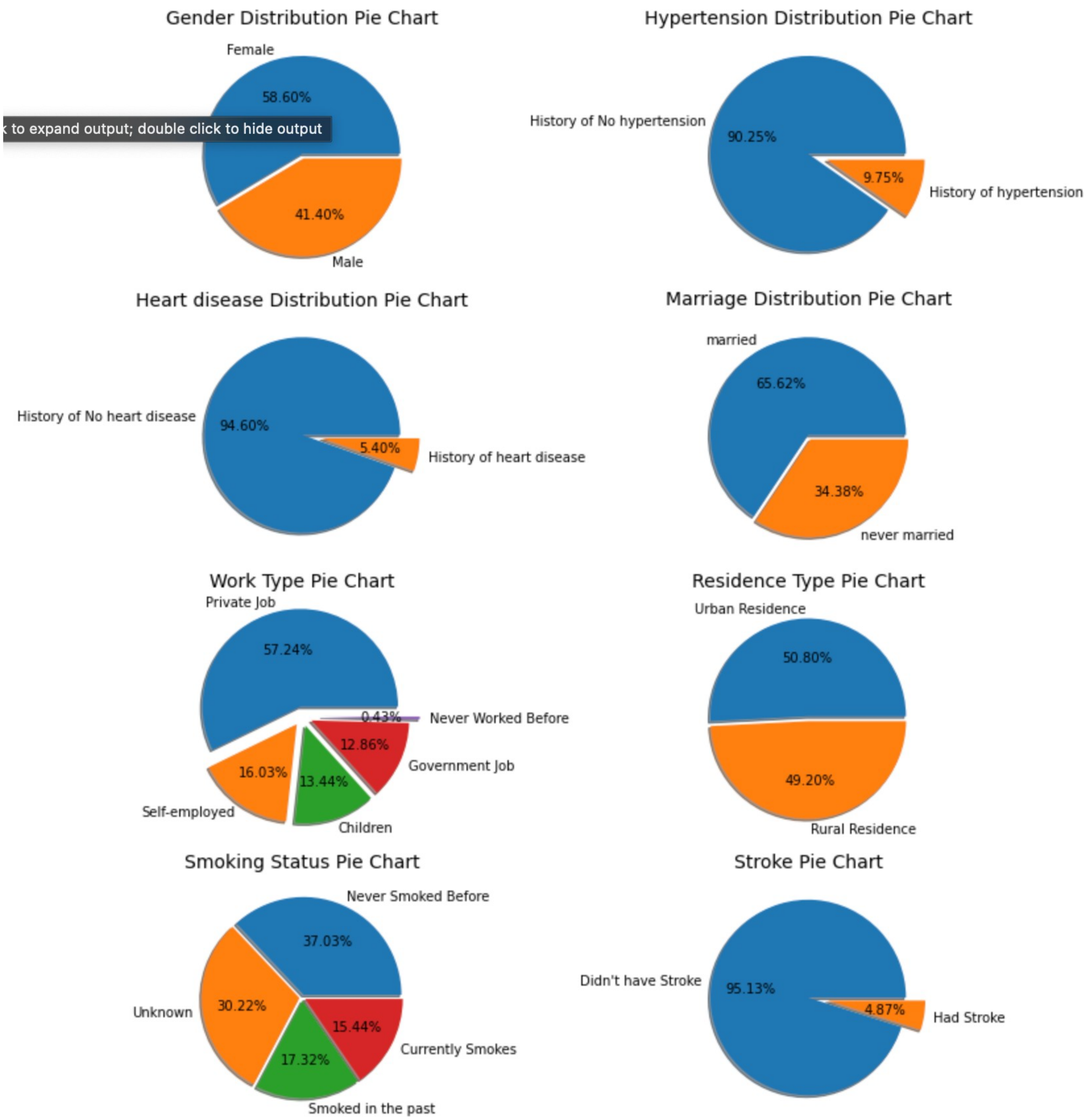

Figure 2 : pie chart plot of Categorical variables

From the above plots, most of the features are well balanced except for the History of heart diseases and the classification of the data around a stroke event or not (my target variable). As part of modeling, I would probably need to rebalance the training dataset by using a rebalancing technique like SMOTE or look for additional data to balance the two classes of target variables of the stroke event. As for the heart rate, I have done a further look around the heart rate data and confirmed some significant association with stroke prediction per below Figure 3.

