

Stroke Prediction Dataset: Capstone Two Project Proposal

In the field of healthcare, much emphasis is focused on secondary and tertiary prevention, where a patient has already been diagnosed and treatment is focused on preventing the disease from getting worse and reducing its symptoms. However, if we shift our perspective to making primary prevention a priority in which we prevent the disease or incident from occurring in the first place, this would greatly improve a patient's quality of life and reduce treatment costs. Rather than treating a stroke, which is extremely pricey if we factor in the cost of hospitalization, medications, and rehab treatments, a more cost effective solution would be to identify patients that are at high risk of getting a stroke and incorporating lifestyle modifications to lower their risk of getting a stroke.

In this project, I work with a stroke prediction dataset to implement a machine learning model that can predict with over 90% accuracy whether a patient is likely to have a stroke based on parameters like gender, age, presence of disease, BMI, glucose level, and smoking status. Using attributes that are readily accessible, data for this model can be easily collected at doctors' visits and quickly return results without requiring invasive testing or waiting for lab work turnaround time. This model would be useful for health care professionals to efficiently identify patients that are at high risk of getting a stroke and determine which parameter contributes to the high risk. Using this information, the patients can be counseled on how to lower their risk, either through lifestyle modifications or medications. By identifying risk factors early on, primary prevention could greatly reduce the cost of treatment for a stroke.

This dataset can be accessed through Kaggle and has a clear machine learning idea--to predict whether a person is likely to have a stroke based on various attributes. This model would work well to assist healthcare professionals identify high risk patients to help prevent them from getting a stroke and reduce the burden of cost of treatments. Although the data size is a little over 5,000, which may not be ideal, with an accuracy of over 90%, it can be reassuring to know the results of this predictive model will be reliable.

According to the World Health Organization (WHO), stroke is the second leading cause of death globally, responsible for approximately 11% of total deaths. Life after a stroke can be

profoundly debilitating--having to deal with loss of mobility and cognitive function--leading to an overall lower quality of life and oftentimes, depression. This model collects data that can be easily and quickly attainable without requiring expensive tests or blood work in addition to being over 90% accurate. This predictive model would be highly useful for focusing on primary prevention of a stroke and therefore contribute to providing optimal patient care by reducing treatment costs and pill burden, ultimately improving a patient's quality of life.