DSC 478 Project Proposal

Project Type (Healthcare Analysis- Stroke Prediction)

DSC 478

Jacob Crawford, Jaimi Patel, Seif Abuhashish

**Overview:**

Stroke is the second driving reason for death internationally, accounting for around 11% of absolute passouts. This dataset is utilized to foresee whether a patient is going to suffer a stroke given the information like gender, age, various diseases, and smoking status. Each column in the information gives relevant data about the patient. The idea is to analyze the stroke data and predict the possibility of having a stroke. Our data set contains 12 attributes and 5110 data points. We plan to preprocess, organize, and clean the set to derive knowledge from our set with descriptive statistics and vizulations. After gaining a better understanding of our data we build multiple classification models that predict our dependent variable. The models are compared and the best fitting model is chosen for analysis and interpretation.

**Analysis Approach:**

DM Task 1: Pre-processing

* Load, format and visualize the dataset with the correct data points and variables into our python script. Process the dataset for inconsistencies, incomplete values, missing values, and noise.
* We will explore the different variables to get a better understanding of what they represent. We also explore the independent variables’ distribution; what type of feature it is (categorical or numeric)? What is the distribution of the values in each variable (descriptive statistics) ? Is the data skewed or imbalanced, and what does that signify? Can we identify outliers and noisy data? Do any variables seem to correlate to one another and why?
* We will also make determinations if some columns are not useful to our purpose. If the feature is not useful we will eliminate it.

DM Task 2: Exploratory phase

* After preprocessing we begin to explore our data set by creating various graphs, charts, and matrices to help visualize and draw conclusions. We start gathering descriptive statistics on the variables and creating visualizations to explain and understand the nature of this data. Bar graphs help explain and compare the categorical variables against one another, box plots provide a five-number summary and a visualization of spread which can help detect outliers. Histograms help us visualize the frequency of continuous numeric data such as age, average glucose levels and BMI. Correlation matrices help us organize, understand, and visualize the similarities between the independent variables. This helps in identifying multicollinearity. When variables are similar we eliminate them due to unnecessary redundancy in the set. Exploring techniques such as PCA and clustering can help reduce the number of features and smooth out noisy data while eliminating multicollinearity, and reducing the number of features. Dimensionality reduction can help interoperability and computational efficiency.
* Data normalization of the numeric attributes to reduce data redundancy and improve data integrity. All of these steps give us a richer familiarity of our data and give us ideas of what to prioritize and how to proceed.

DM Task 3: Supervised Learning

* Predictive models can be built to classify our dependent variable (stroke) and then compared against other models built. A classification model such as logistic regression model, SVM, or Random Forest can be built to identify the class of our dependent variable.

**Data Schema and Size:**

Healthcare dataset about strokes.

The Stroke dataset contains 12 attributes and 5110 observations. Stroke is the binary dependent variable. The rest of the data set is composed of 11 independent variables, five of those are categorical (gender, ever\_married, work\_type, Residence\_type, smoking\_status), three are continuous variables (age, BMI, avg\_glucose\_level), and two are binary variables (hypertension, heart\_disease).

The attributes are 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

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

**Plan for Evaluation**:

We will evaluate the models against one another using classification charts and confusion matrices to compare F1 scores, accuracy, error rate, precision, and ROC. The best model is then selected from the group to explain the significant findings. The final model is then translated from a technical view into practical insight for the average person. Model interpretation is very important because it is essential in bridging practical results from technical professionals to a non technical audience.

**Plan Work Distribution:** Jacob Crawford, Jaimi Patel, Seif Abuhashish

We will organize our efforts using discord chat, discord telephone and email. We will collaborate on how to specifically clean our data and agree to a common dataset that we will all interrogate. We plan to explore our three data mining tasks individually and then compare our findings and observations as a group.