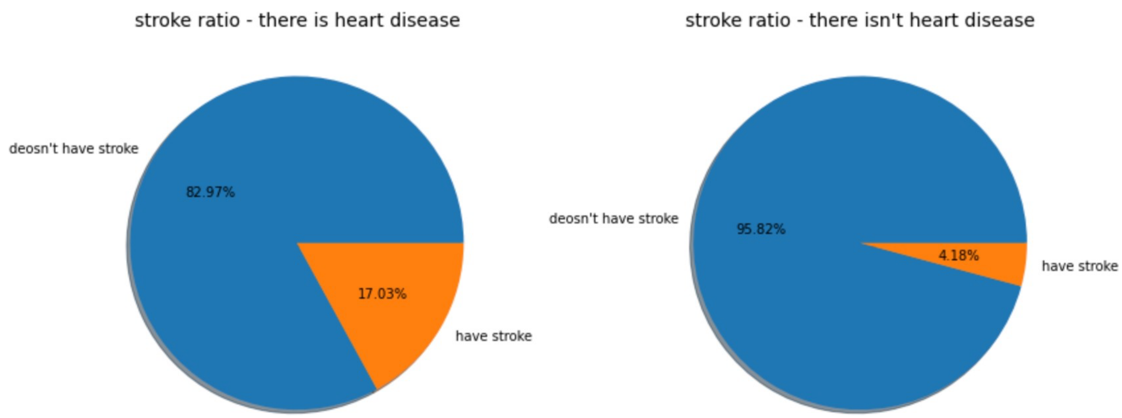

Figure 3: pie chart plot of heart diseases vs. Stroke

Below is a histogram of numerical variables in the feature set.

