TechEagle Team Project Proposal

Nigel Wang, Faustin Nzitonda, Susie Li, Jennifer Kim

1. Business Problem

For our business scenario, we will be working with the stroke dataset to estimate whether a person is likely to get a stroke or not. We will look at different variables that pertain to patients' medical and personal historical data such as gender, age, hypertension, heart disease, marital status, work type, residence, average glucose level, body mass index, and smoking status to determine the likelihood of stroke for each person.

Based on Centers for Disease Control and Prevention, more than 795,000 people in the United States have a stroke every year, while 1 in 6 deaths from cardiovascular disease was due to stroke. In addition, the total stroke-related costs between 2017 and 2018 resulted in approximately $53 billion, which includes cost of healthcare services, medicines, and missed days of work.

By delving deeper into the stroke-related factors and each variable through data science, we can evaluate the statistics and patterns that may be more prevalent for stroke patients. For example, we may find that the likelihood of a person getting a stroke varies depending on a patient’s race or gender. Modeling and pattern-finding can provide more insight into this issue. Most importantly, identifying a higher likelihood of a person’s stroke presence based on similar patient’s records may lead to early action for treatment and increased prevention of stroke. In result, the telling of the data may lead to reduced stroke-related costs for health care services, patients, and business.

2. Modeling Ideas

Stroke prediction is a data science classification problem. Since the task has a target variable, it can be modeled as supervised learning. However unsupervised methods such as clustering can be used as a part of data exploration to understand the various segments in data. The stroke prediction problem has these columns in one training example; age, gender, hypertension, heart_disease, ever_married, work_type, Residence_type and avg_glucose_level,bm, smoking_status, and stroke. Given these features age, gender, hypertension, heart_disease, ever_married, work_type, Residence_type and avg_glucose_level,bm, and smoking_statusi; the task is to predict the presence or absence of the stroke disease. The presence of stroke is encoded as 1 and the absence is encoded as 1.

3. Data Details

We import the data from a csv file downloaded in Kaggle.com (https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset) . In our dataset, the id specifies each patient, and we have 10 attributes for each patient to predict the target variable: strock or not strock. We have a total of **5110** instances in our dataset. As we conduct supervised learning, we will split the dataset to training, validation and test dataset in the future.

We have 12 variables in dataset:

- The ‘id’ columns is used to to specify each instance
- the ‘age’ is a numerical attribute
- the ‘gender’ is a polynomial attribute
- The ‘hypertension’ is a binomial attribute indicates if a patient has hypertension
- The ‘heart_disease’ is a binomial attribute indicates if a patient has heart disease
- The ‘ever_married’ is a binomial attribute indicates if a patient married or married before
- The ‘work_type’ is a polynomial attribute
- The ‘Residence_type’ a binomial attribute indicates if a patient from urban or rural
- The ‘average_glucose_level’ is a numerical attribute
- The ‘bmi’ is a numerical attribute
- The ‘smoking_status’ is a polynomial attribute
- the ‘stroke’ is our binomial, **target variable**, 0 indicates not stroke and 1 indicates stroke

For this classification problem, the most prevalent class for our target variable is 0 (not stroke), the ratio is **0.95**. The class 1 has 4861 instances, the class 0 has 249 instances.