كلية علوم وهندسة الحاسب
College of Computer Science & Engineering
جامعة جدة
University of Jeddah

Predicting strokes using Machine Learning

By

Omar Khalid Alharthy	**1945081**
Fahad Bakoor	**1945869**
Abdalmjed Alqurashi	**1945839**

Supervisor

Dr: Zakariya Nasser

College of Computer Science and Engineering, Department of Information Technology

Jeddah University, SAUDI ARABIA

Content table

Chapter 1

Introduction

Strokes occur when there is a significant reduction or obstruction in the blood flow to the brain. Fat build-up is typically the cause of obstruction symptoms can vary. Mild symptoms are present in some people. Others display severe symptoms. Some individuals experience nausea—shortness of breath, dizziness, or sudden dizziness without exhibiting other symptoms. Older individuals are especially vulnerable.[1] Machine learning (ML) offers a quick and accurate prediction result, and it has developed into a potent tool in healthcare settings, providing stroke patients with individualized therapeutic care. Although there is a definite need for study, some research fields are not receiving adequate attention for scientific exploration, despite the expanding use of ML and Deep Learning in healthcare. Therefore, the purpose of this work is to carefully review papers for each category after classifying state-of-the-art ML approaches for brain stroke into four groups based on their functionalities or similarities. Results from the ScienceDirect web scientific database on ML for brain stroke from 2007 to 2019 revealed 39 studies. SVM (Support Vector Machine), or Support Vector Machine, is determined to be ideal. Models for stroke issues in ten investigations. In addition, most studies are on stroke diagnosis, whereas the majority are on stroke treatment, indicating a research gap that needs to be filled. Similarly to this, CT pictures are a standard dataset for strokes. Finally, effective methods employed for each category include SVM and Random Forests. The current study highlights the value of various ML techniques used in brain stroke.[2] In this work, we will investigate the use of Machine Learning to predict strokes before they occur using features such as age, blood pressure, ergonomics, etc. We hope this helps physicians save more lives.

Objectives:

This research project aims to improve patient healthcare outcomes and lower healthcare spending on heart-stroke treatment. These objectives can be achieved by:

- detecting the disease before complications.
- Identify risk factors and reduce the risk of heart strokes.

Target users:

- The Patients

- Hospital Visitors

- Physician

The tools used in this work:

- Python
- Microsoft word
- Microsoft Excel
- Tableau

Chapter 2:

Studies and research on Strokes

The risk of stroke (fatal and nonfatal combined) increased with age by 9% (95% CI, 9% to 10%) in males and by 10% (9% to 10%) in women. Both men and women experienced an increase in risk associated with a 10-mm Hg increase in systolic blood pressure (28%; 24% to 32%). Smoking increased risk similarly for males (82%; 66% to 100%) and women (104%; 78% to 133%). Increases in the body mass index had very little of an impact. Higher levels of HDL (High-Density Lipoprotein) cholesterol reduced the risk of stroke more in women than in men (hazard ratio per mmol/L 0.58; 0.49 to 0.68) (0.80; 0.69 to 0.92). The effects of the various risk variables varied slightly between nations and regions, highlighting the importance of high blood pressure in Central Europe (Poland and Lithuania). The estimates of the relative risks for stroke caused by the traditional risk factors are broken down by age, sex, and area in different parts of Europe. An important lesson in public health is that smoking raises the risk of Stroke across Europe [4].

In a medical condition known as a stroke, the blood vessels in the brain are ruptured, harming the brain. Symptoms may appear when the brain's flow of blood and other nutrients is disrupted. The World Health Organization (WHO) claims that stroke is the leading cause of death and disability worldwide. Early detection of the numerous stroke warning symptoms can lessen the stroke's severity. Various machine learning (ML) Models have been created to forecast the probability that a brain stroke would occur. This study develops four distinct models for accurate voting using a variety of physiological characteristics and machine learning methods, such as Logistic Regression (LR), Decision Tree (DT) Classification, Random Forest (RF) Classification, and Voting Classifier, Prediction. With an accuracy of almost 96%, Random Forest was the most accurate algorithm for this challenge. The open-access Stroke Prediction dataset was used in the method's development. The accuracy % of the models employed in this study is substantially greater than that of earlier research, demonstrating the reliability of the models utilized. Numerous model comparisons have demonstrated their robustness, and the scheme may be inferred from the study analysis [5].

Over 16 years, patients were prospectively examined using a common diagnostic technique. The patient was evaluated upon admission, seven days, one, three, and six months after release, and then once a year for up to ten years following the stroke. Body mass index (BMI) was used to categorize the study participants into three groups: average weight (25 kg/m2), overweight (25–29.9 kg/m2), and obese (30 kg/m2). The main goal was overall survival throughout the follow-up. The overall composite cardiovascular events over the trial period served as the secondary endpoint. Our inclusion criteria led to the recruitment of 2785 patients. BMI revealed that 504 (18.1%) patients were obese, 1113 (41.0%) were overweight, and 1138 (40.9%) patients were of normal weight. The average NIHSS(NIH Stroke Scale) score at admission (11.28 8.65) did not differ between the research groups. In comparison to patients of average weight (90.2%; 95% CI, 88.4%-92.0%), early (first week) survival was substantially greater in obese (96.4%; 95% CI, 94.8%-97.9%) and overweight (92.8%; 95% CI, 91.2%-94.4%) patients. Similar to this, the 10-year survival rate for obese patients was 52.5% (95% CI, 46.4%-58.6%); for overweight patients, it was 47.4% (95% CI, 43.5%-51.3%), and for normal-weight individuals, it was 41.5% (95% CI, 39.7%-45.0%) (log-rank test=17.7; P0.0001). Compared to patients of average weight, those who were overweight or obese had a considerably lower chance of dying within ten years (HR, 0.82; 95% CI, 0.71-0.94; and HR, 0.71; 0.59-0.86). Removing all potentially confounding factors [6].

The top cause of death in China and Japan, Stroke is the second-highest cause globally. Its avoidance is a primary objective. To choose the right level of intervention, it is essential to identify primary stroke risk, mainly through newly personalized risk variables like indicators of considerable artery damage like arterial stiffening. The biology of arterial stiffness, its predictive value for stroke, and the treatment implications of this risk factor for stroke prevention are the main topics of this review. In a longitudinal study that included 1715 patients with essential hypertension and measures of carotid-femoral pulse wave velocity (PWV), an indication of arterial stiffness at admission showed the predictive significance of arterial stiffness for stroke. A nasty follow-up PWV significantly predicted stroke over a mean follow-up period of 7.9 years, when 25 fatal strokes occurred (relative risk = 1.39 [(95% CI 1.08, 1.72]; p = 0.02 for each four m/sec increase). This prediction was made independent of traditional cardiovascular risk factors, such as age, cholesterol level, diabetes mellitus, smoking, and mean blood pressure. To confirm the prognostic efficacy of aorta stiffness in primary and secondary events, in low- and high-risk groups, in different countries, and using diverse arterial stiffness measuring technologies, additional longitudinal studies are required. In addition to addressing cardiovascular risk factors such as hypertension, dyslipidemia, diabetes mellitus, and smoking, all linked to arterial stiffening, drug therapy may prevent stroke by reducing arterial stiffness due to the significant local initiatives. Drugs that interfere with the renin-angiotensin-aldosterone pathway should be especially beneficial in light of the significant local activities of angiotensin II on arterial stiffness. Using statins' non-lipid-lowering effects and direct targeting of the molecular processes that cause arterial stiffening, such as the formation of advanced glycation end products, are two promising therapeutic approaches to reduce arterial stiffness [7].

Despite mounting evidence that food-based dietary patterns may lower the risk of cardiovascular disease, little is known about the food amounts that have the most significant impact on the risk of various cardiovascular outcomes, as well as the caliber of the meta-analysis. The goal of this meta-analysis was to combine information on the risk of coronary heart disease (CHD), stroke, and heart failure and the consumption of 12 major food groups, including whole grains, refined grains, vegetables, fruits, nuts, legumes, eggs, dairy, fish, and red meat. Processed meat and sugar-sweetened beverages (SSB) (HF). Methods: Up until March 2017, we systematically searched PubMed and Embase for prospective studies. Using a random effects model, summary risk ratios (RRs) and 95% confidence intervals (CIs) were calculated for both linear and non-linear relationships and for the highest versus lowest intake categories [8].

After post-stroke atrial fibrillation surveillance, the prevalence of atrial fibrillation increases to about one-third of all ischemic strokes. According to data from stroke registries, most of these strokes—either fatal or severely disabling—are caused by atrial fibrillation, which is both undiagnosed and untreated or inadequately treated. The majority could be avoided if efforts were focused on early atrial fibrillation detection through screening or case finding, as well as treatment of all patients with atrial fibrillation at high risk of stroke with carefully monitored vitamin K antagonists or non-vitamin K antagonist anticoagulants. Unless determined to be a shallow risk by detailed validated risk ratings, such as CHA2DS2-Vasc, the default practice should be administering anticoagulant thromboprophylaxis to all patients with atrial fibrillation. Evaluation of bleeding risk utilizing the HAS-BLED score should highlight the risk factors for reversible bleeding. Last but not least, patients require encouragement from their doctors and a variety of other sources to begin anticoagulant treatment and ensure long-term adherence and persistence with treatment [9].