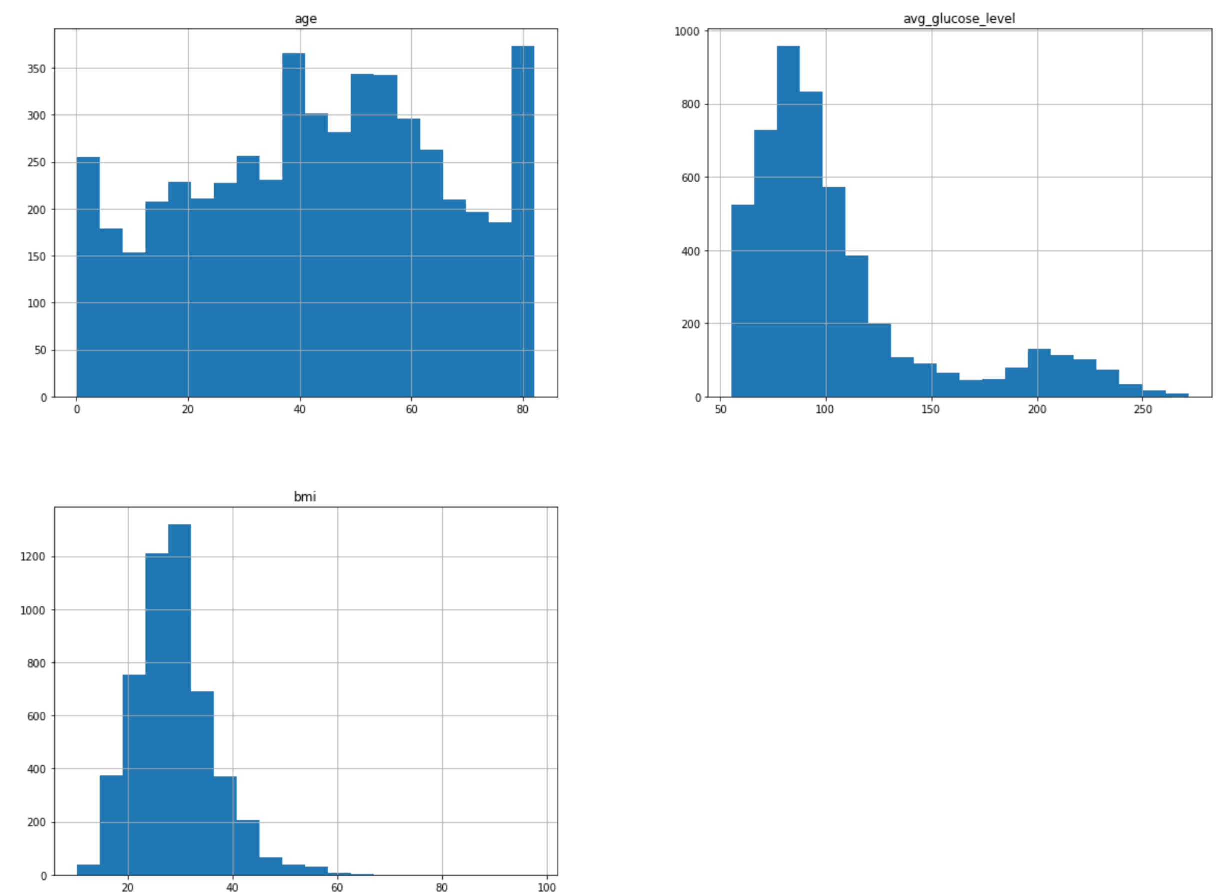

Figure 4: Histograms of Numerical variables

Figure 4 shows that we have a reasonable variation in the age in the input data set. Also, none of the numerical variables have outliers to clean. For example, glucose value and BMI are centered around 90 and 30 weights, respectively. This data behavior is expected as most incoming patients tend to be centered around this mode value as slightly borderline glucose (100 max) and Overweight/obese bmi range of 30.

Here is a heat map correlation of categorical variables against Stroke.

