Capstone Two Project Proposal: Predictive Modeling for Stroke Risk Assessment Using Machine Learning

Problem Identification

Problem statement formation

Strokes are a leading cause of death and disability globally, affecting millions of people each year. Even with all the progress we've made in medicine, accurately predicting the risk of stroke remains a challenge to date. The traditional method mainly relies on a handful of well-known factors that lead to inaccurate risk assessments, potentially hindering preventive interventions and personalized healthcare strategies. This capstone project aims to develop a predictive model that accurately assesses the risk of stroke in individuals based on various health-related features.

Context

Stroke prevention is crucial to reduce the disease burden and overall healthcare costs. Early identification of high-risk individuals can lead to timely interventions and minimize the disability by implementing effective preventative measures. Using machine learning in healthcare could pinpoint the crucial health factors linked to stroke risk.

Criteria for success

The following criteria will measure the success of this project:

- *Predictive Accuracy*: The machine learning model should be able to identify individuals at high risk of stroke with high sensitivity and specificity

- *Clarity:* The model should provide clear and actionable information to healthcare providers regarding the most significant risk factors for stroke

- *Usability:* The final appropriate model should be user-friendly and easily incorporated in the existing healthcare system for practical use

Scope of solution space

This project will focus on developing and evaluating machine learning models for stroke risk prediction. It will involve data preprocessing, feature engineering, model selection, training, and evaluation. It will explore various machine learning algorithms, including logistic regression, decision trees, random forests. Feature selection techniques will be used to identify the most relevant predictors of stroke risk. Then, the model will be trained and validated using a comprehensive dataset.

Constraints

The availability and quality of data will constrain the project. Access to a comprehensive dataset with relevant health and lifestyle factors is essential. Another constraint is that the model must comply with healthcare regulations and data privacy laws, protecting the confidentiality and security of patient's healthcare information.

Stakeholders

The primary stakeholders for this project include:

- *Healthcare providers:* Physicians, Nurses, etc. who can use the model to identify high-risk patients
- *Healthcare Organizations*: Hospitals and clinics looking to improve patient outcomes
- *Researchers*: Academics and medical researchers
- *Patients*: Individuals at risk of stroke

Data Sources

This project will utilize publicly available Kaggle datasets, including a wide range of variables relevant to stroke risk assessment, ensuring a comprehensive dataset for model training and validation.

> https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset?select=healthcare-dataset-stroke-data.csv

Proposed solution approach

1. *Data Collection and Preprocessing:* Gather and preprocess data from the identified sources and perform data cleaning, handling missing values and transforming variables as needed.

2. *Exploratory Data Analysis (EDA):* Perform EDA to understand the distribution of variables and identify potential correlations and patterns relevant to stroke risk.

3. *Feature engineering:* Extract relevant features from the dataset, including traditional risk factors (age, sex, blood pressure, etc.) and potentially novel features (genetic markers, lifestyle patterns, etc.) for predicting stroke risk.
4. *Model selection and training:* Explore various machine learning algorithms (e.g., logistic regression, decision trees, random forests) and train models on the preprocessed data.

5. *Model evaluation and selection:* Evaluate the performance of different models using appropriate metrics (e.g., accuracy, sensitivity, specificity). Then, select the best-performing model based on these metrics and its interpretability.

6. *Model Interpretation*: Analyze the model to understand the significance of various predictors and provide actionable insights for healthcare providers.