Diabetes causes the circulatory system to age more quickly. Compared to people without diabetes, people with diabetes have a stroke risk of around twice as high. Although hyperglycemia is a significant risk factor for poor outcomes following stroke, it may not be the cause of those outcomes. It has not been demonstrated that reducing glucose is related to a better prognosis. Similarly, long-term glucose-lowering medications do not reduce the risk of stroke in diabetic patients. Diabetic individuals continue to receive standard care for stroke prevention and therapy. However, with the recent availability of numerous novel agents, the future might be more promising. We are anticipating how these drugs may affect macrovascular problems like stroke.[10]

Although meta-analyses are required to quantify this risk, stroke is a recognized risk factor for dementia from any cause. We looked for studies comparing the risk of dementia from any cause versus a comparison group without Stroke on Medline, PsycINFO, and Embase. Meta-regression was utilized to look into potential effect modifiers, and random effects meta-analysis was used to pool adjusted values from some research. With 1.9 million individuals, we found 36 research on common stroke and 12 studies on incident stroke (1.3 million participants). The pooled hazard ratio for dementia from all causes of prevalent stroke was 1.69 (95% confidence interval: 1.49-1.92; P .00001; I2 = 87%). The pooled risk ratio for an incident stroke was 2.18 (95% CI: 1.90-2.50; P .00001; I2 = 88%). Study-specific factors had little impact on these correlations, except for sex, which accounted for 50.2% of the variation in the prevalence of stroke between studies. Strong, independent, and possibly modifiable risk factors for dementia from all causes include stroke [11].

According to current recommendations, intravenous thrombolysis is only used to treat acute stroke if it can be shown that it has been less than 4.5 hours after the onset of symptoms. We wanted to know if individuals with strokes that had an unclear time of start and characteristics on magnetic resonance imaging (MRI) that suggested a recent cerebral infarction would benefit from thrombolysis with intravenous alteplase. We randomly assigned patients with an uncertain stroke time to start receiving either intravenous alteplase or a placebo in a multicenter trial. MRI diffusion-weighted imaging revealed an ischemic lesion in all patients. Although fluid-attenuated inversion recovery (FLAIR) showed no parenchymal hyperintensity, indicating that the stroke had occurred within the previous 4.5 hours. Patients whose thrombectomy was scheduled were omitted. At 90 days, a favorable outcome was considered to have occurred if the modified Rankin scale of neurologic disability (which ranges from 0 [no symptoms] to 6 [death]) had a score of 0 or 1. Altereplase's potential to result in lower ordinal scores on the modified Rankin scale than would placebo (shift analysis) [12].

Patients with acute stroke are overrepresented in both diabetes and hyperglycemia. Poor stroke outcomes are linked to hyperglycemia. Although symptomatic cerebral hemorrhagic transformation is more likely in diabetes and hyperglycemia, blood sugar levels do not appear to affect how well thrombolysis works. Although there is limited evidence from supportive randomized controlled trials describing the advantages and disadvantages of insulin administration for hyperglycemia in stroke, evidence from general patients treated in intensive care units suggests that intensive control of hyperglycemia may improve early outcomes. As a result, this evidence cannot be directly extrapolated to patients with acute stroke. The American guidelines are weaker than the European ones, which advise that glucose control may be prudent and set a definitive intervention threshold of 10 mmol/l. Adjusted insulin infusions or glucose-potassium infusions have their supporters, and both entail a slight risk of hypoglycemia despite success. It would seem prudent to follow a regimen that has been approved locally [13].

Stroke has a significant socioeconomic influence on society worldwide. In terms of media attention, patient and caregiver understanding, service advancements, and research, stroke is assuming an increasing influence. Over 9 million stroke survivors make up the 4.5 million stroke fatalities annually around the world. If they live to be 85 years old, nearly one in four men and nearly one in five women their age can anticipate having a stroke. A stroke occurs from 2 to 25 times per thousand people nationwide. 15–40% of cases return within five years. Compared to 1983, it is predicted that by 2023, there will be a 30% absolute increase in the number of people having their first stroke. The overall prevalence rate is about 5 per 1,000 people. Stroke is the leading cause of adult disability, with 65% of survivors functionally independent one year after a stroke [14].

The decision to deliver thrombolysis, a treatment that might lead to a favorable recovery or deterioration due to symptomatic cerebral hemorrhage, is crucial in the emergency treatment of ischemic Stroke (SICH). When combined with clinical factors, specific imaging characteristics from early computer tomography (CT) have been proven to predict SICH, albeit with only fair accuracy. In this proof-of-concept trial, we explore whether machine learning of CT images can identify patients who will develop SICH rather than show clinical improvement without hemorrhage after receiving tPA. Retrospective clinical records and CT scans of 116 individuals with acute ischemic stroke who received intravenous thrombolysis were gathered (including 16 who developed SICH). The sample was divided repeatedly for 1760 distinct combinations into training (n = 106) and test sets (n = 10). Clinical severity and CT brain pictures were inputs to a support vector machine (SVM). The SVM's performance was compared to well-known prognostication techniques (SEDAN and HAT scores, original or after adaptation to our cohort). The area under the receiver-operating-characteristic curve (AUC) measurements of the SVM's predictive ability (0.744) showed a favorable comparison to prognostic scores (original and modified versions: 0.626-0.720; p 0.01). Additionally, assuming a 10% SICH frequency, the SVM detected 9 out of 16 SICHs instead of 1–5 using prognostic ratings (p 0.001).In conclusion, machine learning techniques applied to acute stroke CT images provide automation and perhaps even better performance for SICH prediction after thrombolysis. Such techniques should be evaluated with larger cohorts and the inclusion of sophisticated imaging [15].

The leading top organ of the human body is the brain. In a medical condition known as a stroke, the blood vessels in the brain are ruptured, harming the brain. Symptoms may appear when the brain's flow of blood and other nutrients is disrupted. Stroke is considered an emergency medical issue since it frequently results in death, long-term neurological impairment, and complications. According to the World Health Organization (WHO), stroke is the leading cause of death and disability. The severity of a stroke can be lowered by early recognition of the numerous warning signs. The primary goal of this project is to use deep learning and machine learning techniques to predict the likelihood of a brain stroke happening at an early stage. A trustworthy dataset for stroke prediction was downloaded from the Kaggle website to test the algorithm's efficacy. Extreme Gradient Boosting (Ada Boost, Light Gradient Boosting Machine, Random Forest, Decision Tree, Logistic Regression, K Neighbors, SVM (Support vector machine)-Linear Kernel, Naive Bayes, and deep neural networks (3-layer and 4-layer ANN) were among the classification models that were successfully used in this study for classification tasks. The classification accuracy of the Random Forest classifier is 99%, which came in at the top (among the machine learning classifiers). When using the chosen features as input, the three-layer deep neural network (4-Layer ANN) technique gave a greater accuracy of 92.39%. The study demonstrated that machine learning methods performed better than deep neural networks [16].

An unanticipated restriction in blood flow to the brain and heart brings on most strokes. By being aware of the numerous stroke warning signs beforehand, stroke severity might be decreased. If blood flow to a part of the brain abruptly stops, a stroke may follow. In this study, we describe an approach for applying Logistic Regression (LR) algorithms to predict the early onset of stroke illness. Preprocessing methods like SMOTE, feature selection, and outlier handling was used on the dataset to enhance the model's performance. This technique assisted in balancing the distribution of classes, locating and eliminating unnecessary features, and dealing with outliers. Elevated blood pressure, body mass, cardiac problems, normal blood sugar levels, smoking status, previous stroke, and age play a role. Depending on whatever part of the brain is impacted by the decreased blood flow, impairment happens as the neurons in that area gradually die. Early symptom identification is crucial for predicting stroke and promoting a healthy lifestyle. Additionally, we conducted an experiment utilizing logistic regression (LR). We contrasted it with several other research using the same dataset and machine learning model, namely logistic regression (LR). The results showed that our method had the most excellent F1 score and area under the curve (AUC) score compared to the other five researchers in the same field. Making it a valuable tool for predicting stroke disease with an accuracy of 86%. The predictive model for stroke has potential uses; thus, it is still essential to researchers and professionals in the medical and health sciences [17].

Currently, clinicians must manage a significant amount of complex clinical, laboratory, and imaging data, which calls for advanced analytic techniques. This enormous amount of data can be used to build forecasting models using machine learning-based methods. We used machine learning models with clinical, laboratory, and quantitative imaging data as inputs to predict short- and medium-term functional outcomes in acute ischemic (AIS) patients with proximal, middle cerebral artery (MCA) occlusions. Consecutive AIS patients with proximal M2 and MCA M1 occlusions were also included. To forecast the result, xgboost, LightGBM, CatBoost, and Random Forest were employed. When choosing features, minimum redundancy and maximum relevancy were employed. The primary outcomes were the modified Rankin Score (mRS) at 90 days and the National Institutes of Health Stroke Scale (NIHSS) shift. The algorithm that predicted the favorable and unfavorable result groups at 90 days had the most significant area under the receiver operating characteristic curve (AUROC). This method is called When predicting the favorable and unfavorable groups based on the NIHSS shift; Random Forest obtained the highest AUROC. We successfully predicted the functional outcome of AIS patients with proximal MCA occlusions using clinical, laboratory, and imaging characteristics in combination with machine learning [18].

