MSSP 608: Practical Machine Learning Project Report –Stroke Prediction
Group Member: Yanxi Zeng, Xinyuan Hu

Google Colab Notebook:

https://colab.research.google.com/drive/1zDGUliY_vuKWDiIQ2n5nSVRmNMPyh8bg?usp=sharing

Part 1: Introduction

The question to be considered

This project is expected to predict whether a patient is likely to get a stroke based on their basic information and health conditions like gender, age, diseases history, smoking status and so forth via decision tree model.

Background / the meanings of this project

According to the World Health Organization (2020), stroke is the second leading cause of death globally, responsible for approximately 11% of total deaths. And it has been estimated that with early intervention, half of all strokes could be prevented by controlling modifiable risk factors in such individuals (Brainin et al., 2018). Therefore, it is vital to explore the risk factors of stroke and identify adults at high risk of stroke for primary prevention. Machine learning is a valid way to achieve these goals, which is also more convenient and cost-effective than traditional methods (Liu et al., 2019).
In recent years, numerous machine learning and data analysis models have been applied to assess stroke risk factors and outcomes. They include evaluating a mixed-effect linear model to predict the risk of cognitive decline poststroke (Hbid et al., 2021) and developing a deep neural network (DNN) model, applying logistic regression and random forest to predict poststroke motor outcomes(Kim et al., 2021).
This project using machine learning tools to predict whether an individual has a high risk of stroke based on their possible stroke-related information therefore is also worthy of pursuit.

Part 2: Primary task

Task Description

The primary task aims to train a decision tree model to predict whether a patient is likely to get a stroke using a stroke prediction dataset from Kaggle.

Data

The dataset to be used in this project is the stroke prediction dataset from Kaggle (https://www.kaggle.com/fedesoriano/stroke-prediction-dataset). This dataset includes 5110 observations with 12 attributes, which can be divided into 11 features related to the basic information and health conditions of patients, and one feature showing whether the patient had a stroke.
The detailed description of variables is as follows:

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

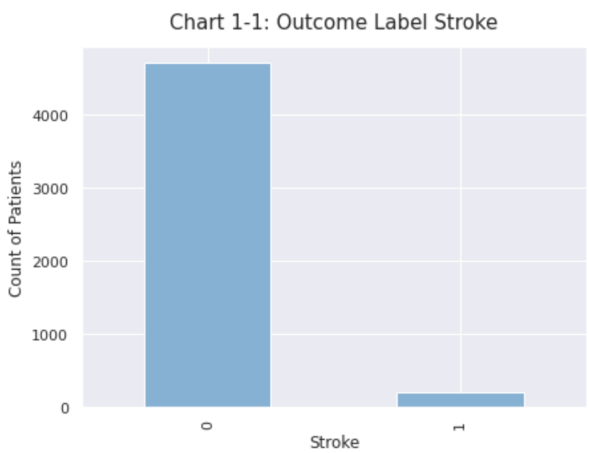
Outcome Variable

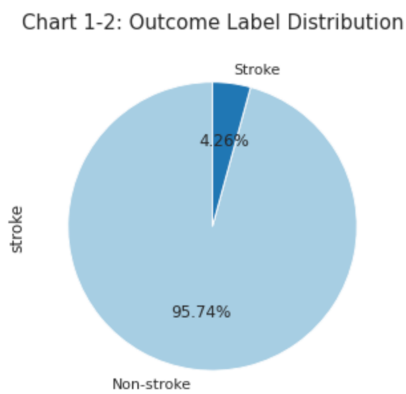
The binary label stroke includes 1 which indicates the patient had a stroke and 0 when the patient did not.

Table 1

Summary of Outcome Label

Variable	Label	Format	Obs	Percentage
stroke	0	int64	4700	0.96
	1	int64	209	0.04





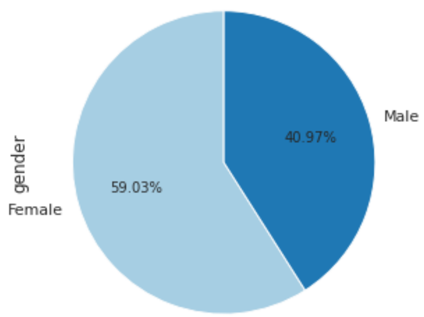
Features: Exploratory Data Analysis

There are

- **Gender**

Almost 59.03% samples in this dataset are female.

Chart 2-1: Distribution of Gender



- **Age**

The age distribution is shown in Chart 3-1. And according to this box plot, it seems that the elderly are more likely to get stroked.

Chart 3-1: Age Distribution of Different Patients

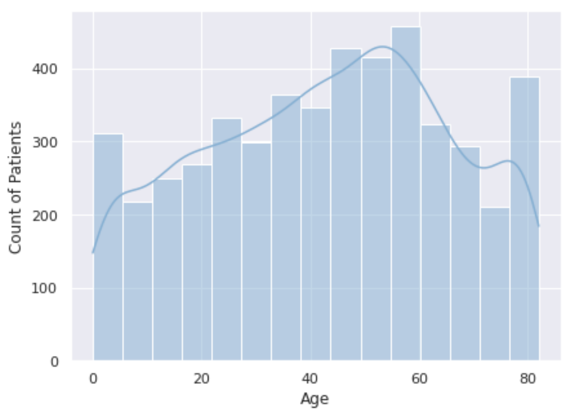
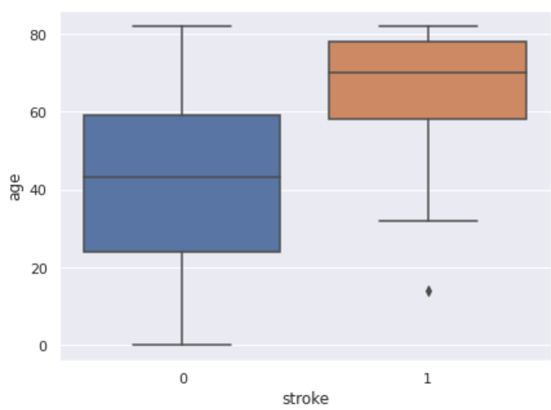


Chart 3-2: Box-plot of Age



- **Martial status & smoking status**

Almost 65% samples in this dataset is married; and most people has never smoked.