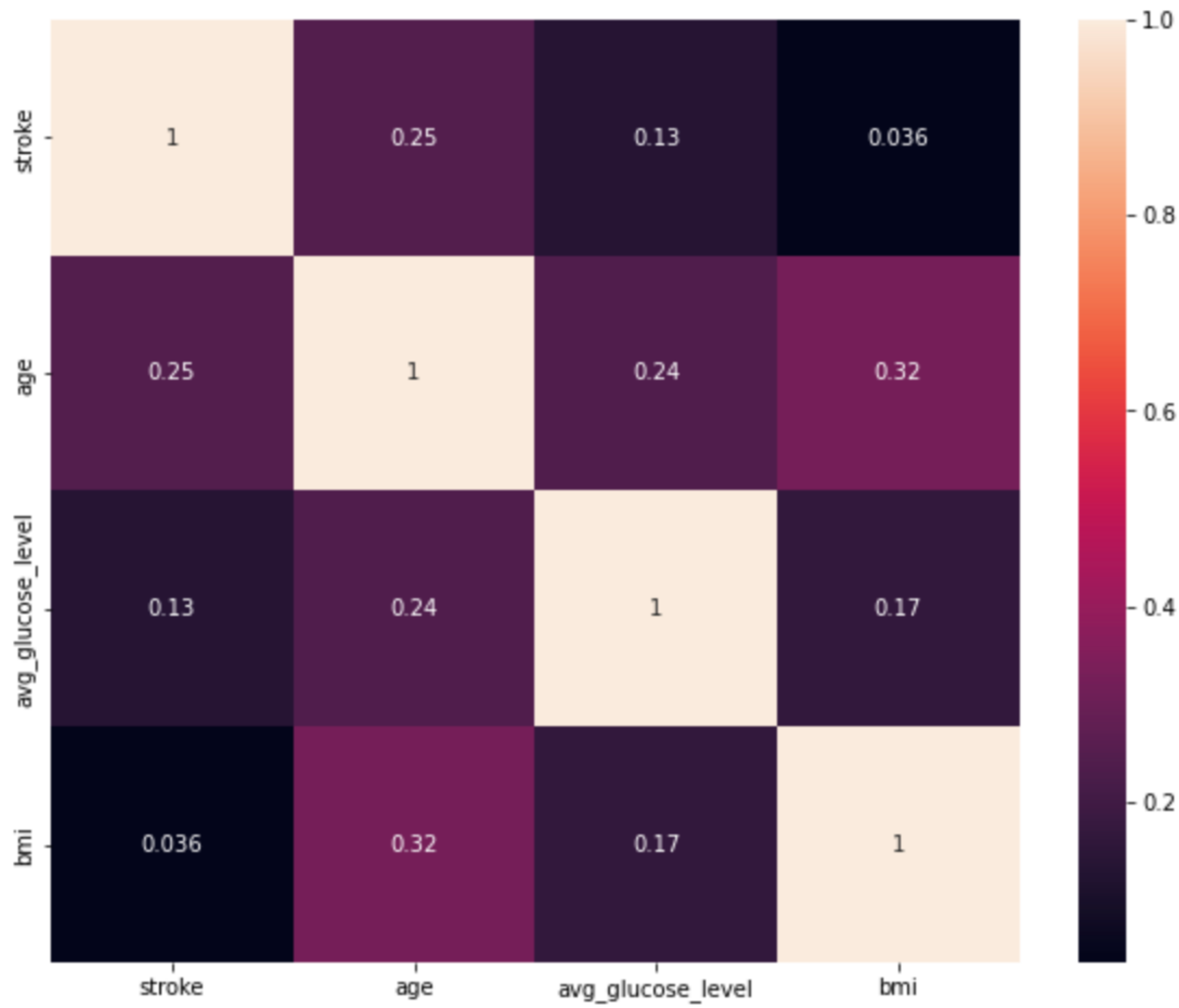

Figure 5: Heat Correlation against numerical features

Figure 5 shows a strong correlation with some of the categorical variables like age against stroke prediction.

Here is Figure 6, outlining the correlation heat map of categorical features against Stroke.