According to the World Health Organization, strokes are the second most common cause of death worldwide (WHO). In several areas of stroke management, information technology (IT), particularly machine learning (ML), may be advantageous and beneficial. However, the vast majority of research now concentrates on creating ML models for dealing with such instances without examining the built models' level of confidence and dependability. Diverse metric functions must be estimated to improve model performance and identify the underlying datasets' critical elements. So, to ensure that different ML models produce accurate and trustworthy results, this research investigates whether or not their results are accurate and realistic... In light of this, numerous models, including Support Vector Classifier, K-Nearest Neighbors, Logistic Regression, Random Forest, XGB Classifier, and LGBM Classifier, were developed to forecast the likelihood of stroke. The most appropriate and precise model for stroke prediction is reached after comparing all the collected findings using the selected metric functions [19].

One of the most common causes of death across the globe is stroke. The inability of the brain to get nutrients owing to blocked or bleeding blood vessels is known as a stroke. Time is essential to successful stroke treatment since early identification and treatment of stroke are linked to better patient outcomes. In this chapter, we use deep learning architectures to explore the classification of strokes using the Brain Stroke CT Dataset. A total of 2501 brain stroke computed tomography scans were used in the investigation (CT). For training and testing, photographs were employed. The classification of brain stroke CT scans as usual and as the stroke was accomplished using a variety of well-known pre-trained convolutional neural networks (CNNs), including GoogleNet, alexnet, VGG-16, VGG-19, and Residual CNN. The classifier's performance is measured using several performance metrics, including accuracy (ACC), specificity (SPE), sensitivity (SEN), and F-score. VGG-19 has the best classification results, with ACC 97.06%, SEN 97.41%, SPE 96.49%, and F-score 96.95% [20].

Chapter 3:

Data Quality.

Data understanding (collecting)

Data collection:

After searching for the most suitable dataset, we found the Stroke Dataset on Kaggle website [3]. The dataset is publicly available for researcher.

Data Description:

The World Health Organization (WHO) estimates that stroke is the second leading reason of death worldwide, responsible for around 11% of all fatalities.

Based on input variables, including gender, age, different illnesses, and smoking status, this dataset is used to determine if a patient is prone to getting a stroke. Each row of the data contains pertinent information about the patient.

Data Exploration

The data has 12 columns without the id and class 10.

Attribute Information

1) id: unique identifier

2) gender: "Male," "Female," or "Other."

3) age: age of the patient

4) Hypertension: 0 if the patient does not have hypertension. One of the patients has hypertension

5) Heart Disease: 0 if the patient does not 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

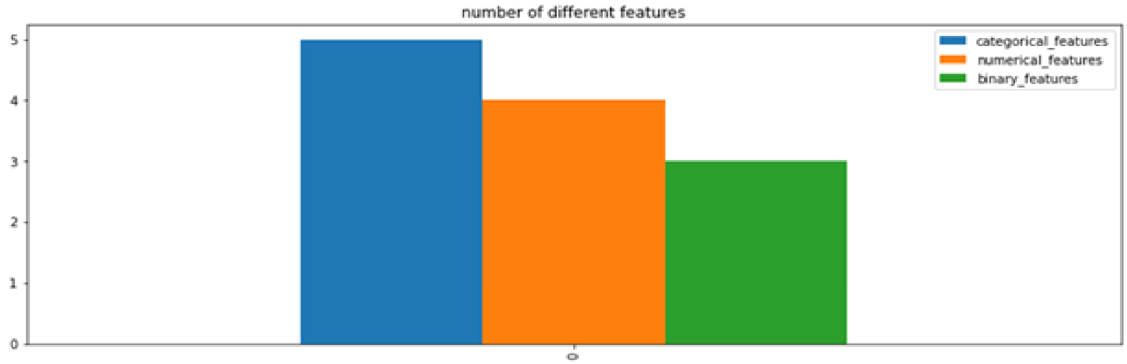
10) BMI: body mass index

11) smoking status: "formerly smoked," "never smoked," "smokes," or "Unknown"*

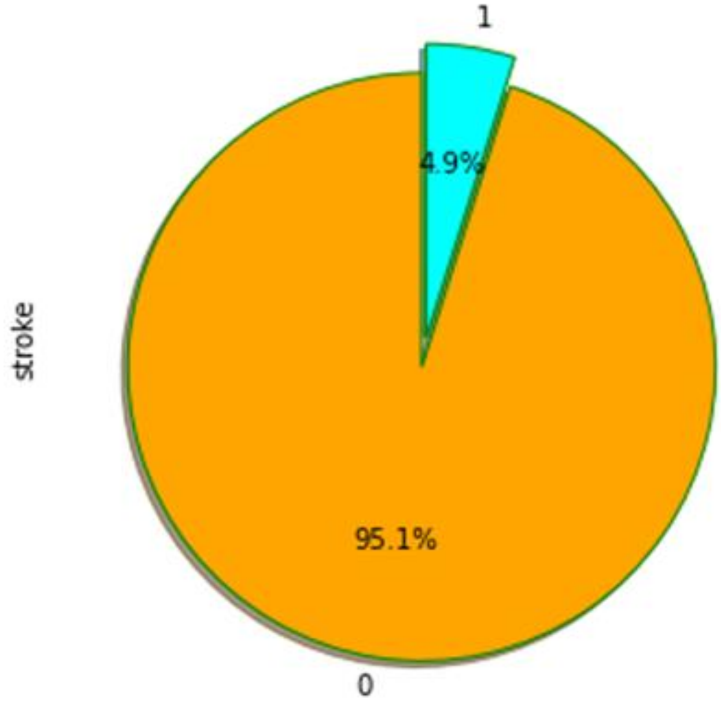
12) Stroke: 1 if the patient had a stroke or 0 if not

The data contains some missing values in the BMI (Body Mass Index) attribute (201) and some values labeled (unknown) in the smoking status attribute.

As mentioned above, the data is relatively new, containing 12 columns and 5110 rows, with 250 stroke risks and 4,860 no-stroke risks, so 5% of the rows are labeled to have a stroke risk.

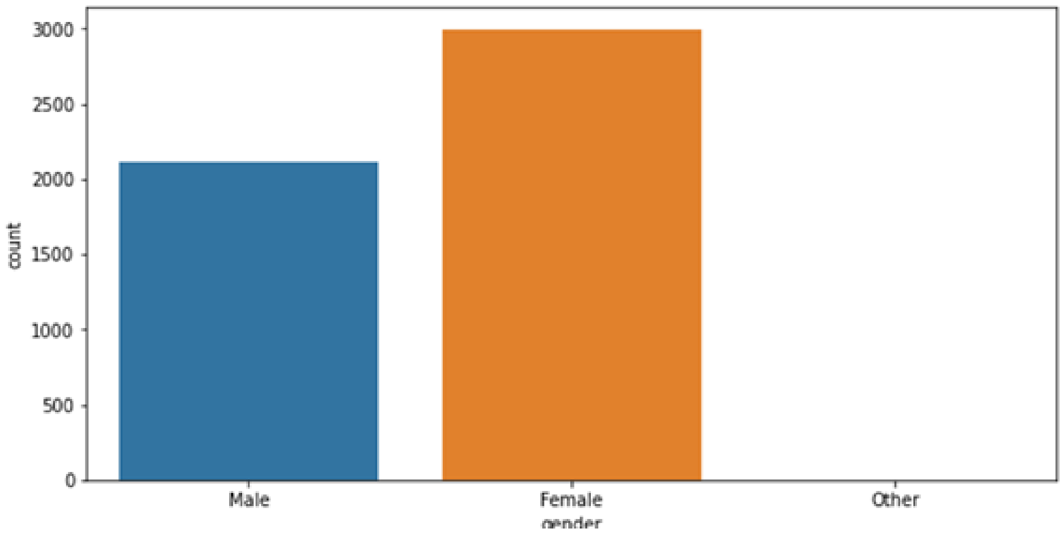


- The majority of our data is categorical data.