Chart 4: Distribution of Marital Status

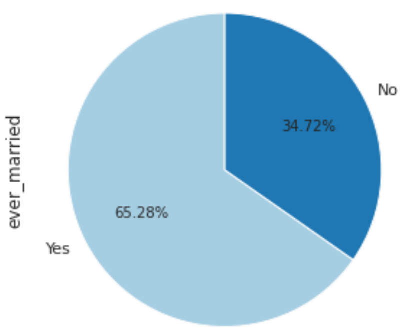
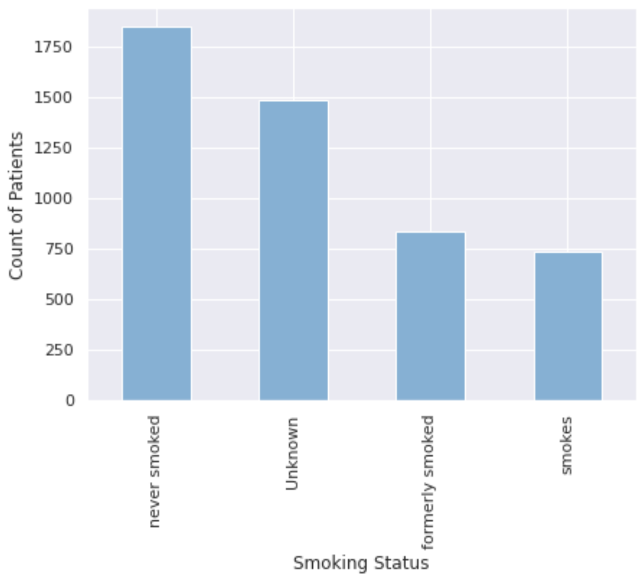
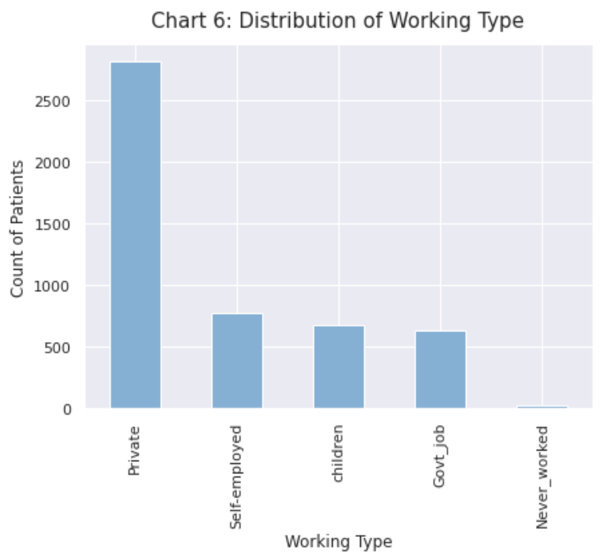


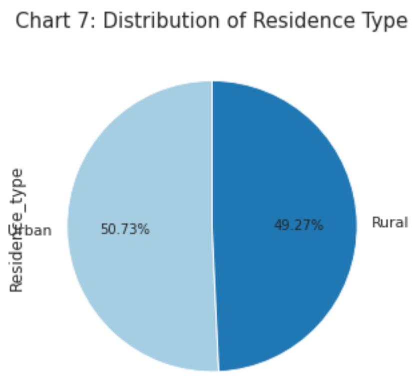
Chart 5: Distribution of Smoking Status



- **The types of working and types of residence**

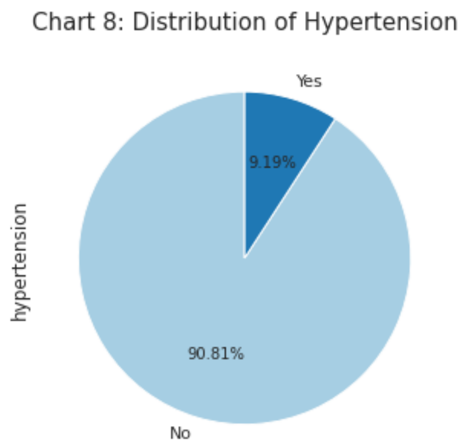
Most samples in the dataset work in a private company; and half of the sample lives in rural areas and half lives in urban areas.

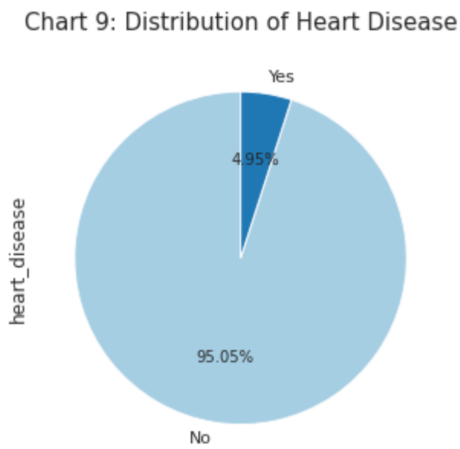




- **Hypertension and Heart disease**

The majority of samples in this dataset do not have hypertension or heart disease.





