Project number: *B10*

Group members: *Yusif Israfilbayov*

Peter Olusegun Aina

Repository Link: *https://github.com/Yusif99/IDS_2021_HeartStroke*

Heart Stroke Prediction

BUSINESS UNDERSTANDING

Heart stroke is one of the leading cause of deaths in the world. If happened, it requires significant attention and care to recover from. Apart the emotional distress caused to the patients and their circle of people, it also results in huge amount of expenses both for the patients and the healthcare system. We believe that informing the people who has high chance of stroke would save everyone from large efforts and expenses.

Our aim is to improve the mentioned conditions through application of machine learning algorithms for forecasting the chances of stroke in order to take precautions on time. Family practitioners can keep track of some life habits and physiological measurements in order to be assessed by the built model. People under risk can be alarmed in order to take serious steps to decrease their chance of having stroke.

Definition of success would be achievement a significant decrease by means of costs, number of stroke cases and amount of time spent for the cases. This model can benefit national healthcare system and people under risk by cutting off the related expenses. Besides, health insurance companies can also use this project to fine tune their business risks and costs using this prediction tool.

Our project is a mini version of what can be applied in real life. The data needed is to be gathered by family practitioners or any other convenient methods by two to three months interval. A person must be in charge of registering the gathered data to the system. Although everything looks on track, but ethical issues must also be tackled in the project. This includes registration of people's life habits and decisions, such as smoking, work style etc. The participants must be informed beforehand about the usage and privacy of their personal information. To ensure the smooth progress, there must be diligent effort put into gathering the data from the people considering they might question the ethical side of the things.

Our project uses minimum amounts of resources for gathering the data which saves on costs. The equipment required is quite general. abundant and inexpensive in hospitals. The quality of assessment in our model is focuses on accuracy and false negative rate. Accuracy is important for the concerning organization, whereas false negative rate is important for not losing people's trust in the project. If a person agrees to share their information with the organization, although is not informed by means of his risk, this may result in loss of reliability towards the project. Therefore, success criteria also includes the minimization of false negative rate.

DATA UNDERSTANDING

Our project is a mini version of what can be applied in a larger scale, and for this purpose we have gathered the data from already available sources. We need categorical and numeric data related to life habits and some general physiological tests. This data is easily achievable, and already available. After the first model building step is completed, extra data can be gathered as new inputs emerge. We retrieve an example dataset from Kaggle platform.

Our data has eleven features and two of them are IDs of the patients and labels as if the patient had a stroke or not. Others include: gender, age, hypertension, heart disease, ever married, work type, residence type, average blood glucose level, body mass index and smoking status. Hypertension is high blood pressure and the feature registers whether the patient has this condition. Heart diseases shows whether the person has any heart condition or not. Work type includes private, government job or self-employed. Residence describes where the person lives; urban or rural. Average blood glucose level is a continuous feature that accounts for average amount of glucose measured in the person's blood through biochemical tests. Body mass index is mathematically calculated from person's height and weight, which is also a continuous variable. Under the smoking status, we have never smoked, formerly smoked or smokes.

When we further explore the data, we can find missing values for many entries. The missing values are under body mass index (bmi) variable. Another problem with the dataset is that a large portion of smoking status is described as unknown. We must find a way to work through these problems. Our dataset is not huge enough to drop all these entries, therefore we will have to assign some values to them. Selected method for value assigning is linear regression for body mass index and class prediction for smoking status values using other variables. Validity of our data can be checked through some attributes, i.e., whether the samples reflect the population. Average age in our data is 43 and average bmi appears to be 29. This is very close to the average bmi (30) in that age range, in the USA. This holds for the

average blood glucose level. More than one model will be used to choose the best performing one with different parameters.

At this point our data is ready to be fit into possible models. Several algorithms will be used in order to choose the best working one.

PROJECT PLAN

1) Data retrieval (2 h)

choosing and downloading the right data

2) Data exploration (3 h)

Researching and understanding the data features

Understanding the data types of the features

3) Data pre-processing (5 h)

Checking for missing values

Researching about how to treat missing values

Method selection and assignment of missing values

4) Statistical analysis of data features (5 h)

Researching about possible statistical analysis methods

Correlation test between various features of the data

Distribution of various data features

5) Model selection (8 h)

Researching about different types of models

Testing different models by means of accuracy and false positive rates

6) Model building and evaluation (5 h)

Hyperparameter tuning

Model evaluation

7) Presentation (3 h):

_ *Preparation of presentation deliverables*

TOOLS AND METHODS:

- Kaggle
- Jupyter notebook
- Microsoft Excel
- Microsoft Word

- Sci-kit Learn
- Linear Regression
- Random Forest Classifier
- K Nearest Neighbours Classifier
- Support Vector Classifier
- Pearson correlation analysis