BELLEVUE UNIVERSITY

DSC630-T301 Predictive Analytics (2223-1) – Project Milestone2

Project Proposal on predicting a likelihood of stroke based on risk factors

Veera Reddy Koppula

10/12/2021

Table of Contents

1. Introduction

1.1 Background

A stroke occurs when the blood supply to part of the brain is suddenly interrupted or when a blood vessel in the brain bursts, spilling blood into the spaces surrounding brain cells. Brain cells die when they no longer receive oxygen and nutrients from the blood or there is sudden bleeding into or around the brain. The symptoms of a stroke include sudden numbness or weakness, especially on one side of the body; sudden confusion or trouble speaking or understanding speech; sudden trouble seeing in one or both eyes; sudden trouble with walking, dizziness, or loss of balance or coordination; or sudden severe headache with no known cause. There are two forms of stroke: ischemic - blockage of a blood vessel supplying the brain, and hemorrhagic - bleeding into or around the brain.

According to CDC website [1], here are some of the facts about stroke:

- In 2018, **1 in every 6 deaths** from cardiovascular disease was due to stroke.
- Someone in the United States has a stroke every **40 seconds**. Every **4 minutes**, someone dies of stroke.
- Every year, more than **795,000 people** in the United States have a stroke. About 610,000 of these are first or new strokes.
- About 185,000 strokes—**nearly 1 of 4**—are in people who have had a previous stroke.
- About **87%** of all strokes are ischemic strokes, in which blood flow to the brain is blocked.

- Stroke-related costs in the United States came to nearly **$46 billion** between 2014 and 2015.2 This total includes the cost of health care services, medicines to treat stroke, and missed days of work.
- Stroke is a leading cause of serious long-term disability.2 Stroke reduces mobility in more than half of stroke survivors aged 65 and over.

Stroke Statistics by Race and Ethnicity

- Stroke is a leading cause of death for Americans, but the risk of having a stroke varies with race and ethnicity.
- Risk of having a first stroke is **nearly twice** as high for blacks as for whites, and blacks have the highest rate of death due to stroke.
- Though stroke death rates have declined for decades among all race/ethnicities, Hispanics have seen an increase in death rates since 2013.

Stroke Risk Varies by Age

- Stroke risk increases with age, but strokes can—and do—occur at any age.
- In 2009, **34%** of people hospitalized for stroke were **less than 65 years old**.

Signs of Stroke in Men and Women

- Sudden **numbness** or weakness in the face, arm, or leg, especially on one side of the body.
- Sudden **confusion**, trouble speaking, or difficulty understanding speech.
- Sudden **trouble seeing** in one or both eyes.
- Sudden **trouble walking**, dizziness, loss of balance, or lack of coordination.
- Sudden **severe headache** with no known cause.

Anyone can have a stroke at any age. But certain things can increase chances of having a stroke. The best way to protect someone from a stroke is to understand risk of stroke and how to control it.

Some individual factors, medical problems, lifestyle decisions that can increase one's risk of stroke include:

- Age
- Sex
- Race or Ethnicity
- High Blood Pressure/Hypertension
- High Cholesterol
- Heart Disease
- Diabetes
- Unhealthy diet/Physical inactivity
- Obesity
- Tobacco use
- Alcohol Use

1.2 Problem Statement

Although stroke is a disease of the brain, it can affect the entire body. A common disability that results from stroke is complete paralysis on one side of the body, called hemiplegia. A related disability that is not as debilitating as paralysis is one-sided weakness or hemiparesis. Stroke may cause problems with thinking, awareness, attention, learning, judgment, and memory. Stroke survivors often have problems understanding or forming speech. A stroke can lead to emotional problems. Stroke patients may have difficulty controlling their emotions or may express inappropriate emotions. Many stroke patients experience depression. Stroke survivors may also have

numbness or strange sensations. The pain is often worse in the hands and feet and is made worse by movement and temperature changes, especially cold temperatures.