- **The average Glucose level & Glucose level in blood**

For most people in the dataset, their average Glucose level distributed in the range of 50 to 100.

Chart 10-1: Distribution of Average Glucose Level

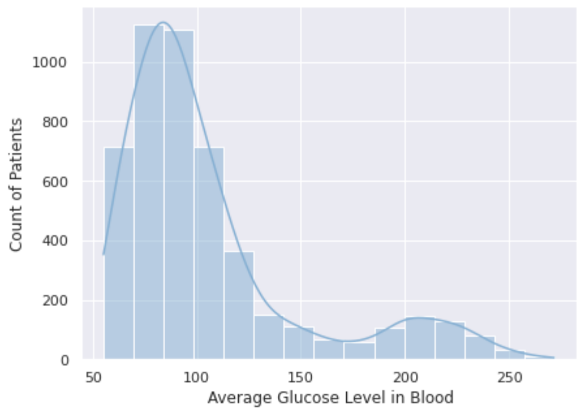
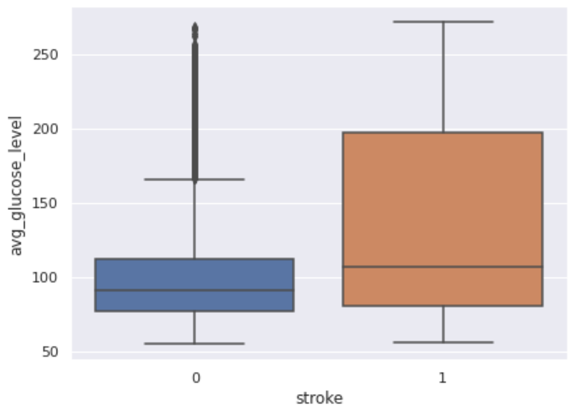


Chart 10-2: Box-plot of Glucose Level in Blood



- **BMI**

Chart 11-1: BMI Distribution of Different Patients

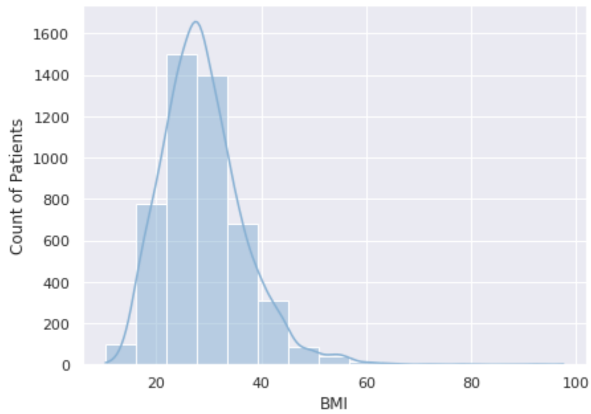
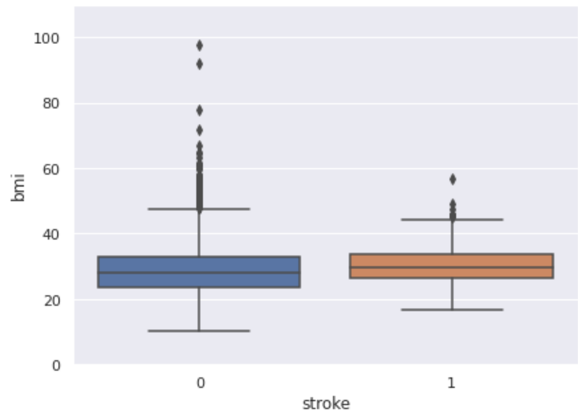


Chart 11-2: Box-plot of BMI



Summary of Features

Variable	Format	Obs	Mean	Std	Min	Max
gender	object	4909	——	——	——	——
age	float64	4909	42.87	22.56	0.08	82.00
hypertension	int64	4909	0.09	0.29	0.00	1.00
heart_disease	int64	4909	0.05	0.22	0.00	1.00
ever_married	object	4909	——	——	——	——
work_type	object	4909	——	——	——	——
Residence_type	object	4909	——	——	——	——
avg_glucose_level	float64	4909	105.30	44.43	55.12	271.74
bmi	float64	4909	28.90	7.85	10.30	97.60
smoking_status	object	4909	——	——	——	——

Experimental Setup
The original model will train a decision tree classifier and it will split features into training and testing sets as 4:1.

Model

- **Metrics**

The model performance of individual model will be evaluated by several criteria: the overall accuracy of the model; the precision, recall and f1-score of both label '0' and '1'; the overall macro avg and weighted avg of the precision, recall and f1-score. And whether this task is successfully completed will be determined by the model performance.

