Project Proposal

CS 131

Nathan Johnson

Lance Le

Title: Predicting Strokes with Machine Learning

Introduction:

Our project will be covering the topic of strokes and how we can predict if they are likely to occur in individuals based on variables such as age, sex, body mass index, average blood glucose level, and more. For years, strokes have been one of the leading causes of death worldwide. This project is significant because it helps show that we can use certain traits to predict the likelihood that an individual will suffer a stroke. These traits are included as features in the dataset we found on Kaggle. This dataset includes 5,110 entries and 11 columns, not including the target.

Problem Statement:

The problem that this project aims to solve is to be able to predict if a patient is likely to get a stroke based on simple traits/factors. This issue is perverse, as according to the World Health Organization (WHO), stroke is the 2nd leading cause of death globally, responsible for approximately 11% of total deaths. Being able to easily and non-invasively assess if a patient is likely to get a stroke would be invaluable and could save lives, especially for those who do not know that they may be at risk for a stroke.

Data Description:

The Stroke Prediction Dataset includes the following columns:

- id: a unique identifier
- gender: “Male”, “Female”, or “Other”
- age: the age of the patient in years
- hypertension: presence of hypertension (1) or lack thereof (0)
- heart_disease: presence of heart disease (1) or lack thereof (0)

- ever_married: “Yes” or “No”
- work_type: "children", "Govt_jov", "Never_worked", "Private" or "Self-employed"
- Residence_type: “Rural” or “Urban”
- avg_glucose_level: average glucose level in blood
- bmi: body mass index
- smoking_status: "formerly smoked", "never smoked", "smokes" or "Unknown"
- stroke: 1 if the patient had a stroke or 0 if not (this is the target column)

For data processing, we will need to take care of missing values in the BMI and smoking_status columns by either dropping them or replacing them with the average of the column (the latter only works for the BMI column). We can also replace the “No” and “Yes” classifications in the ever_married column with 0 and 1, respectively.

Methodology:

Our models will be binary classification models, as the goal is to determine whether or not the patient has a cardiovascular disease.

The three machine learning models we plan to use are logistic regression, neural network, and random forest.

For the neural network, we will utilize TensorFlow in order to build our network and Sci-Kit Learn KFold for hyperparameter tuning to decide on the most optimal width and depth for our model.

Objectives:

Our objectives for this project are to build and train the best model at predicting strokes as we can, with evaluation metrics in the 90s for each metric, and then present our model results.

Evaluation Metrics:

We will use the evaluation metrics accuracy, precision, recall, F1-score, and ROC AUC to evaluate our models as our problem is a classification problem and we want to make sure our models are not overfitting to the training data. precision, recall, F1-score, and

ROC AUC are particularly important as our dataset is imbalanced and these metrics will help assess how our model is dealing with the imbalanced data. We will also utilize matplotlib for visualization of our model results.

Timeline:

Create a timeline or schedule for your project, indicating milestones and deadlines for each task (e.g., data preprocessing, model training, evaluation, and documentation).

Task	**Deadline**
Download dataset	4/11/25
Preprocess data for training	4/16/25
Create base models	4/16/25
Train models	4/23/25
Validation and last tuning	4/30/25
Create slideshow presentation	4/30/25

Conclusion:

In this project, we plan on building a machine learning model that is able to predict if a patient is likely to have a stroke with high accuracy. We will use logistic regression, neural network, and random forest models along with k-fold hyperparameter tuning to build the best model possible. Creating this model will have a significant impact as strokes are one of the leading causes of death worldwide, and the ability to predict them before they happen has the potential to save countless lives.

References:

https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset/data