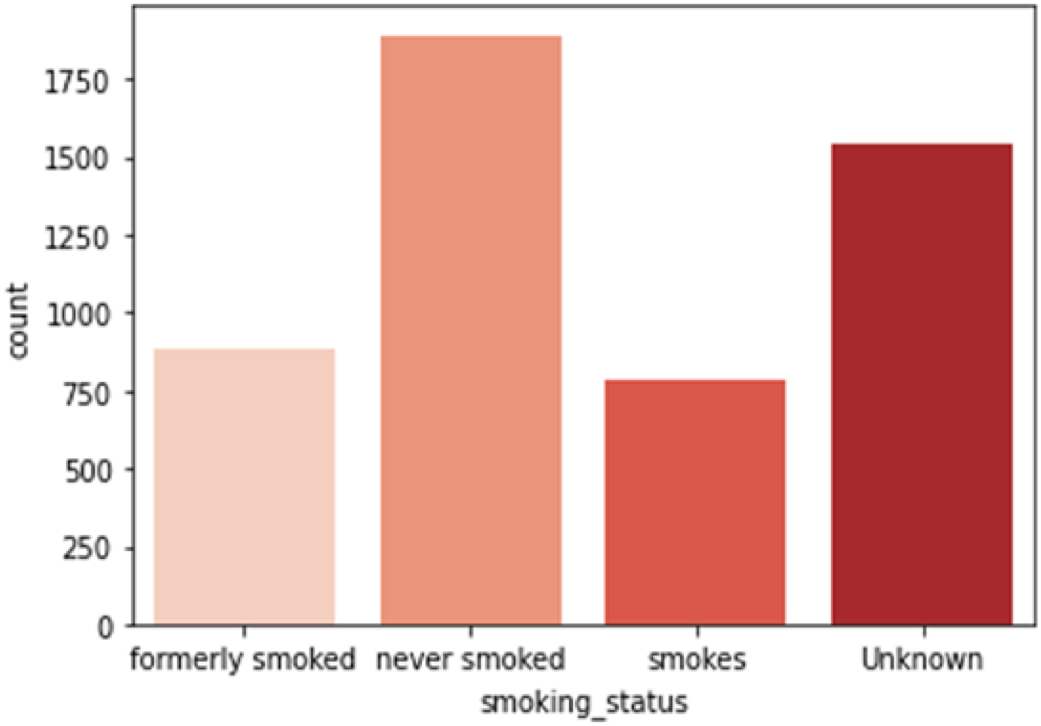


Only 5% of people have a stroke! So, our data is highly imbalanced, and we resolve this later by doing a resampling technique. So, after applying that, it will resolve our issue. For now, we have :

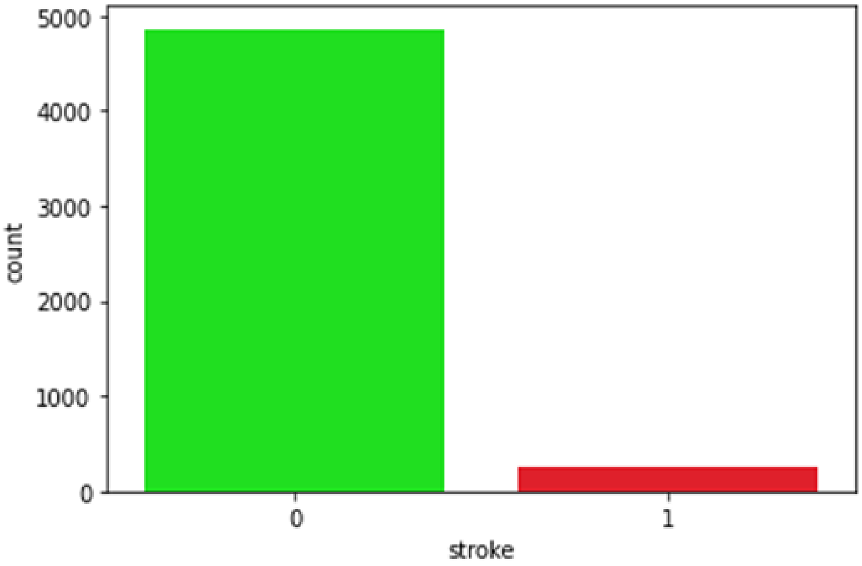
- 95% of data is not stroke. Moreover, only 5% of People in the data have a Stroke.



- There is about a 1000 difference between Females and Males in the data
- By seeing that visualization, we can see that we have only around 2000 Males in the data
- There are massive Female data, which is around 3000

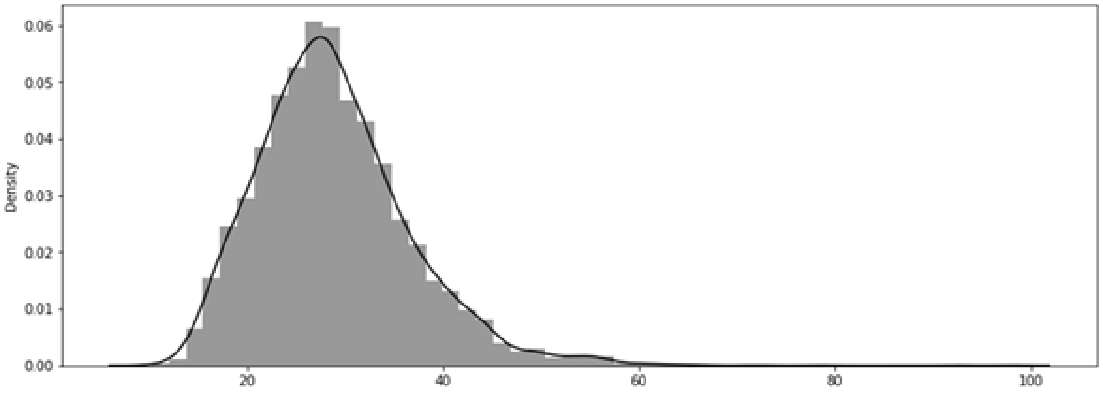


- • Many people never smoked in their lives, but we also do not know the exact status of unknowns in our dataset.
- • We have a vast number of data on those who never smoked, almost 2000.
- • the number of people who formerly smoked is around 800.
- • We have people who only smoke, who are around 750.



From the above dependent variables, fewer people suffered a stroke. But this also

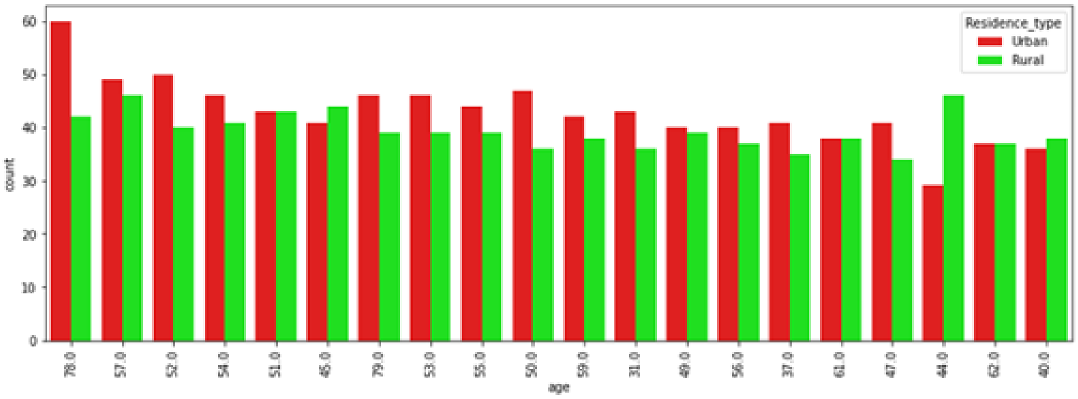
This means that our dataset needs to be balanced. We will likely have to use sampling techniques to make the data balance.

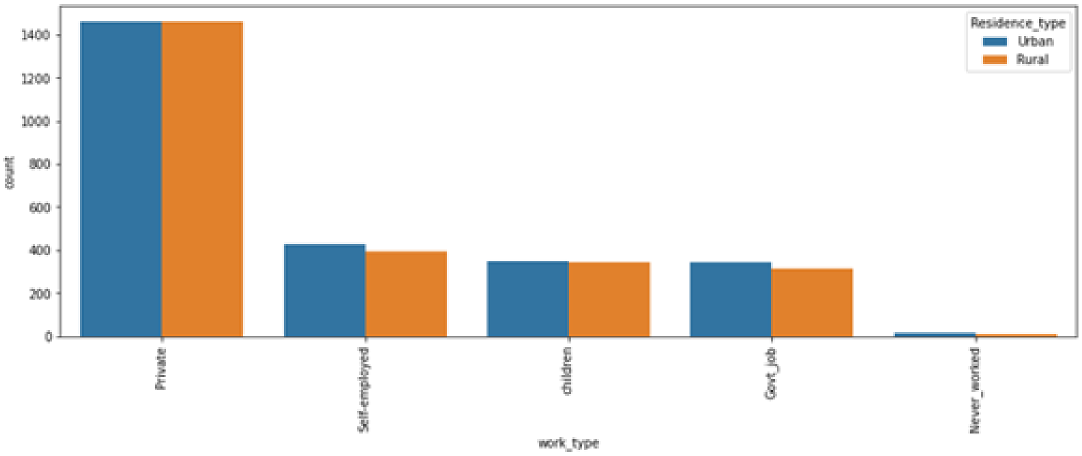


- Dealing with data imbalance
- The minority class may be oversampled using SMOTE (Synthetic Minority Oversampling Technique), whereas the majority class can be under-sampled using Tomek Links due to the stroke's extreme imbalance.
- SMOTE (Synthetic Minority Oversampling Technique) will be my tool for this. All categorical variables must be changed to ints