- **The process of building the best-performed model**

The process of building the best-performed model includes:

- **Features adjustment**
  For column 'gender''Residence_type' and 'ever_married', their original two values, which are expressed in the form of string, are changed to '0' and '1'. And 9 new dummy features, which are the numeric forms, are created to replace 'smoking_status' and 'work_type' column. Especially, column 'work_type_Govt_job', 'work_type_Never_worked', 'work_type_Private', 'work_type_Self-employed' and 'work_type_children' replaced the 'work_type' column; column 'smoking_status_Unknown', 'smoking_status_formerly smoked', 'smoking_status_never smoked' and 'smoking_status_smokes' replaced the 'smoking_status' column.
- **Balance the dataset / Unsampling**
  Since the counts of label '1' and the counts of label '0' are unbalanced in this original-size dataset (the results of this problem is that the precision, recall, f1-score for label'1' and '0' are very different in any model using the original-size dataset), we used the SMOTE Algorithm to balance the dataset.
- **Classification comparison between decision tree model and logistic regression model**
  We created a decision tree model and a logistic regression model to compare their model performance.
- **Hyperparameter tuning for both the decision tree model and logistic regression model**
  We then created a decision tree model after hyperparameter tuning and a logistic regression model after hyperparameter tuning to compare their model performance with that of two original models.
- **Decide the final model**
  Lastly, we decided the final model with the best model performance.

Results
The performance comparison of these four models are as follows:

The accuracy of four models

Model	Accuracy
The logistic regression model before hyperparameter tuning	87.32612
The logistic regression model after hyperparameter tuning	87.24884

The decision tree model before hyperparameter tuning	81.06646
The decision tree model after hyperparameter tuning	82.38022

- **The precision, recall, f1-score and support of label '0' and '1' in the logistic regression model before hyperparameter tuning:**

```
          precision    recall  f1-score   support

       0       0.89      0.85      0.87       647
       1       0.86      0.89      0.88       647
```

- **The precision, recall, f1-score and support of label '0' and '1' in the logistic regression model after hyperparameter tuning:**

```
          precision    recall  f1-score   support

       0       0.89      0.85      0.87       647
       1       0.86      0.90      0.88       647
```

- **The precision, recall, f1-score and support of label '0' and '1' in the decision tree model before hyperparameter tuning:**

```
          precision    recall  f1-score   support

       0       0.76      0.90      0.83       647
       1       0.88      0.72      0.79       647
```

- **The precision, recall, f1-score and support of label '0' and '1' in the decision tree model after hyperparameter tuning:**

```
          precision    recall  f1-score   support

       0       0.77      0.93      0.84       647
       1       0.91      0.72      0.80       647
```

After comparison, we selected the logistic regression model before hyperparameter tuning as the best-performed model. On the one hand, the precision, recall, f1-score and support of label '0' and '1' in this model are relatively high and balanced in these four models; on the other hand, its accuracy is the highest (87.32612) among these four models.

Errors

With 87.33% accuracy, there are in total 164 cases in the test set that the model assigned the wrong prediction about stroke. Certain pattern can be seen in the error cases. The descriptive statistics of the following features in the error set are different from those in the correct set.

Table 3-1

Summary of Correct Set

	id	gender	age	hypertension	heart_disease	ever_married	Residence_type	avg_glucose_level	bmi
count	1130.000000	1130.000000	1130.000000	1130.000000	1130.000000	1130.000000	1130.000000	1130.000000	1130.000000
mean	37540.316814	0.268142	58.162229	0.098230	0.099115	0.758407	0.383186	124.530172	30.524927
std	22800.840338	0.443188	18.925013	0.297757	0.298949	0.428238	0.486378	53.923015	6.075288
min	129.000000	0.000000	10.000000	0.000000	0.000000	0.000000	0.000000	55.470000	15.300000
25%	17303.500000	0.000000	46.000000	0.000000	0.000000	1.000000	0.000000	82.838281	26.647675
50%	35352.500000	0.000000	62.242259	0.000000	0.000000	1.000000	0.000000	102.842070	29.680296
75%	58827.000000	1.000000	74.337319	0.000000	0.000000	1.000000	1.000000	163.925425	33.560165
max	72861.000000	1.000000	82.000000	1.000000	1.000000	1.000000	1.000000	254.630000	78.000000

Table 3-2

Summary of Error Set

	id	gender	age	hypertension	heart_disease	ever_married	Residence_type	avg_glucose_level	bmi
count	164.000000	164.000000	164.000000	164.000000	164.000000	164.000000	164.000000	164.000000	164.000000
mean	35032.804878	0.207317	65.835828	0.152439	0.207317	0.847561	0.432927	126.371483	30.245343
std	21420.111750	0.406626	10.457707	0.360547	0.406626	0.360547	0.496998	57.031417	6.188026
min	768.000000	0.000000	42.000000	0.000000	0.000000	0.000000	0.000000	60.980000	17.600000
25%	19632.000000	0.000000	56.967221	0.000000	0.000000	1.000000	0.000000	81.440000	26.416738
50%	32172.000000	0.000000	65.994972	0.000000	0.000000	1.000000	0.000000	101.915000	29.578322
75%	51591.750000	0.000000	75.975672	0.000000	0.000000	1.000000	1.000000	160.926540	34.345425
max	72081.000000	1.000000	82.000000	1.000000	1.000000	1.000000	1.000000	255.170000	56.000000

Part 3: Extension task

Task#1 Description

The extension task#1 aims to conduct a K Means clustering analysis (unsupervised learning). This task would involve determining the optimal number of clusters and a quantitative analysis of the clusters that are produced.

We selected the value of k at the “elbow”, for example, the point after which the distortion/inertia start decreasing in a linear fashion. Thus, for the given data, we concluded that the optimal number of clusters for the data is 2. And the silhouette score for k=2 is also the highest.

We also checked the silhouette plots for k=2. The silhouette plots show that the value of 2 is a good pick as all the clusters have silhouette scores above the average.