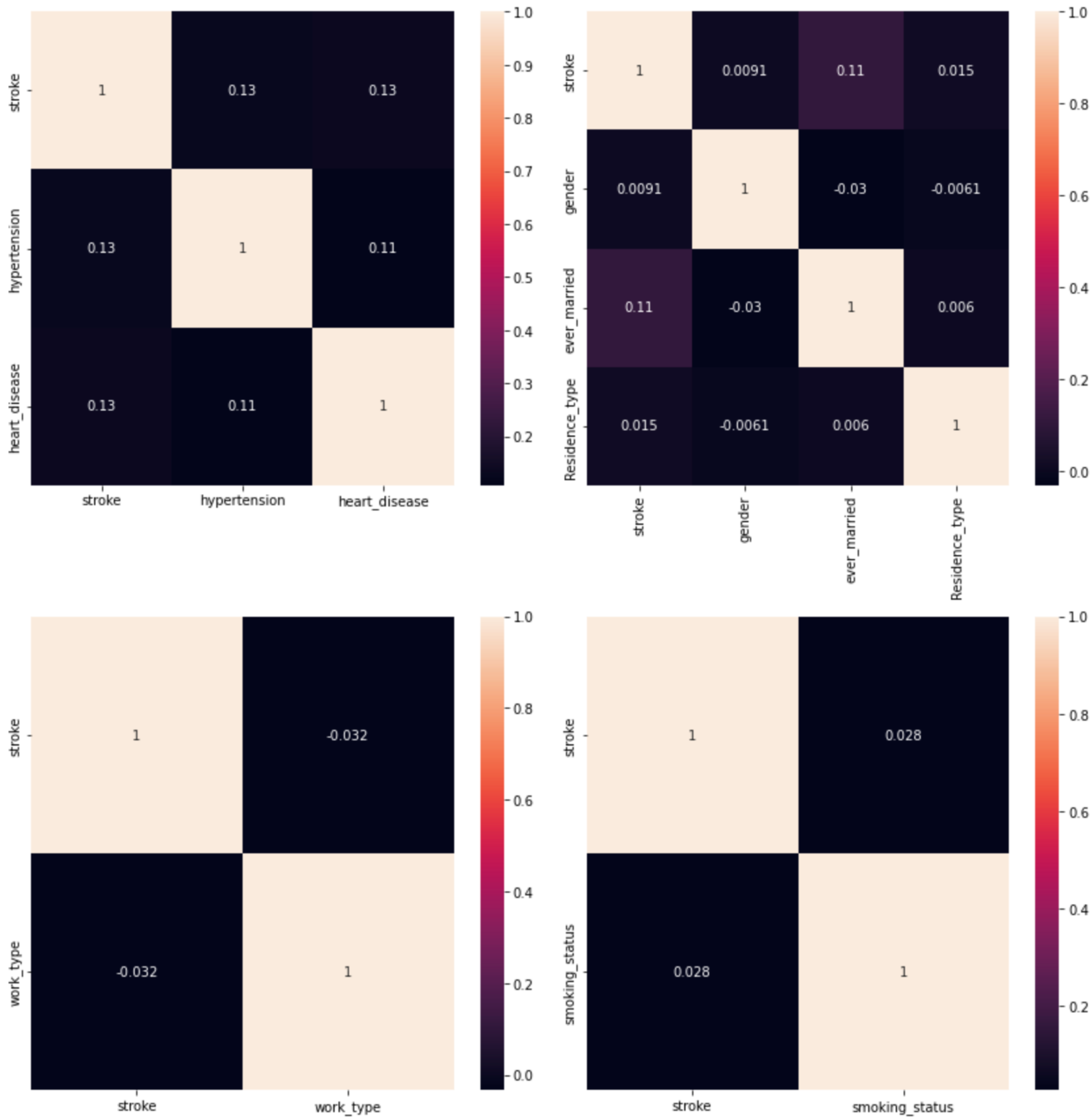

Figure 6: Correlation heatmap for categorical variables

Figure 6 shows a correlation of categorical values against Stroke. Some slightly significant correlations with prior heart disease hypertension if the patient is married. Very first look, it shows as if there is no significant correlation with residence type, work type, or smoking status of the patient. But a lot of anecdotal evidence states that stress in life adds risk factors to Stroke.

Also, based on several medical research outcomes, smoking is one of the severe risk factors for heart diseases and hypertension. As these two factors seem to have some significant correlation with Stroke, I would like to retain smoking status and an input feature to my model. This could indicate an impending cascading event of heart disease or Stroke.

After the exploratory data analysis, the focus shifted to model development. The first model built was using the KNN classifier with a grid search using 1-10 k value search space. The model showed significant accuracy and corresponding scores, with k=8 as best performing. However, a closer look at the confusion matrix proved that the model is not identifying true positives in the target test set. Here is a look at the model outcome.

```
Accuracy: 0.9515
precision: 0.9538
recall: 0.9515
f1 Score: 0.9286
Confusion Matrix for Prediction:

array([[1215,    0],
       [  62,    1]])
```

Figure 7: KNN model performance metrics before Class balancing

This seemed like a result of imbalanced classes in the train data set for my model. I have rebalanced by training data set using SMOTE as a next step. Here is a look at before and after for my train data set,

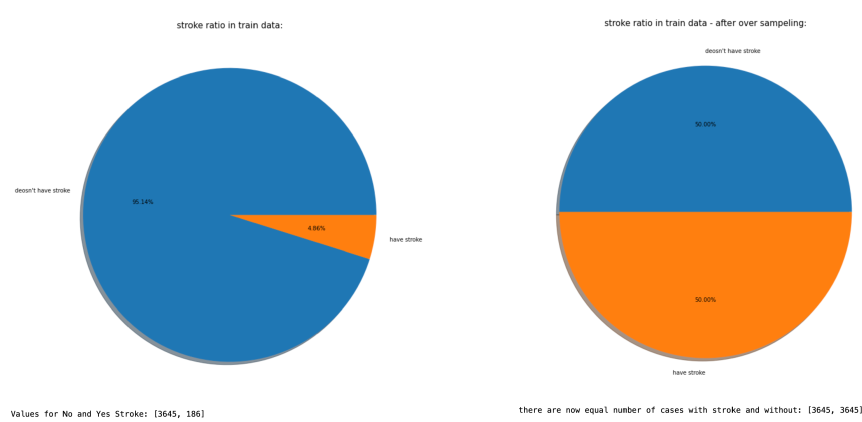

Figure 8: Train data set before vs. After Class balance using SMOTE

After class rebalances, I have trained various models to see the performance of the data set like Logistics regression, SVM, GassianNB, BernoulliNB, KNN classification (without grid search), Decision tree, and Random search. All the models performed with varying degrees of accuracy from 0.56 to 0.89 with different degrees of success in identifying true negatives and positives. Out of all the models, K-nearest neighbor balanced measurement scores and identified true positives and negatives.

After that, the K-nearest neighbor classifier was employed again with a grid search of 1-10. This time, the model performed better by pushing accuracy from 0.81 to 0.89 and keeping true positive and negative identification compared to my original trained KNN model before class rebalancing.

```
Accuracy: 0.8889
precision: 0.9118
recall: 0.8889
f1 Score: 0.8999
Confusion Matrix for Prediction:

array([[1127,   88],
       [  54,    9]])
```

Figure 9: KNN model performance metrics after Class balancing

Data understanding

This project aims to create a predictive model and analyze the data of people having a stroke. It is essential to note this dataset itself is not very large and has an Imbalance in target class distribution. But it was something of my interest and the best data I could find out there for free. For a person to develop a disease, hundreds of factors play the role. By no means could I create a model that can predict Stroke based on a few parameters. However, it provided some insights into Stroke disease itself and the process of developing a predictive model.