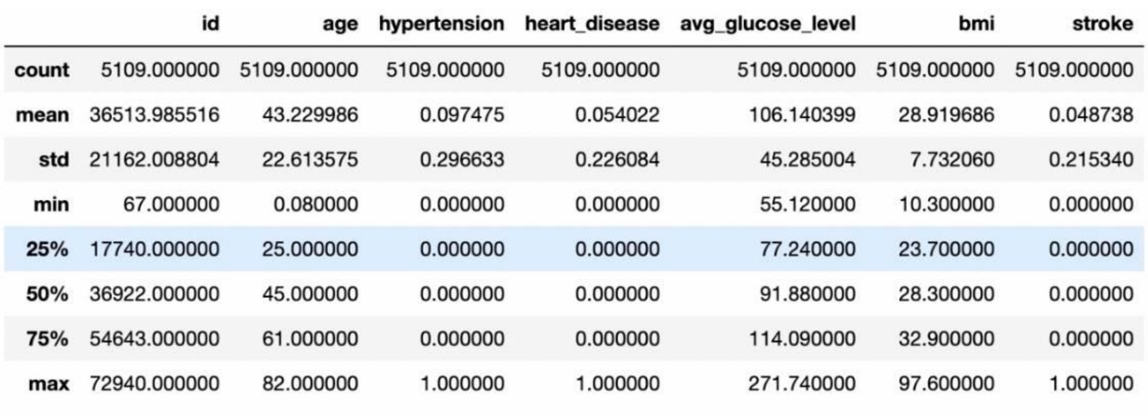
To use SMOTE. We will not be One-Hot-Encoding them since we will later create a model.

We will be Label-Encoding each of them.

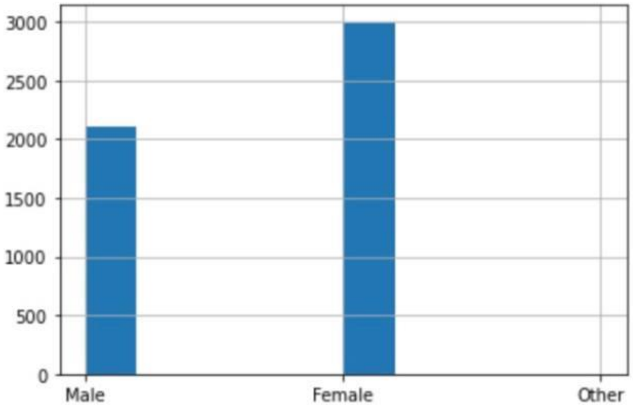




Describe Table



	id	age	hypertension	heart_disease	avg_glucose_level	bmi	stroke
count	5109.000000	5109.000000	5109.000000	5109.000000	5109.000000	5109.000000	5109.000000
mean	36513.985516	43.229986	0.097475	0.054022	106.140399	28.919686	0.048738
std	21162.008804	22.613575	0.296633	0.226084	45.285004	7.732060	0.215340
min	67.000000	0.080000	0.000000	0.000000	55.120000	10.300000	0.000000
25%	17740.000000	25.000000	0.000000	0.000000	77.240000	23.700000	0.000000
50%	36922.000000	45.000000	0.000000	0.000000	91.880000	28.300000	0.000000
75%	54643.000000	61.000000	0.000000	0.000000	114.090000	32.900000	0.000000
max	72940.000000	82.000000	1.000000	1.000000	271.740000	97.600000	1.000000

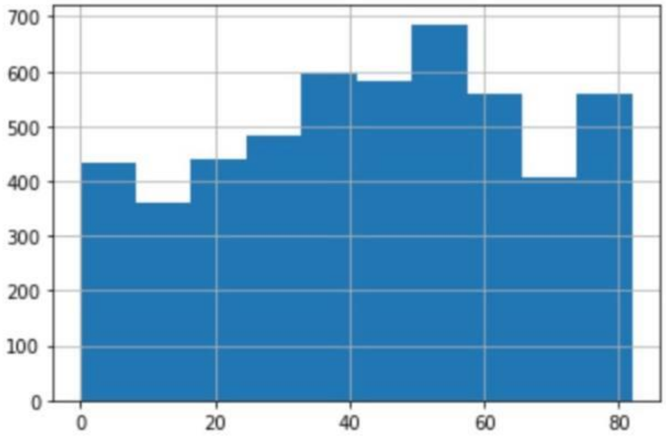


By seeing that visualization, we can see that we have only around 2000 Males in the data.

We have massive Female data, which is around 3000

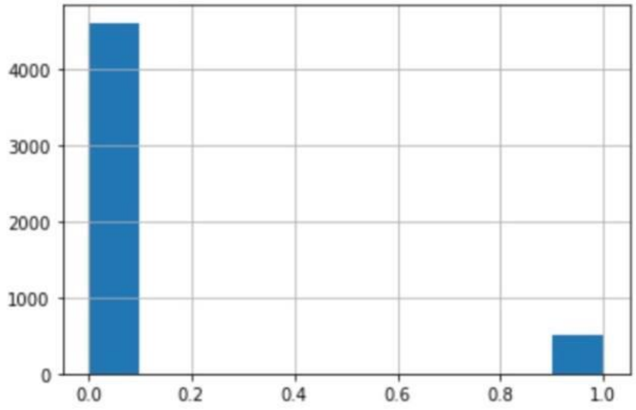
And on the other hand, we do not have any row

Histogram of Age :



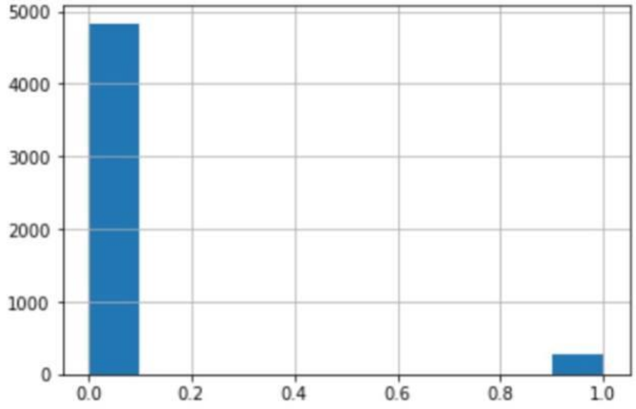
- Age is usually distributed.

Histogram of Hypertension:



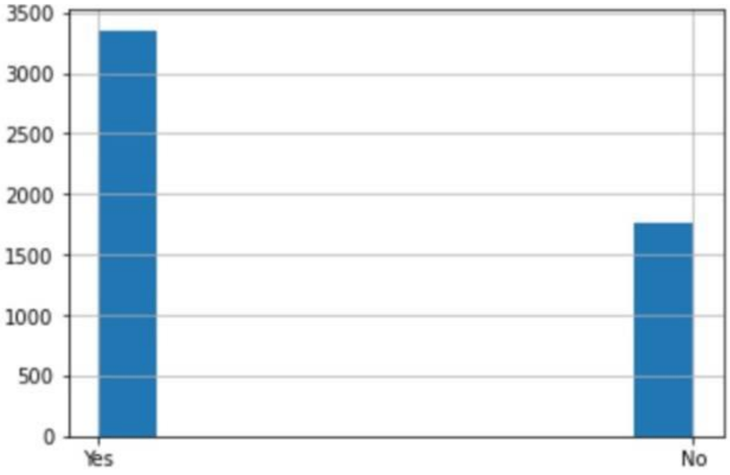
- A small percentage of the data have hypertension.

Histogram of Heart disease:



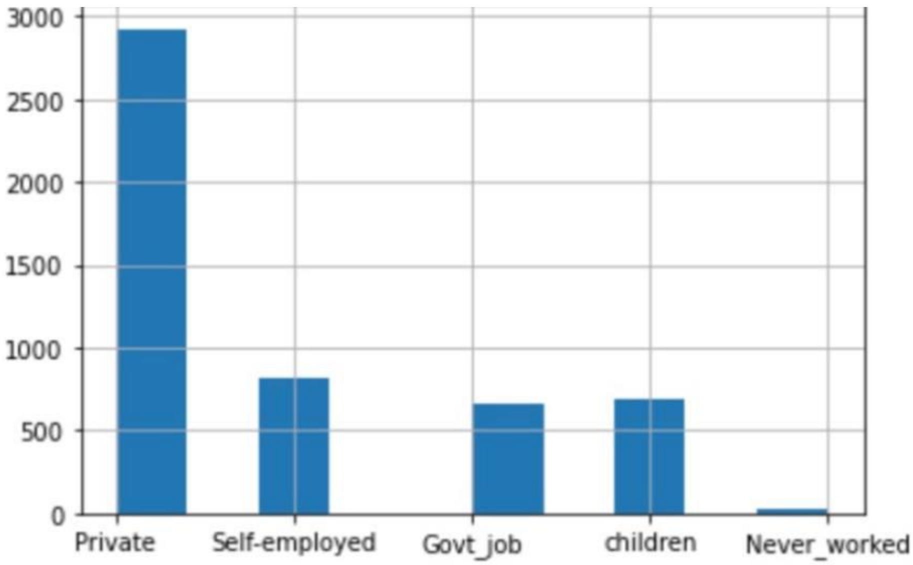
- A small percentage of the data have any heart diseases.