Chart 12: Elbow Analysis for Optimal k

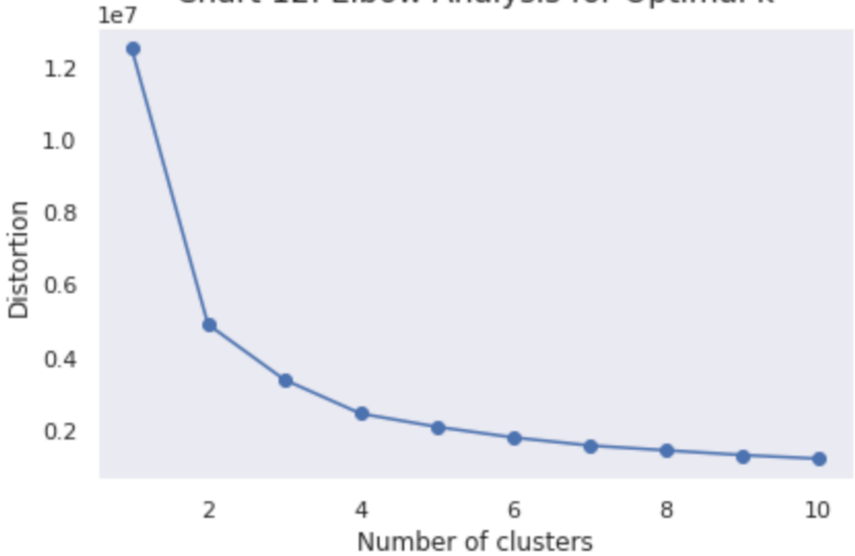


Chart 13: Silhouette Analysis for Optimal k

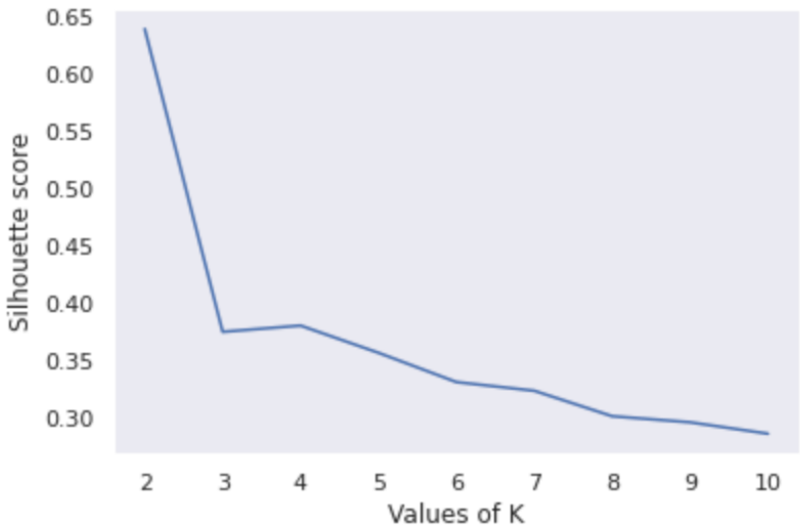
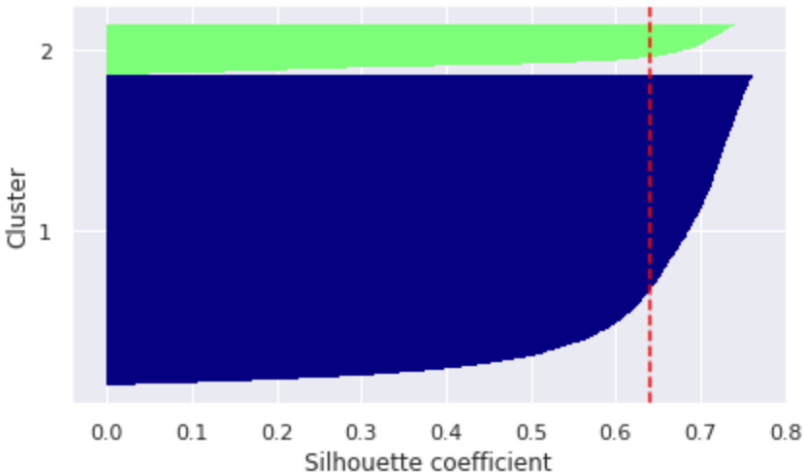


Chart 14: Silhouette Coefficient Plot (k=2)



From the tables below, we can find that the patients in the cluster#2 are elder, are more likely to have hypertension and heart disease, have higher average glucose level and bmi level. And they are more susceptible to stroke.

Table 4-1

Summary of Cluster #1

	id	Cluster	age	hypertension	heart_disease	avg_glucose_level	bmi	stroke
count	4212.000000	4212.0	4212.000000	4212.000000	4212.000000	4212.000000	4212.000000	4212.000000
mean	36616.347341	0.0	40.423818	0.068139	0.035138	89.492557	28.301353	0.031102
std	21028.450708	0.0	22.269671	0.252014	0.184150	19.976114	7.671979	0.173613
min	77.000000	0.0	0.080000	0.000000	0.000000	55.120000	10.300000	0.000000
25%	18065.250000	0.0	22.000000	0.000000	0.000000	74.602500	23.100000	0.000000
50%	36839.000000	0.0	41.000000	0.000000	0.000000	87.150000	27.600000	0.000000
75%	54797.500000	0.0	57.000000	0.000000	0.000000	102.272500	32.400000	0.000000
max	72940.000000	0.0	82.000000	1.000000	1.000000	150.030000	97.600000	1.000000