I should be thankful that the data I acquired was clean with no missing or duplicate values, which saved me a lot of time and effort. I have used mainly the matplotlib and seaborn as the visual tools. They've produced some good visuals in the notebook.

I chose pie charts over other plots for categorical visuals because it's easier to see the distribution and symmetry at a glance. For the numerical variables, though, I generated a set of histograms for each variable.

Data preprocessing

Dealing with missing entries in the bmi column was achieved by imputing the data with a median value, as it's a numerical attribute to avoid skewing the model training. However, my significant effort was to rebalance the imbalanced classes in the train data set. I achieved this by oversampling underrepresented classes by using SMOTE package.

Feature engineering

During the feature engineering, categorical features were encoded using the one-hot encoding method of Pandas library. This enabled the conversion of all the categorical variables into numbers. After that, all the numbers were scaled using standard scalar to transform the data into a scaled attribute.

Model development

Model development was the most critical and insightful part of the project as I had to go through a bunch of classifiers and experiment with them. Unfortunately, some of the classifiers performed so poorly that I did not include that in my notebook.

Even though I knew the data was imbalanced in the first attempt, I have tried a KNN classifier training using grid search hyperparameter tuning for n_neighbors to see how the model would perform. This resulted in a higher accuracy of 95% but looking at the confusion matrix of results showed that the model failed to identify any true positive outcomes. This reinforced my learning of the importance of data and review of train data fed into a model.

I had them rebalance the train data set using SMOTE to oversample underrepresented true positive classes. I used this technique as I could not find a free data set that could augment and rebalance my dataset.

Once the data set was rebalanced, I have trained various models to see the performance of the data set like Logistics regression, SVM, GassianNB, BernoulliNB, KNN classification (without grid search), Decision tree, and Random search. These models have been performed with varying accuracy and true positive and negative identification. Out of all these models, KNN was the most well-balanced.

To boost the model further, I have trained a KNN algorithm with hyperparameter tuning using n_neighbors, which resulted in better results of close to 90% accuracy and better true positive and negative identification with precision/recall and f1 scores all-around 90% mark.

Conclusion

Best predictive model found out to be a KNN classifier with hyperparameter tuning, which has the highest classification accuracy of 90%. I must admit that predicting a stroke is a very sensitive undertaking for anyone. I am confident that this model is the best model for the given limited features as in this study. But it is not the best for predicting a stroke in general because the model doesn't consider other important risk factors like exercise, alcohol consumption, nutrition, etc. Although this project was academic, it gave me practical perspectives on data mining and predictive modeling. Some of the challenging parts of the project were preprocessing the dataset, dealing with an imbalance in the dataset, fine-tuning the various classification algorithms.

Recommendation

This model can predict stroke disease based on 11 features included in this study with 90 percent accuracy. Still, it can be improved further if more data and features have all other important risk factors of Stroke and a more balanced dataset with higher volume for train data. Therefore, I would recommend any stakeholder to use this model with caution.

Risks:

Following the potential risks involved in this project:

- This project assumes that the dataset sourced is valid and accurate.
- There is always a chance of data loss due to incompleteness or extremities/anomalies.
- The number of observations in the dataset used is low (5110) for this study.
- This dataset does not include all the risk factors of a stroke like diabetes, alcohol consumption, physical activity etc., hence the model cannot be fully accurate.

References:

1. Centers for Disease Control and Prevention. Underlying Cause of Death, 1999–2018. CDC WONDER Online Database. Centers for Disease Control and Prevention; 2018. Accessed March 12, 2020.
2. Tsao CW, Aday AW, Almarzooq ZI, Alonso A, Beaton AZ, Bittencourt MS, et al. Heart Disease and Stroke Statistics—2022 Update: A Report From the American Heart Associationexternal icon. *Circulation*. 2022;145(8):e153–e639.
3. Jackson G, Chari K. National Hospital Care Survey Demonstration Projects: Stroke Inpatient Hospitalizationsexternal icon. *Natl Health Stat Report.* 2019 Nov;(132):1-11.
4. Fang J, Keenan NL, Ayala C, Dai S, Merritt R, Denny CH. Awareness of stroke warning symptoms—13 states and the District of Columbia, 2005. *MMWR.* 2008;57(18):481–5.
5. https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset?resource=download