Histogram of Ever-Married :



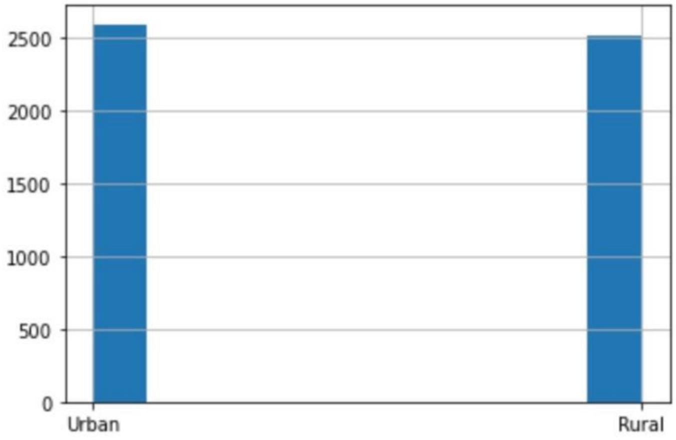
- More people are married in the data set.

Histogram of Work-type :



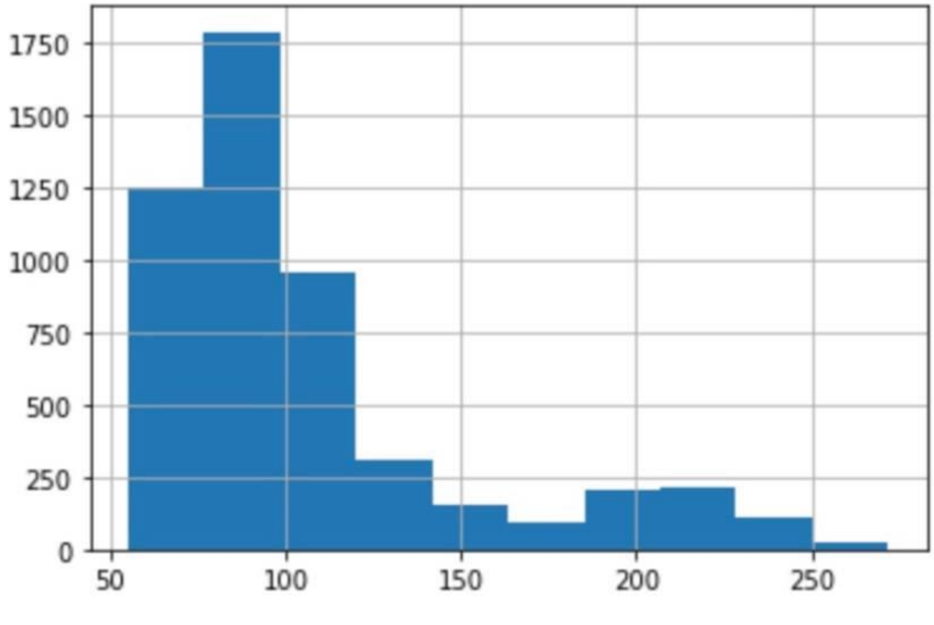
- Most of the work type is private
- The minor work type is never worked

Histogram of Residence-type :



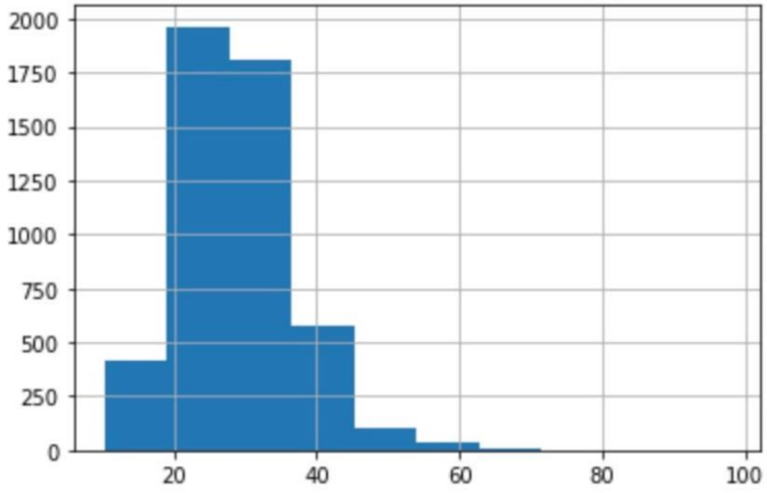
- There is slightly more urban residence.
- Generally, the two categories are almost equal.

Histogram of Avg-glucose-level :



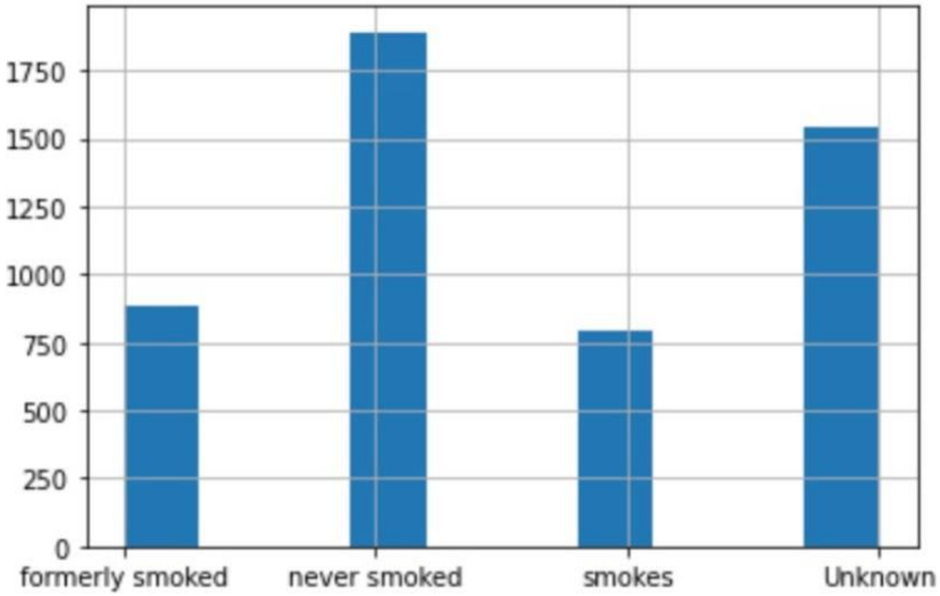
- The graph is skewed to the right.

Histogram of BMI :



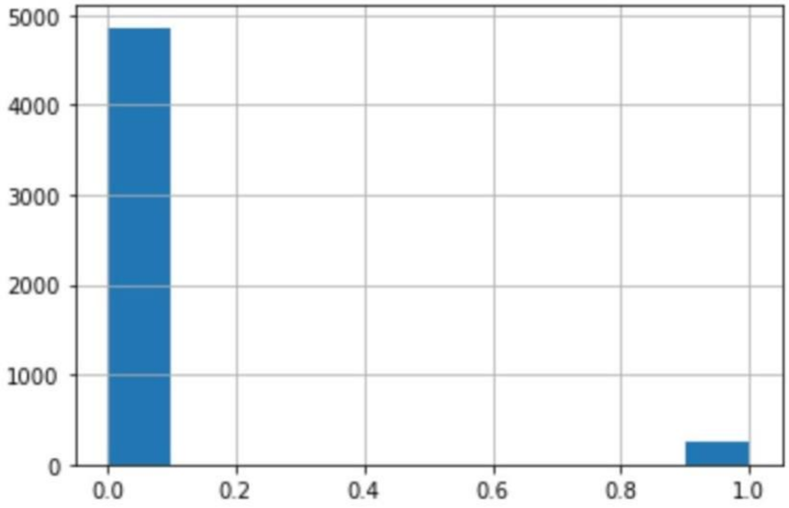
- The graph is skewed on the right.

Histogram of Smoking-status :



- Most of the values are labeled as never smoked
- The least is smokes
- A significant percentage of the values are labeled unknown

Histogram of Stroke :



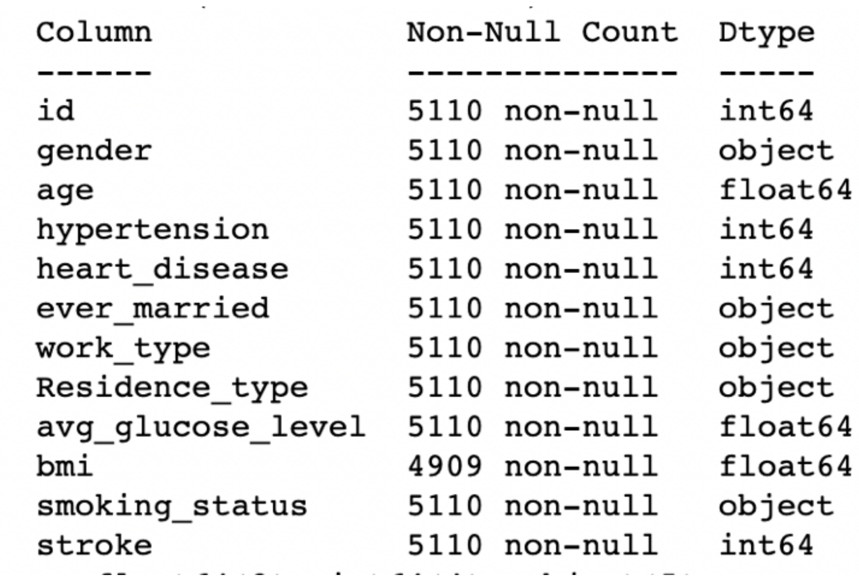
- As we mentioned before, there is an unbalance in the class

So, we need to handle these distributions by performing transformations and making them closer to normal distributions to improve the model performance.