Table 4-2

Summary of Cluster #2

	id	Cluster	age	hypertension	heart_disease	avg_glucose_level	bmi	stroke
count	696.000000	696.0	696.000000	696.000000	696.000000	696.000000	696.000000	696.000000
mean	39747.850575	1.0	57.665230	0.235632	0.136494	200.943966	32.484483	0.112069
std	20606.264026	0.0	18.216318	0.424698	0.343560	29.010041	7.993363	0.315678
min	239.000000	1.0	0.720000	0.000000	0.000000	142.630000	12.800000	0.000000
25%	22352.250000	1.0	49.000000	0.000000	0.000000	180.790000	27.000000	0.000000
50%	42551.500000	1.0	60.000000	0.000000	0.000000	203.960000	31.400000	0.000000
75%	56385.000000	1.0	71.000000	0.000000	0.000000	221.807500	36.825000	0.000000
max	72915.000000	1.0	82.000000	1.000000	1.000000	271.740000	71.900000	1.000000

Task#2 Description

The extension task#2 aims to conduct a fairness audit of this dataset, especially on three features related to the patients' basic information: gender, age, and residence types.

Gender. Strokes affect differently on gender in several ways. The U.S. Department of Health & Human Services points out women are more likely to have recurrence than men within 5 years of the first stroke and some stroke risk factors are more common in women. Women usually have more events and are less likely to recover while age-specific stroke rates are higher in men.

Age. According to the Stanford Health Care, most strokes occur in people who are 65 or older. However, in the US, 10% of people experience a stroke younger than 45, and that number is rising.

Residence Types. Some studies have concluded that risk factors were more prevalent but less likely to be controlled in rural than in urban residents without prior stroke, whereas in those with prior stroke, risk factor prevalence and treatment were similar. (Kapral et al., 2019; Kamin et al., 2021)

From the above statistics, we recognize that historical biases may have affected the quality of our data. It could be a reasonable concern when misclassification on the younger people or other underrepresented communities, who will be most impacted by making wrong results through automation. Their risks of getting a stroke could be ignored, leading to missing the best time for prevention and treatment. If we were building this system in a professional setting, we would contact and work with the data provider which could be the hospitals with records of stroke patients and high-risk individuals. Getting more balanced data would be an effective solution to promote fairness and give more attention to underrepresented groups.

Methods

For Age, we will transform it to a binary feature, under_45 (below 45 years old) and 45_above (45 years old and above). The threshold 45 is taken from the statistics by Stanford Health Care.

The fairness is measured by the following metrics:

- **Demographic Parity.** Calculate how many members of each demographic subgroup get assigned into each label from the classifier. If the distribution of labels is not balanced among the groups, the model is unfair.
- **Prevalence.** Calculate the percentage of each subgroup in the dataset, then see whether the distribution changes at each class label. Fairness is expected to see an even distribution.

Results

- **Demographic Parity.** According to the charts below, we can see that the distribution of labels is balanced among the residence-type groups. Specifically, urban and rural patients have similar proportions of suffering from a stroke. However, older patients with 45 years old and above and those with less than 45 years old have quite different proportions. The same is for the male and female patients.

Chart 15-1. Men's Stroke Distribution

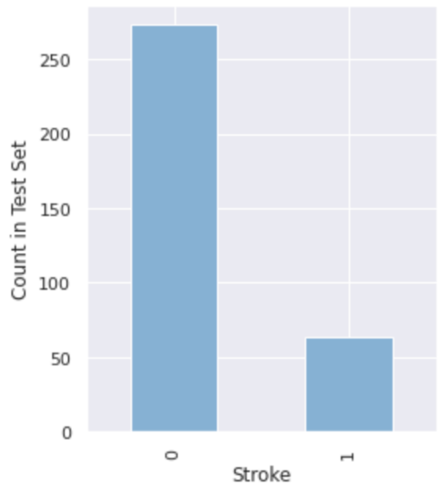
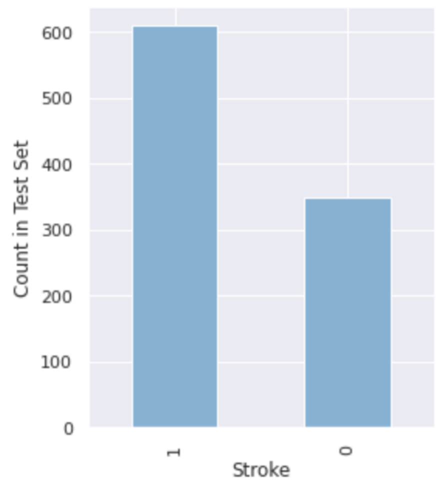
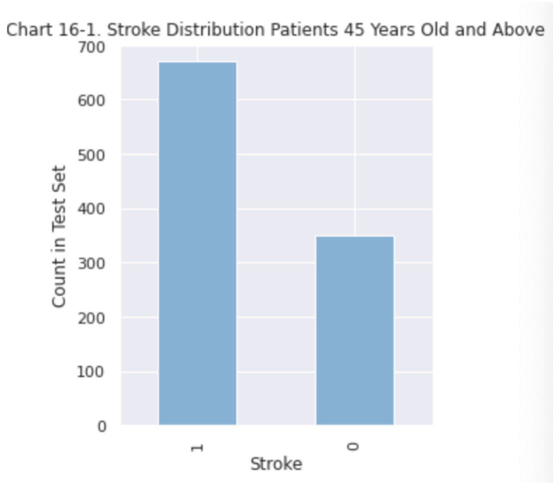
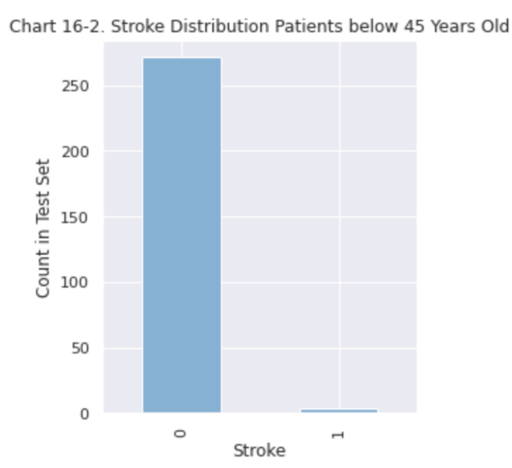
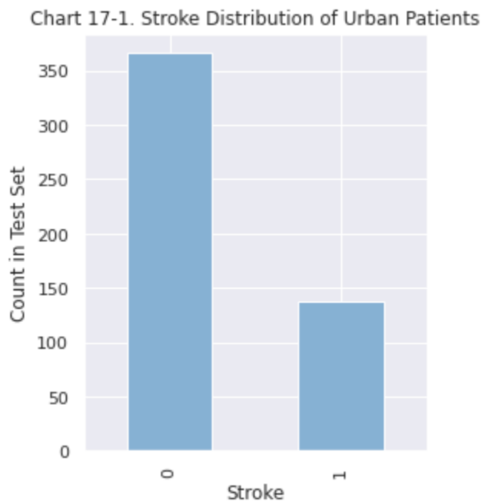


