Team 05: Lukyan Sukhachevskyi, Jacob Thomas Miller

Predicting Stroke Risk in Elderly Patients Using Machine Learning

Motivation

Strokes are a leading cause of death and disability worldwide, particularly among the elderly population. Early prediction and prevention of strokes could significantly improve patient outcomes and reduce healthcare costs. Our project aims to develop a machine learning model that can predict the risk of future strokes in elderly patients based on their blood pressure, heart rate, and other physical characteristics. This project is motivated by the need for more accurate and personalized stroke risk assessment tools, as current methods may not fully capture the complex interplay of risk factors in individual patients. By leveraging machine learning techniques, we hope to identify subtle patterns and relationships in patient data that may not be apparent through traditional statistical methods. If time permits, we aim to expand our model to predict and monitor other cardiological conditions (coronary heart disease, heart attacks, kidney disease, etc.). Our work differs from existing approaches by incorporating a wider range of physiological parameters and utilizing advanced ML algorithms to improve prediction accuracy.

Methodology

We will use a supervised learning model to predict stroke risk:

1. **Data Collection**: For initial model development and validation, we will use the dataset from Systolic Blood Pressure Intervention Trial (SPRINT) mentioned in “Blood Pressure Management in Stroke: Viewpoint” study by Philip B. Gorelick from Northeastern University. This data includes blood pressure, heart rate, and other relevant patient characteristics that will be helpful in predicting strokes. If possible, we will seek additional datasets from public health repositories.
2. **Data Preprocessing**: We'll clean the data, handle missing values, and perform feature engineering to create relevant input variables.
3. **Feature Selection:** We'll use techniques like correlation analysis and principal component analysis to identify the most relevant features in the blood pressure data that correlates to stroke prediction.
4. **Model Development:** We plan to implement and compare multiple algorithms, including Logistic Regression, Neural Networks, and Random Forest.
5. **Model Training and Validation:** We are thinking of using K-fold cross-validation to train and validate our models, optimizing our model’s hyperparameters using gradient descent, grid search, or random search techniques.
6. **Model Evaluation:** We will assess the performance of our model using accuracy, precision, and recall. We will evaluate our model's performance against existing stroke risk assessment tools (e.g., CHADS2 score) using the same dataset. Perform external validation using a separate dataset to assess the model's generalizability.

7. **Timeline:**

   Week 1-2: Data collection, preprocessing, and exploratory data analysis

   - Collect blood pressure data from public health databases
   - Sort and categorize relevant data

- Correlate data (age, weight, ethnicity, etc.) to blood pressure data

Week 3-4: Feature selection and initial model development

- Business logic code
- Library setup
- Exploratory Data Analysis and correlation analysis
- Determine feature set for modeling
- Start random forest implementation
- Indicate safe and problematic ranges for systolic and diastolic measurements

Milestone Report Deliverables (by end of Week 4):

- Completed data preprocessing and feature selection
- Implemented Logistic Regression and Random Forest models
- Initial evaluation results for these two models
- Preliminary insights from exploratory data analysis
- Updated timeline and any adjustments to project scope

Week 5-6: Model refinement, hyperparameter tuning, and cross-validation

- Start Logistic Regression model implementation
- Identify origin of errors and bias
- Implement NN
- Setup k-fold cross examination
- Implement false positives, and misclassifications to training dataset

Week 7-8: Model evaluation, comparison with existing tools, and sensitivity analysis

- Perform initial training and evaluation
- Compare initial results across all models
- Calculate performance metrics

Week 9-10: External validation and final report preparation

- Seek additional dataset for external validation
- Prepare final project report and presentation

Resources

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7666043/