JSS Mahavidyapeetha

JSS Science and Technology University, Mysuru 570006

DEPARTMENT OF COMPUTER APPLICATIONS

STROKE ANALSIS AND PREDICTION

Master of Computer Application

by

Name	***USN***
ASHOK KUMAR H G	01JST20PMC006
ANAND REDDY N	01JST20PMC003

Submitted to,

Dr. S K Niranjan

Professor,

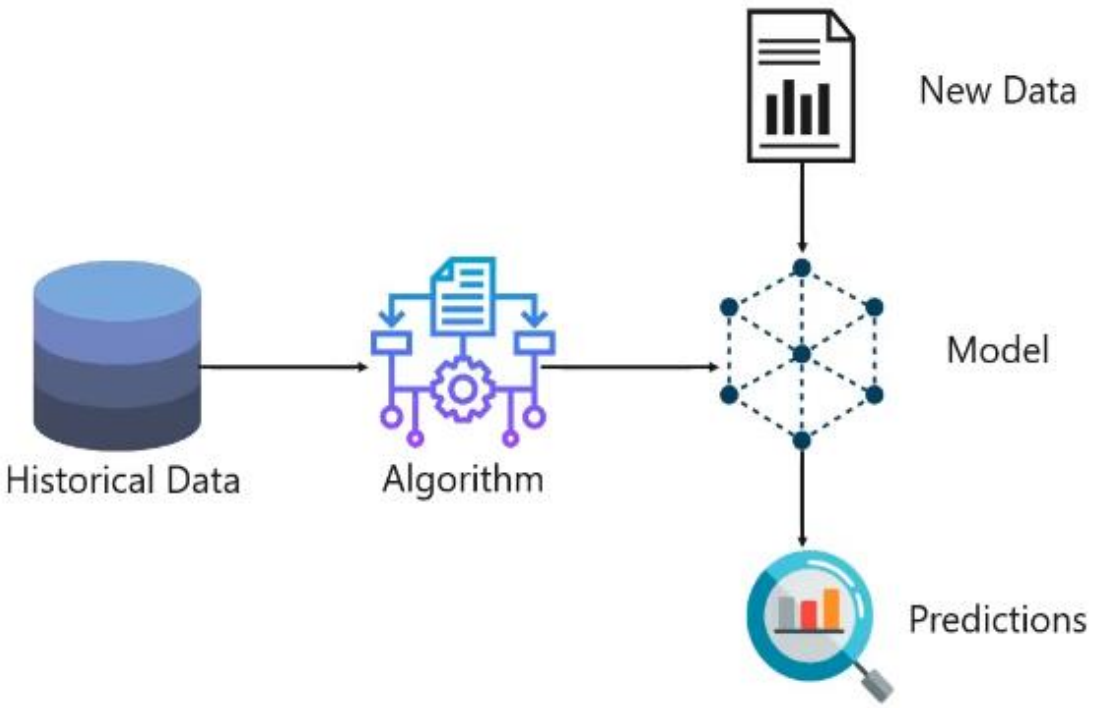
Department of Computer Application

JSSSTU, Mysore

Introduction:

The objective of this activity is to develop a preliminary screening tool which can be used to identify the likelihood of an individual having a stroke based on general contributing attributes. A Dataset from **Kaggle** was used as the basis for the analysis model. Cerebrovascular accidents (strokes) in 2020 were the 5th leading cause of death in the Country. A stroke occurs when the blood supply to a region of the brain is suddenly blocked or when a rupture occurs starving the brain cells of oxygen and nutrients. Blockage obstructing the flow of blood to a region of the brain is called an ischemic stroke and accounts for 87% of all strokes. The rupturing of a blood vessel is called a hemorrhagic stroke and accounts for 13% of all strokes.

The dataset used for this analysis did not identify the type of stroke for each respective individual. To stay consistent with the dataset, the general word stroke will be used to describe the occurrence being predicted. A third category of stroke called a transient ischemic attack (TIA), or "mini stroke", caused by a temporary clot can also occur. The TIA has contributing factors similar to those of the ischemic and hemorrhagic stroke and is included in the general term stroke when identifying a potential outcome.



Requirement Analysis:

- Data Selection:

  A dataset from Kaggle was selected for the machine learning process. The data was reviewed to identify trends and cleanup requirements. The primary data cleanup activities identified were to address "N/A" values associated with body mass index and "Unknown" smoker status. The "N/A" values represented 3.9% of the dataset and were addressed by using the mean body mass index and assigning that to the "N/A" values. The "Unknown" smoker status represented 30.4% of the dataset. Literature review verified that "Unknown" values were considered an accepted data point and therefore the "Unknown" values were left as presented in the raw data.

- Data Source:

  The attributes with the dataset are:

  - id: a unique identifier for each set of information
  - gender: "Male, "Female", "Other"
  - age: age of the patient
  - hypertension: 0 assigned if hypertension not present, 1 if patient has hypertension
  - heart_disease: 0 assigned if heart disease not present, 1 if patient has heart disease
  - ever_married: "No" or "Yes"
  - work_type: "children", "Govt_job", "Never_worked", "Private", or "Self-employed"
  - Residence_type: "Rural" or "Urban"
  - avg_glucose_level: average glucose level in blood
  - bmi: body mass index
  - smoking_status: "formerly smoked", "never smoked", "smokes", or "Unknown"
  - stroke: 0 if patient has not had a stroke, 1 if patient has had a stroke

- Data Review:

The raw dataset for machine learning consists of 5110 unique rows. Each row contains patient information designated by a unique id. There were 2,994 (58.60%) “Females”, 2,115 (41.40%) “Males” and 1 “Other” in the gender attribute. The “Other” gender was dropped from the dataset for a resulting dataset of 5,109 unique rows.