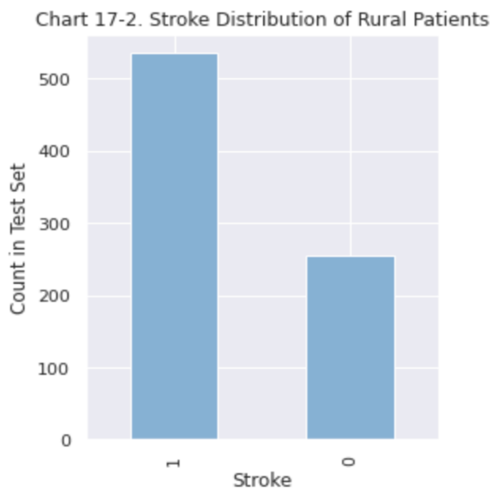
Chart 15-2. Women's Stroke Distribution











- **Prevalence.** As shown in the charts below, the stroke distribution among gender and age groups for prevalence are similar to the demographic parity. This time the stroke distribution of the residence-type groups is unbalanced either.

Chart 18-1. Baseline Distribution Divided by Gender - Count

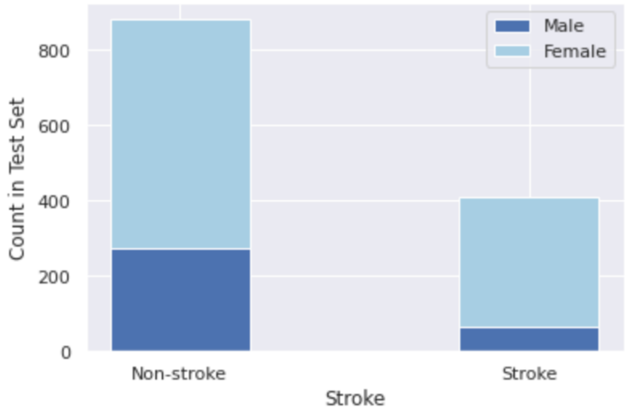


Chart 18-2. Baseline Distribution Divided by Gender - Percentage

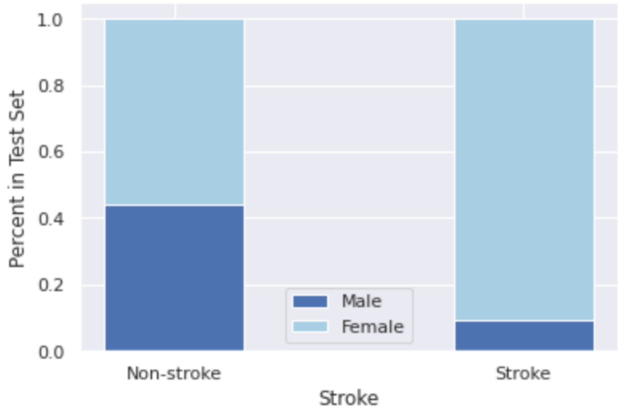


Chart 19-1. Baseline Distribution Divided by Age - Count

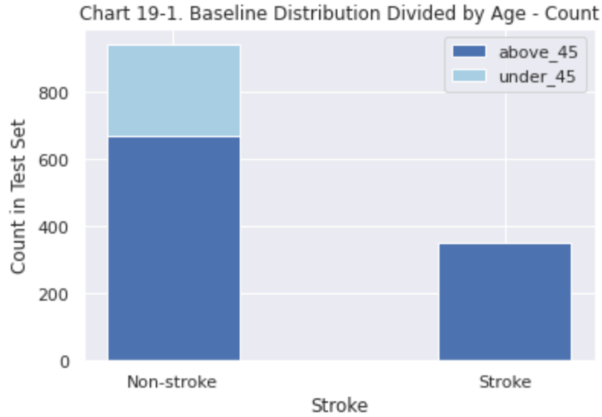


Chart 19-2. Baseline Distribution Divided by Age - Percentage

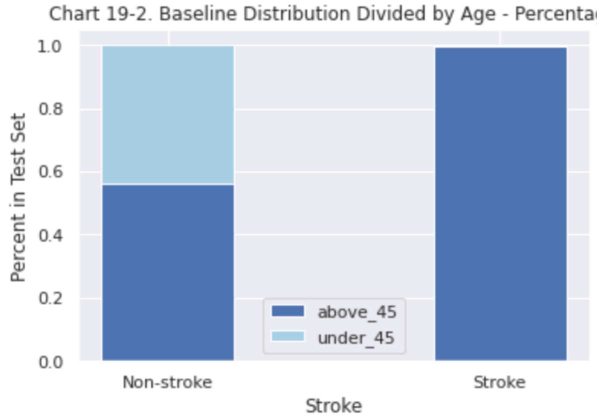


Chart 20-1. Baseline Distribution Divided by Residence Type - Count

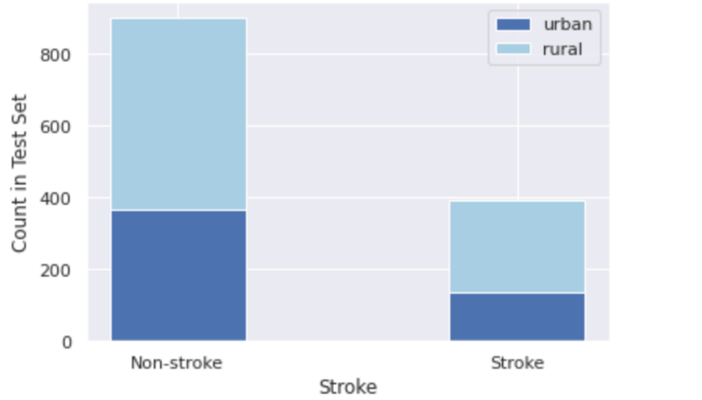
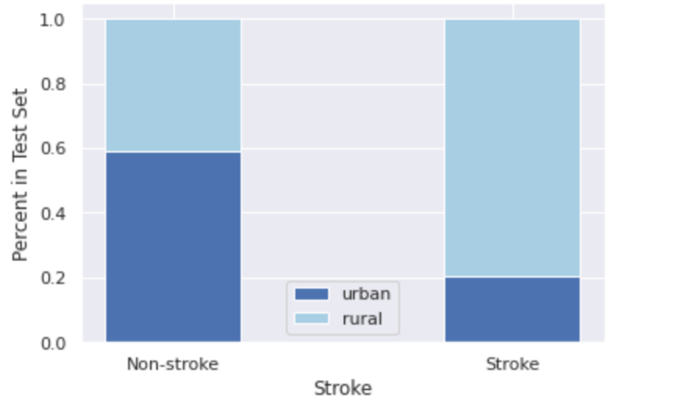


Chart 20-2. Baseline Distribution Divided by Residence Type - Percentage



Reference

Brainin, M., Feigin, V., Martins, S., Matz, K., Roy, J., Sandercock, P., ... & Wiseman, A. (2018). Cut stroke in half: polypill for primary prevention in stroke. *International Journal of Stroke*, 13(6), 633-647.

Hbid, Y., Fahey, M., Wolfe, C., Obaid, M., & Douiri, A. (2021). Risk Prediction of Cognitive Decline after Stroke. *Journal of stroke and cerebrovascular diseases: the official journal of National Stroke Association*, 30(8), 105849.

Kapral, M. K., Austin, P. C., Jeyakumar, G., Hall, R., Chu, A., Khan, A. M., Jin, A. Y., Martin, C., Manuel, D., Silver, F. L., Swartz, R. H., & Tu, J. V. (2019). Rural-Urban Differences in Stroke Risk Factors, Incidence, and Mortality in People With and Without Prior Stroke. *Circulation. Cardiovascular quality and outcomes*, 12(2), e004973.