Chapter 4:

Data preparation

Some information about the columns in the dataset

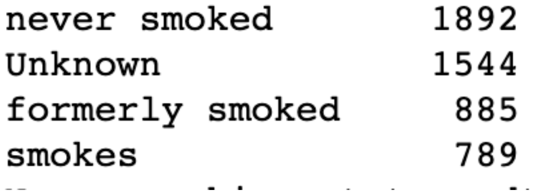


```
Column             Non-Null Count  Dtype
------             --------------  -----
id                 5110 non-null   int64
gender             5110 non-null   object
age                5110 non-null   float64
hypertension       5110 non-null   int64
heart_disease      5110 non-null   int64
ever_married       5110 non-null   object
work_type          5110 non-null   object
Residence_type     5110 non-null   object
avg_glucose_level  5110 non-null   float64
bmi                4909 non-null   float64
smoking_status     5110 non-null   object
stroke             5110 non-null   int64
```

The column "BMI" has 201 missing values.

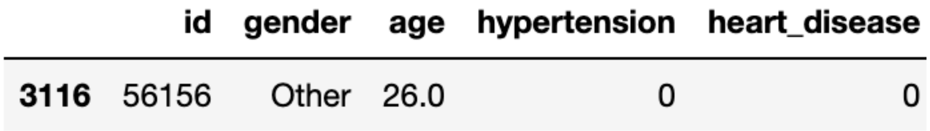
Here we can use many methods to deal with the missing values; for example, dropping the missing values from the table or filling the missing data with the average. Nevertheless, first, we decided to use the machine learning algorithm decision trees to predict the missing value.

We all tried using the decision tree algorithm to predict the missing values; this is only an initial solution, and other methods will be considered and experimented with.



```
never smoked        1892
Unknown             1544
formerly smoked      885
smokes               789
```

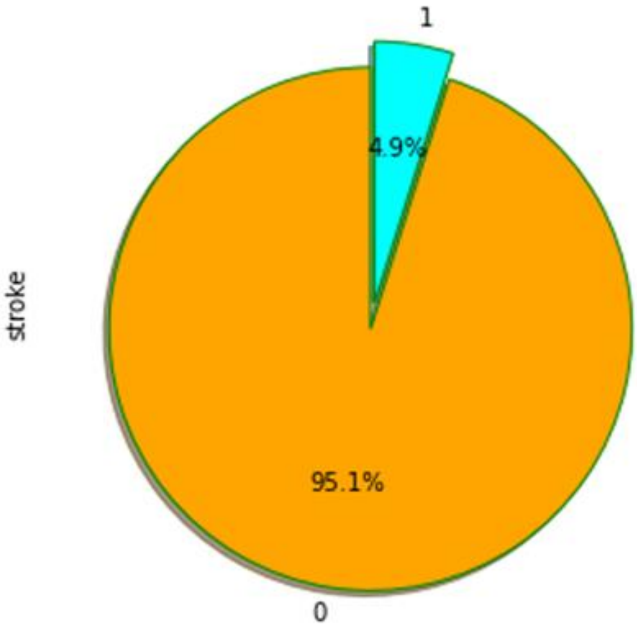
We have some smoking statuses labeled unknown.



	id	gender	age	hypertension	heart_disease
3116	56156	Other	26.0	0	0

One of the values in the column gender is labeled the other.

The stokes (the class).



The ratio of the data is 95% no strokes, and only 5% confirmed strokes.

Oversampling can solve this issue, and we will consider more different approaches throughout the project.

Conclusion

Our Plans for Next Term:

Our Plans for Next Term: We will experiment with multiple classifications algorithmics such as Naïve Bayes, KNN (K-nearest neighbors), Decision Trees, Random Forest, SVM (Support vector machine), and Deep Learning Neural Networks Measuring the accuracy between them, and we will be considering other algorithms as well.

We will try different pre-proc ing approaches to optimize the model's accuracy and avoid overfitting.

Chapter 5:

Model Building

Interdiction

This chapter will cover the tools we used to build the models and the algorithms we have used. And the oversampling technique that we chose

Experiments setup and tools :

All of our work was done using python and multiple machine learning and deep learning libraries such as TensorFlow, PyTorch, Keras, and Scikit-learn, and data manipulation and data pre-proc ing libraries such as Pandas and numpy; we also used data visualization libraries such as matplotlib, Seaborn.

The models we built are all classification models, seeing that our data has only two possible classes, whether the patient has a risk of a brain stroke or not. However, we noticed some class imbalances earlier, so we decided that oversampling could solve this problem using the library sklearn. Utils resample. We also tried to improve the model using hyperparameter tuning.

Initial parameters and selection criteria :

As mentioned earlier, we focused mainly on classification algorithms to compare them and find the best results. The algorithms in question are Random Forest, Support Vector Machine, Logistic Regression, and Artificial Neural Networks. We will also compare the results between the models before and after the oversampling.

Conclusion

We addressed the tools that we have used to deal with multiple problems that we have faced during the building of this project, the algorithms that we are using, and hyperparameter tuning to improve the results.

Chapter 6:

Results and Discussions

Introduction:

This chapter will describe the experiment results and performance evaluation metrics in great detail. Listing accuses multiple algorithms before and after oversampling.

Performance Evaluation metrics:

So we will be checking the accuracy of the model to check how close they are to the actual value, the recall to find how many times the model was able to detect a specific category, and precision to find out how good the model is at predicting a specific category, F1 score, and area under the curve which is the measure of the ability of a binary classifier to distinguish between classes. These measures were used for all models before and after oversampling to observe how much the oversampling can affect the model.

Experiments results:

In machine learning, the experiment applies a model to a data set to discover how well the model performs. We used several models and used the train test split as follows:

- Training data: contains 70% of the total dataset.

- Testing data: contains 30% of the total dataset.

The results before over-sampling:

Support vector machine (SVM):

Accuracy	Precision	Recall	F1-score
0.76	0.12	0.77	0.20

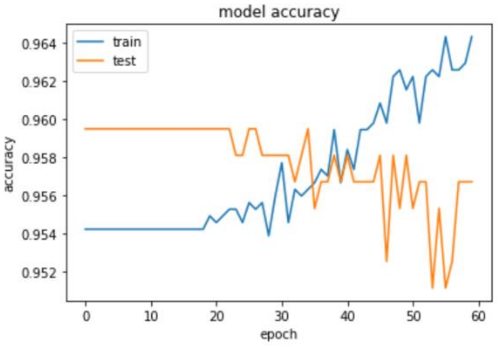
Random forest:

Accuracy	Precision	Recall	F1-score
0.89	0.11	0.20	0.14

Logistic Regression:

Accuracy	Precision	Recall	F1-score
0.75	0	0.71	0.13

Artificial Neural Network:



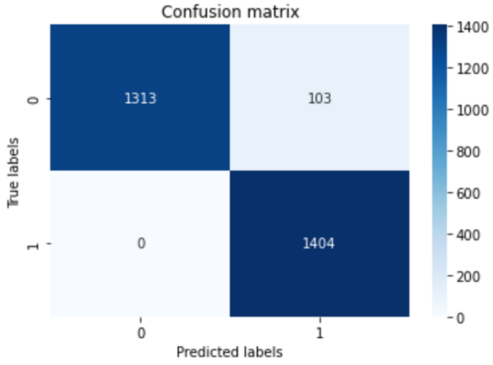
Accuracy	Precision	Recall	F1-score
0.93	0.1	0.01	0.02

Discussion:

After employing multiple models, there was a noticeable decline in the accuracy of the models employed, so we looked at the results. The correct response was to oversample the data, and the accuracy obtained was adequate.

