**CSE422: Artificial Intelligence**

**Project report**

****

**GROUP:5**

**Project Name :** Stroke Prediction

**Submitted by- Submitted to-**

*Afif Alamgir (20301199) Afia Fairoose Abedin*

*Nabiha Tasnim Orchi(20301148) Sumaiya Akhter*

*Jennifer Abedin (20301219)*

*Ishrat Tasnim Awishi(20301031)*

**Introduction :**

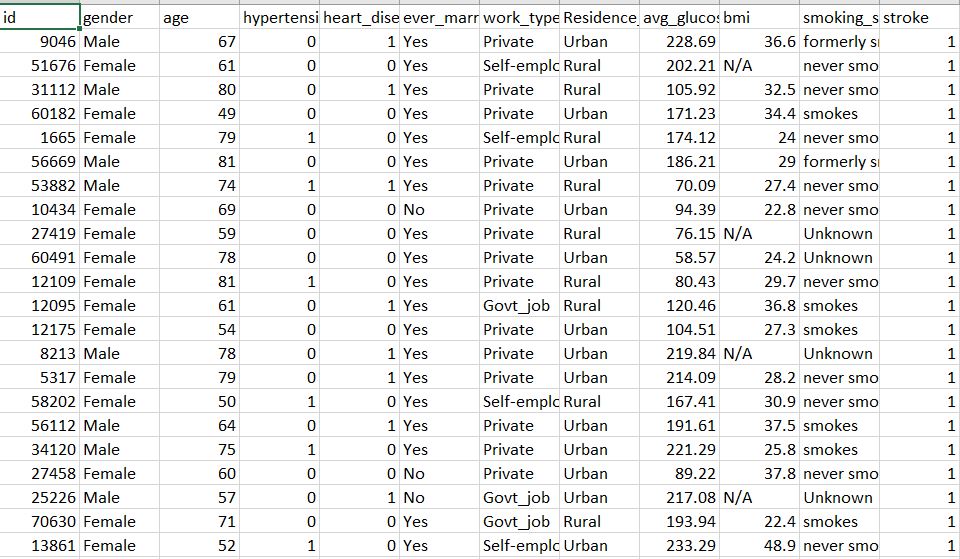
Our brain is a very sophisticated organ that regulates many bodily processes. When anything prevents blood flow to a portion of the brain or when a blood artery in the brain bursts, a stroke, also known as a brain attack, happens. Affected or dying brain tissue is present in both scenarios. The section of the body that regulates a certain function won't operate as it should if a stroke happens and blood flow can't get to that area. A stroke can result in long-term impairment, permanent brain damage, or even death. An emergency situation involving a stroke requires quick medical attention. Early intervention can lessen problems and brain damage. Stroke is the second biggest cause of death worldwide, accounting for around 11% of all fatalities, according to the World Health Organization (WHO).

Based on input characteristics like gender, age, numerous diseases, and smoking status, this dataset is used to determine whether a patient is likely to get a stroke. The data's rows each provide pertinent information about the patient.

**METHODOLOGY:**

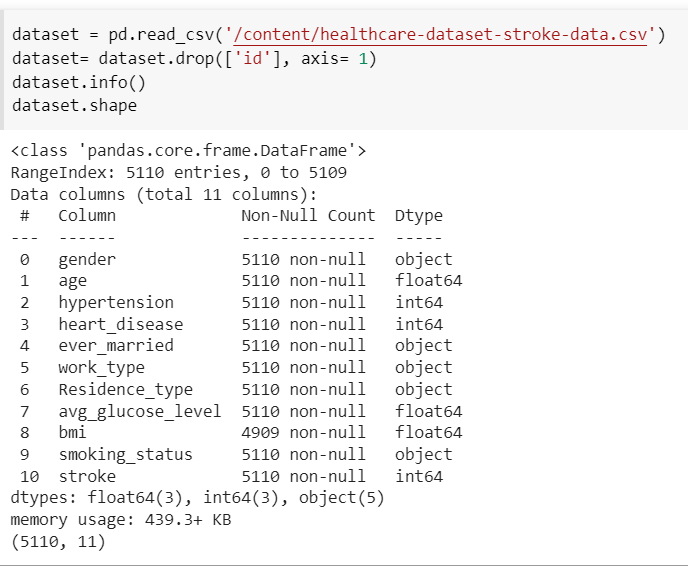
**Dataset Description:**

Our data set was designed to predict stroke based on 5110 patient entries with 11 columns of data they had provided, including "gender, age, hypertension, marital status, work type, residence type, average glucose level, body mass index, and smoking status and stroke chances." Here if we fit the data with ***label*** and ***feature*** according to the information, here stroke column(L) will be counted as ***label*** and Column B to K will be considered as ***feature***.