Kamin Mukaz, D., Dawson, E., Howard, V. J., Cushman, M., Higginbotham, J. C., Judd, S. E., Kissela, B. M., Safford, M. M., Soliman, E. Z., & Howard, G. (2021). Rural/urban differences in the prevalence of stroke risk factors: A cross-sectional analysis from the REGARDS study. *The Journal of rural health : official journal of the American Rural Health Association and the National Rural Health Care Association*, 10.1111/jrh.12608.

Kim, J. K., Choo, Y. J., & Chang, M. C. (2021). Prediction of Motor Function in Stroke Patients Using Machine Learning Algorithm: Development of Practical Models. *Journal of stroke and cerebrovascular diseases: the official journal of National Stroke Association*, 30(8), 105856.

Liu, T., Fan, W., & Wu, C. (2019). A hybrid machine learning approach to cerebral stroke prediction based on imbalanced medical dataset. *Artificial intelligence in medicine*, 101, 101723.

Reeves, M. J., Bushnell, C. D., Howard, G., Gargano, J. W., Duncan, P. W., Lynch, G., Khatiwoda, A., & Lisabeth, L. (2008). Sex differences in stroke: epidemiology, clinical presentation, medical care, and outcomes. *The Lancet. Neurology,* 7(10), 915–926. Derived from https://doi.org/10.1016/S1474-4422(08)70193-5

World Health Organization (2020). The top 10 causes of death. Detrived from https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death

Stanford Health Care (2013). Stroke in Young People. Derived from file://Users/huxinyuan/Downloads/stroke-young-patients-qa.pdf