Recurrent stroke is frequent; about 25 percent of people who recover from their first stroke will have another stroke within 5 years.

For several years, methodologies for an early prediction of likelihood of stroke has been pursued, and several predictions techniques have been utilized to support health care professionals in predicting likelihood of stroke, using known risk factors. Many studies may be conducted on prospective patients to minimize the impact of developing such a disease, and accurate approaches to predict a stoke, such as the methods suggested in this paper, may be critical for saving lives.

For this course project, I will be using some of the prediction modelling techniques to answer whether the person is likely to have a stroke or not based on some of the known risk factors for stroke.

1.3 Scope

The main objective of this paper is to design a robust system which works efficiently and will be able to predict the possibility of stroke accurately. This paper uses the dataset available in Kaggle with 5110 observations. Here are the attributes of the dataset explained:

Attribute Information

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_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

*Note: "Unknown" in smoking_status means that the information is unavailable for this patient

1.4 Document Overview

This document is the initial project proposal, and it will be kept updated constantly as the project progresses and this will serve as a guideline for the project execution.

2. Requirements Development

2.1 Technical Approach

This course project on predictive analytics on heart failure will follow CRISP-DM model for data understanding, modeling, and evaluation. Python programming language will be used to load, explore, and model the prediction system along with necessary machine learning libraries for Python.

2.2 Data Sources

The required dataset for this project will be retrieved from Kaggle website. Here's the link to the data source [2]: https://www.kaggle.com/fedesoriano/stroke-prediction-dataset

2.3 Analysis

Initially, exploratory data analysis will be performed to better understand the data and compute some major descriptive statistics. I will produce some descriptive visualizations representing the dataset as well.

2.4 Model Deployment

For the prediction model, various machine learning algorithms will be employed to see and choose the best resulting model. Particularly, Logistic Regression, Support Vector Machine (SVM), Decision Tree, K Nearest Neighbor (KNN), Naive Bayes, Random Forest Classifier are the potential algorithms to be used.

2.5 Testing and Evaluation

After carefully evaluating the results, I will be using the most accurate train test model to predict heart-related hospitalizations based on the available patient-specific medical history.

3. Expected Results

I expect to curate an effective, accurate and working prediction model which can predict the possibility of occurrence of a stroke.

4. Execution and Management

4.1 Project Plan

This project will follow the timeline below:

Milestone 1 (due Week 1 on 12/5): Done

- Create project plan and begin looking for data for final project.
- Create folder for project in Teams.
- Create timeline for final project milestones

Milestone 2 (due Week 2 on 12/12):

- Create final project outline for idea to be used for other fellow students to review.
- Write up project proposal.
- Work on finalizing data choice.
- Start initial analysis for final project and make sure data will work for the proposal idea.
- Find additional data if need when reviewing initial analysis.
- Finish up the initial analysis for peer review.
- Complete peer review for other independent projects.

Milestone 3(due Week 5 on 1/16):

- Complete further analysis of data and any added data.
- Begin working on the final paper and writing up analysis results and documenting challenges.
- Use this as a rough draft start for summarizing findings.
- Have a final draft report done
- Peer Review

Milestone 4(due Week 9 on 2/13):

- Begin work on the final slide presentation.
- Start to proof rough draft.
- Finish slide show
- Peer Review

Milestone 5(due Week 12 on 3/5):

- Finalize Report
- Record my presentation
- Submit final works
- Peer Review

4.2 Project Risks

This study assumes that the dataset is valid and accurate. There is always a chance of losing some data due to its incompleteness and extremeness.

One of the potential risks associated with this study is low number of observations. Also, this dataset does not account all risk factors associated with stroke for example diet, physical activity, diabetes, and alcohol consumption etc. so the prediction model cannot be fully accurate.

5. References

1. https://www.cdc.gov/stroke/about.htm
2. https://www.kaggle.com/fedesoriano/stroke-prediction-dataset