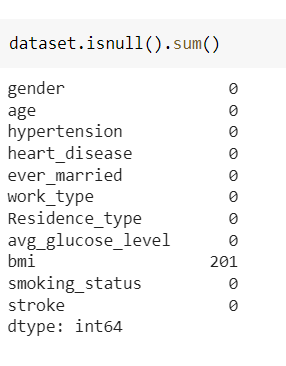


**Preprocessing Techniques Applied:**

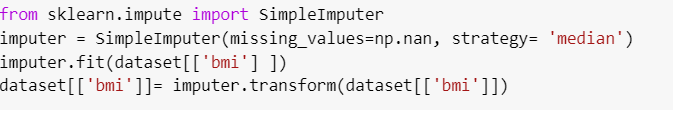
Following an analysis of the discontinued relationship between the label and the dataset using the pandas library, we removed the column id after uploading the dataset.



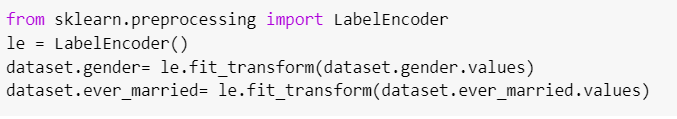
After shaping the dataset finding 5110 entries with 11 columns, we checked the column-by-column data to identify the null numbers.



By determining the median of the total BMI values of the existing entries, the null values were substituted using the Simple Imputing technique from the Sklearn Library.



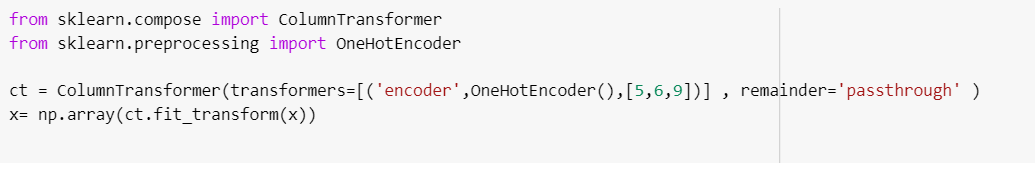
We fitted the data from columns B and F, which include binary values, after preprocessing the data with null values and using the Label Encoder function from the SKlearn library.



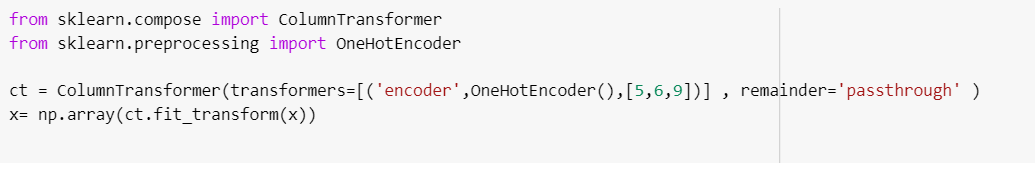
Features and labels are divided into two new data sets.



Because columns 5, 6, and 9 include categorical data, OneHotEncoder from the Sklearn library is used to preprocess these columns.



The test set size was 30% of the total items in the train set, and this was the final step in the data preprocessing procedure.



**Model Description:**

Four models were performed on 70% of total entries.

**KNN Model:**

● One of the simplest Machine Learning algorithms based on the Supervised Learning technique is K-Nearest Neighbour..

● K-Nearest Neighbour algorithm assumes that the new case's data and existing cases are similar, and it places the new case in the category that is most like the existing categories.