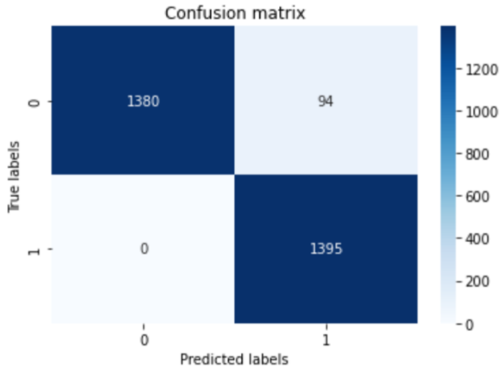
The results after oversampling:

Support vector machine (SVM):



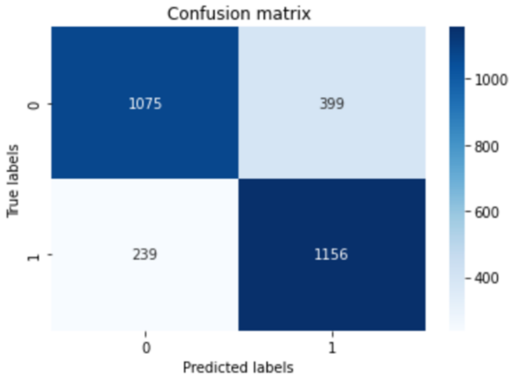
Accuracy	Precision	Recall	F1-score
0.98	1.00	0.96	0.98

Random forest:



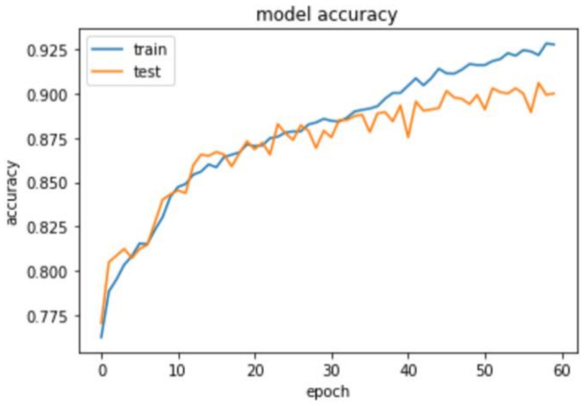
Accuracy	Precision	Recall	F1-score
0.97	0.97	0.94	1.00

Logistic Regression:



Accuracy	Precision	Recall	F1-score
0.77	0.74	0.82	0.78

Artificial Neural Network:



Accuracy	Precision	Recall	F1-score
0.90	True	0.97	0.85

Our support vector machine model acquired an accuracy of 0.98 after trying to enhance the model utilizing hyperparameter tuning, which is an acceptable result and the best accuracy so far. On the other hand, logistic regression performed the worst with 0.77 accuracies after oversampling.

Support vector machine managed to achieve the best result by accuracy metrics, and the results are as follows:

Accuracy	Precision	Recall	F1-score
0.98	1.00	0.96	0.98

Conclusion:

Our dataset was about predicting stroke attacks. Because the dataset was unbalanced, it was not easy to train our model simply. So, we used up-sampling for my minority class to have equal representation from both groups. As a result, we had 50% of both classes after oversampling.

After that, we used a one-hot encoder for category features and a standard scaler for numerical features to bring them all to the same scale. We planned to use Random Forest, support vector machine, Artificial Neural Network, and Logistic Regression for this task. First, we utilized grid search to discover the ideal hyperparameters for our models. Then we trained our final model, which gave us a good accuracy of 97% in the Random forest, an accuracy of 98% in the support vector machine ,an accuracy of 90% in the Artificial Neural Network, and last accuracy of 78% Logistic regression is the simplest algorithms and is used only for learning purposes and is not used in practical applications because of its performance issues.

Chapter 7:

Conclusion & future work

Introduction:

In this chapter, we will discuss our final thoughts and conclusions from this project, the challenges and limitations we faced and how we solved them, and possible ideas to expand on the project in the future work section.

Conclusion:

The brain strokes dataset is significant. So we were intrigued by the subject and wanted to study and understand it more before building the model. We did so by doing multiple exploratory data analyses on each feature. Then we started cleaning the dataset to ensure that it was ready for predictive analysis because we learned in our studies that "Clean input = Clean output," so we had to ensure that the data was ready and clean. as we started building the classification models we faced multiple difficulties that will be discussed in great detail in the next section. However, in the end, we managed to find an optimal accuracy of 98% using a Support vector machine, allowing the early disease prediction with great accuracy, which is the project's main objective.

Difficulties & limitations:

During the model-building phase, we encountered many challenges and limitations with the data; the most significant difficulty and limitation was the dataset size which is 5111 records with a vast class imbalance. For example, only 5% of the records had a stroke (4861 records with no strokes, only 249 with strokes). The solution for this problem was to use an oversampling technique to balance the classes, using sklearn. Utils resample, we managed to balance it and solve this problem. We defiantly wanted to experiment with more deep-learning algorithms, but we needed more computer power to process them.

Future work:

We are interested in experimenting with more deep-learning algorithms, such as anomaly detection. Also, building a proccing image algorithm to predict a brain stroke from patterns in the brain scan images is an exciting experiment. Trying to predict more diseases as well.

Reference:

1. **Thomas (2019). Ultra-modern medicine: Examples of machine learning in healthcare. [online] Built In. Available at: https://builtin.com/artificial-intelligence/machine-learning-healthcare.**

Sirsat, M. S., Fermé, E., & Câmara, J. (2020). Machine learning for brain stroke: a review. *Journal of Stroke and Cerebrovascular Diseases*, *29*(10), 105162.

2.

3. https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

4 . Asplund, K., Karvanen, J., Giampaoli, S., Jousilahti, P., Niemelä, M., Broda, G., ... & Kulathinal, S. (2009). Relative risks for stroke by age, sex, and population based on follow-up of 18 European populations in the MORGAN Project. *Stroke*, *40*(7), 2319-2326.

5 . Tazin, T., Alam, M. N., Dola, N. N., Bari, M. S., Bourouis, S., & Monirujjaman Khan, M. (2021). Stroke Disease Detection and Prediction Using Robust Learning Approaches. *Journal of Healthcare Engineering*, *2021*.

6 . Vemmos, K., Ntaios, G., Spengos, K., Savvari, P., Vemmou, A., Pappa, T., ... & Alevizaki, M. (2011). Association between obesity and mortality after acute first-ever stroke: the obesity–stroke paradox. *Stroke*, *42*(1), 30-36.

7. Laurent, S., & Boutouyrie, P. (2005). Arterial stiffness and stroke in hypertension. *CNS drugs*, *19*(1), 1-11.

8. Bechthold, A., Boeing, H., Schwedhelm, C., Hoffmann, G., Knüppel, S., Iqbal, K., ... & Schwingshackl, L. (2019). Food groups and risk of coronary heart disease, stroke, and heart failure: a systematic review and dose-response meta-analysis of prospective studies. *Critical reviews in food science and nutrition*, *59*(7), 1071-1090.

9. Freedman, B., Potpara, T. S., & Lip, G. Y. (2016). Stroke prevention in atrial fibrillation. *The Lancet*, *388*(10046), 806-817.

10. Hill, M. D. (2014). Stroke and diabetes mellitus. *Handbook of clinical neurology*, *126*, 167-174.

11. Kuźma, Elżbieta, et al. "Stroke, and dementia risk: a systematic review and meta-analysis." *Alzheimer's & Dementia* 14.11 (2018): 1416-1426.

12. Thomalla, G., Simonsen, C. Z., Boutitie, F., Andersen, G., Berthezene, Y., Cheng, B., ... & Gerloff, C. (2018). MRI-guided thrombolysis for stroke with unknown time of onset. *New England Journal of Medicine*, *379*(7), 611-622.

13. Lees, K. R., & Walters, M. R. (2005). Acute stroke and diabetes. *Cerebrovascular Diseases*, *20*(Suppl. 1), 9-14.

14. Wolfe, C. D. (2000). The impact of stroke. *British medical bulletin*, *56*(2), 275-286.

15 . BENTLEY, Paul, et al. Prediction of stroke thrombolysis outcome using CT brain machine learning. *NeuroImage: Clinical*, 2014, 4: 635-640.

16 . RAHMAN, Senjuti; HASAN, Mehedi; SARKAR, Ajay Krishno. Prediction of Brain Stroke using Machine Learning Algorithms and Deep Neural Network Techniques. *European Journal of Electrical Engineering and Computer Science*, 2023, 7.1: 23-30.

17 . GUHDAR, Mohammed; MELHUM, Amera Ismail; IBRAHIM, Alaa Luqman. Optimizing Accuracy of Stroke Prediction Using Logistic Regression. *Journal of Technology and Informatics (JoTI)*, 2023, 4.2: 41-47.

18 . OZKARA, Burak B., et al. Prediction of Functional Outcome in Stroke Patients with Proximal Middle Cerebral Artery Occlusions Using Machine Learning Models. *Journal of Clinical Medicine*, 2023, 12.3: 839.

19 . ZAFEIROPOULOS, Nikolaos, et al. Interpretable Stroke Risk Prediction Using Machine Learning Algorithms. In: *Intelligent Sustainable Systems: Selected Papers of WorldS4 2022, Volume 2*. Singapore: Springer Nature Singapore, 2023. p. 647-656.

20 . DIKER, Aykut; ELEN, Abdullah; SUBASI, Abdulhamit. Brain stroke detection from computed tomography images using deep learning algorithms. In: *Applications of Artificial Intelligence in Medical Imaging*. Academic Press, 2023. p. 207-222.