Data Review				
Data Attribute	Female		Male	
	Count	Percent of gender	Count	Percent of gender
Had a stroke (Y)	141	4.7 %	108	5.1 %
Considered diabetic risk	230	7.7 %	204	9.6 %
Have heart disease (Y)	113	3.8 %	163	7.7 %
Have hypertension (Y)	276	9.2 %	222	10.5 %
Considered obese	1,115	37.2 %	805	38.1 %
Married (Y)	2,001	66.8 %	1,352	63.9 %
Live in Urban areas (Y)	1,529	51.1 %	1,067	50.4 %
Never smoked	1,229	41.0 %	663	31.3 %
Formerly smoked	477	15.9 %	407	19.2 %
Currently smoke	452	15.1 %	337	15.9 %
Unknown smoking status	836	27.9 %	708	33.5 %
Age: 0-19	480	16.0 %	486	22.9 %
Age: 20-39	791	26.4 %	412	19.4 %
Age: 40-49	450	15.0 %	280	13.2 %
Age: 50-59	472	15.7 %	362	17.1 %
Age: 60-69	352	11.8 %	269	12.7 %
Age: 70-79	336	11.2 %	233	11.0 %
Age: 80+	113	3.8 %	73	3.4 %

Preparation:

Data Preparation:

Review of the data identified most of the Yes/No type answers for personal health questions, including if the person had a stroke, were heavily biased to the “No” side. To ensure an effective model learning process, Synthetic Minority Oversampling Technique (SMOTE) was used to synthetically balance the Yes/No results for stroke. SMOTE selects

samples in the minority class that are close and then draws lines between them. New sample points are located on these lines. Other data preparation steps included One-Hot Encoding.

Data operations:

- Data pre –procesing:

  The data found in the different repositories may not all be clean, or may contain errors. A lack of cleaning and treatment in the data reduces the quality of the analysis by generating useless rules in the data mining stage. The problems found in the data repositories considered in this study include incomplete data that are missing attribute values, inconsistent data and even discrepancies between the data. The preprocessing of data is achieved with filters that can be applied in a supervised and unsupervised manner. In both cases, there is the option of cleaning the attribute or instance; this depends on the type of data that the analysis needs. The advantage of prior data preparation is that a smaller data set is generated, improving the efficiency of the data analysis process, especially at the mining application stage.

- Data Cleaning and Imputation:

  Data cleaning was conducted in Jupyter Notebook using Python. As previously noted, the “Other” gender category was dropped from the dataset, resulting in removing 1 row of data. In reviewing the raw data, the bmi attribute was identified as having 201 “N/A” values. This represents 3.9% of the dataset. The mean bmi value of 28.89 was used as the replacement value for the “N/A” values.

  As noted above in the representation data tables, the raw dataset has a total 1,544 “Unknown” smoking status values representing 30.4% of the dataset. A closer look at the data showed 32% of the “Unknown” values were between the ages of 0-10 and 41% was between the ages of 0-15. The Centers for Disease Control and Prevention (CDC) defines a current smoker [9] as an Adult who has previously smoked 100 cigarettes in their lifetime and who currently smokes. Based on the CDC definition and the high percentage of “Unkown” values in the age range 0-10, it was originally discussed to replace those values with “never smoked”. Additional research of online literature to address this issue of “Unknown” labels was conducted and it was found that

"Unknown" is an accepted category. The final decision was made to leave the "Unknown" smoker status values as presented in the raw dataset.

One-Hot Encoding was used for categorical data work_type and smoking status to be used in the linear and tree models as shown below.

- Transformation:

Data transformation consists of creating new attributes from the original attributes. Transformation also includes global transformations, which involve the exchange of rows with columns, where the data is transformed through a certain type of selection. The selection of data can opt for a vertical selection, which acts directly on the attributes of the object of analysis. Horizontal selection then acts on the data instances and, finally, a sample of the population can be used to perform the transformation. These instances are followed by two possible actions. In the first action, several interviews are applied to detect the origin of the data. In the interview, the administrator of the database is responsible for disclosing the age of the records that are online. The second action allows us to know the structure of the tables and the fields that appear in the entity-relationship diagram of each database included in the investigation.

Hypothesis:

A reliable predictive analysis model can be developed if the data and stroke key attributes are correctly identified and prepared for the machine learning process. The importance of features generated by the model selected will be compared against the stroke risk factors identified by the American Stroke Association. If the attributes are correctly identified by the model, the hypothesis will be considered validated.

Basis Risk Factors from American Stroke Association common to the dataset:

- High Blood Pressure
- Smoking
- Diabetes
- Obesity
- Age (cannot be controlled)
- Gender (cannot be controlled)

Actionable Items:

This model is one of many tools which are needed to increase awareness and help reduce stroke incidents. As the noted above, the American Stroke Association states that 80% of strokes are preventable.

Actionable item

- Support stroke prevention awareness programs
  - Exercise
  - Eating correctly
  - Programs to stop smoking

Future Work:

Periodic review and update of the model would be beneficial in creating a more successful tool.

References:

[1] Stroke Prediction Dataset, 11 clinical features por predicting stroke events, https://www.kaggle.com/fedesoriano/stroke-prediction-dataset

[2] American Stroke Association

https://www.stroke.org/en/about-stroke/types-of-stroke/ischemic-stroke-clots

[3] Centers for Disease Control and Prevention, National Center for Health Statistics, https://www.cdc.gov/nchs/nhis/tobacco/tobacco_glossary.htm

[4] American Stroke Association.

https://www.stroke.org/en/about-stroke