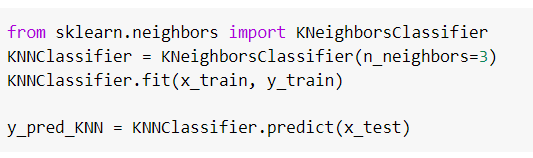
● A new data point is classified using the K-NN algorithm based on similarity after all the existing data has been stored. This means that utilizing the K-NN method, fresh data can be quickly and accurately sorted into a suitable category.

● Although the K-NN approach is most frequently used for classification problems, it can also be utilized for regression. the K-NN approach is most frequently used for classification problems, it can also be utilized for regression.

● Since K-NN is a non-parametric technique, it makes no assumptions about the underlying data.

● It is also known as a lazy learner algorithm since it saves the training dataset rather than learning from it immediately. Instead, it uses the dataset to perform an action when classifying data.

● KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.

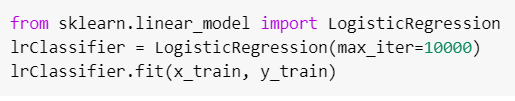


**Logistic Regression:**

A statistical analysis method called logistic regression uses previous observations from a data set to predict a binary outcome, such as yes or no.

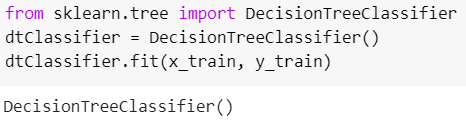
● By examining the correlation between one or more already present independent variables, a logistic regression model forecasts a dependent data variable. A logistic regression could be used, for instance, to forecast whether a candidate for office will win or lose, or if a high school student will be accepted into a particular institution or not. These simple choices between two options allow for binary outcomes.

● Multiple criteria for input can be taken into account using a logistic regression model. The logistic function can take into account the student's grade point average, SAT score, and number of extracurricular activities in the case of college acceptance. It then rates new cases according to their likelihood of falling into one of two outcome categories based on historical information about past outcomes involving the same input criteria.



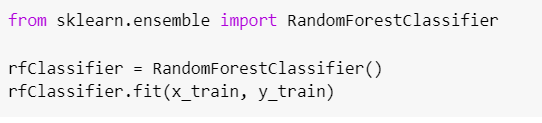
**Decision Tree:**

Decision tree is a flowchart-like structure in which each internal node represents a "test" on an attribute (e.g., whether a coin flip comes up heads or tails), each branch represents the outcome of the test, and each leaf node represents a class label (decision taken after computing all attributes). Classification rules are represented by the routes from root to leaf. The expected values (or expected utility) of competing alternatives are calculated using a decision tree and the closely related influence diagram as a visual and analytical decision assistance tool in decision analysis.



**Random Forest:**

There are three key hyperparameters for random forest algorithms that must be set prior to training. Node size, tree count, and sampled feature count are a few of them. The random forest classifier can then be utilized to address classification or regression issues. The decision trees that make up the ensemble of the random forest algorithm each contain a bootstrap sample of data that is taken from a training set with replacement. One-third of the training sample—also referred to as the out-of-bag (oob) sample—is set aside as test data, we'll return to this sample later. The dataset is subsequently given a second randomization injection by feature bagging, increasing dataset diversity and decreasing decision tree correlation. The prediction will be determined differently depending on the type of issue. The individual decision trees will be averaged for the regression job, and for the classification task, the predicted class will be determined by a majority vote, or the most common categorical variable. The prediction is then finalized by cross-validation using the oob sample.



**Results:**

| **Model Name** | **Accuracy** |
| --- | --- |
| KNN Model | **93.73776908023484** |
| Logistic Regression Model | **95.04240052185257** |
| Decision Tree Model | **91.65035877364645** |
| Random forest | **94.97716894977168** |

**Reference:**

*Data Collection form:*<https://docs.google.com/forms/d/1vLk3z73T3aLIkYZBavuHobAICCaQzVoiAYcr8PPm-2E/edit#responses>

*Dataset:* <https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset>

<https://blog.exsilio.com/all/accuracy-precision-recall-f1-score-interpretation-of-performance-measures/>

<https://www.ritchieng.com/machine-learning-k-nearest-neighbors-knn/>

*Articles:* javapoint.com