Stroke Detection: Prediction with Multiple Classification Models

Group 13 Project Proposal

University of Wisconsin - Madison
Computer Science 539 - Introduction to Artificial Neural Networks

Tony Wang - ywang3298@wisc.edu
Hanchi Chen - Hchen794@wisc.edu

1 Overview

The goal of this project is to use multiple machine learning classifying algorithms to predict whether a patient is likely to stroke based on several parameters like gender, age, various diseases, and smoking status. The ultimate goal will be finding an effective way to predict if the patient has stroke or not.

2 Background

According to the data of the World Health Organization (WHO) in 2019, stroke is the second leading cause of death, which is responsible for approximately 11% of total deaths respectively[1]. What's more, every 40 seconds, someone in the United States has a stroke. Every 3 minutes and 14 seconds, someone dies of stroke[2]. Stroke should be paid attention to more than thought.

The current workflow when patients want to check if they have stroke typically begins with a series of examinations, including Computerized tomography (CT), Blood tests, physical exams and so on. Not only is this approach time-consuming, but it also poses significant financial burdens. Rather than focusing predominantly on stroke treatment, the emphasis should shift towards prevention, especially considering the abrupt onset of stroke symptoms which often renders patients incapacitated, unable to seek timely assistance. A promising solution lies in leveraging neural networks to analyze pertinent data, thereby predicting individuals at heightened risk of experiencing a stroke.

3 Statement of Work

3.1 Datasets

We found the data from https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset/data. The data contains 5110 observation. There are 11 features in the dataset and there are some missing data for bmi. Because we are not sure what features will influence the risk of having stroke, for data pre-processing, we will use many techniques like one-hot-encoding to encode the categorical columns: 'gender', 'work type', and 'smoking status', feature scaling to standardize the features, etc. We are initially planning to 80/20 split the dataset and test accuracy against the 20% test sets.

S.No.	Attribute	Code given	Value ranges	Data type
1	Gender	gender	1, 0	Binary
2	Age	age	in years	Numeric
3	Hypertension binary feature	hypertension	1, 0	Binary
4	Heart disease binary feature	heart_disease	1, 0	Binary
5	Has the patient ever been married	ever_married	1, 0	Binary
6	Work type of the patient	work_type	0, 1, 2	Nominal
7	Residence type of the patient	Residence_type	0, 1	Binary
8	Average glucose level in blood	avg_glucose_level	55.1–272	Numeric
9	Body Mass Index	bmi	10.3-97.6	Numeric
10	Smoking status of the patient	smoking_status	0,1,2	Nominal
11	Stroke event	stroke	0,1	Binary

Table 1: Description of Dataset Attributes

3.2 Methods

Stroke prediction models predict an individual's risk of having a stroke based on various factors such as age, blood pressure, cholesterol levels, smoking status, and other medical history. Machine learning algorithms can analyze large amounts of data and identify patterns and correlations that can be used to make predictions with a high degree of accuracy.[3] Our aim is to optimize the predictive power of various machine-learning classifiers for stroke detection. Some existing work already reached more than 90 percent predictive accuracy in Kaggle. For example, RachidYZ achieved an accuracy of more than 95 percent in Figure 1.[4] This may be considered as an baseline result for us. We'll explore multiple models such as Support Vector Classifier (SVC), Random Forest, XGBoost etc. and specific Artificial Neural Network (ANN) models (may include some pre-trained models), compare the predictive accuracy and design the most proficient model for stroke detection. We may explore more models in the following week based on the course materials. For data pre-processing, we will use many techniques like one-hot-encoding to encode the categorical columns: 'gender', 'work type', and 'smoking status', feature scaling to standardize the features, etc. And since adjusting hyperparameters of these classifiers may affect the overall performance, we will dive deep about configuration.

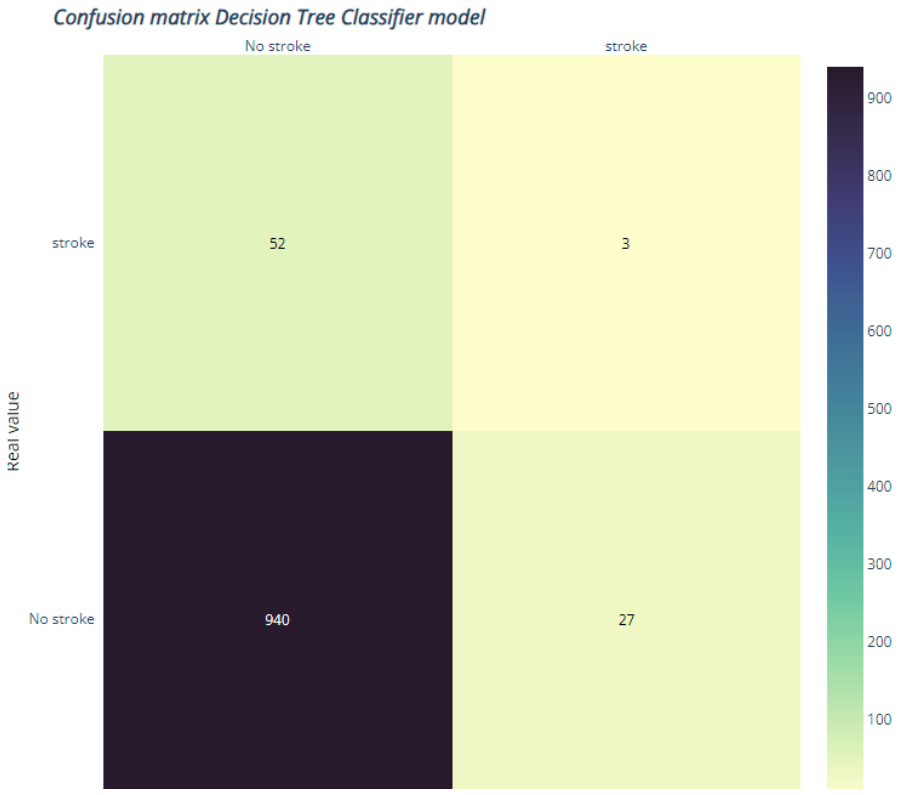

Figure 1.

3.3 Outcome and Performance Evaluation

Our primary objective is to design an efficient model for stroke detection by exploring a wide spectrum of classification models and determining the most proficient model based on predictive accuracy. To evaluate our model's performance, we will use both confusion matrices and predictive accuracy as our performance metrics. Past works in this field have set a high benchmark in terms of prediction accuracy. Given these implementations, our project aspires to achieve, or even surpass, a predictive accuracy of at least 95 percent.

4 Project Plan

4.1 Timeline

We breakdown the project into a set of tasks within a duration of 7 weeks. A Gantt chart is provided to visualize our schedule. We also prepare a GitHub repository with the URL link to the GitHub project.

Week 1

1. **Set Up Project Repository and Tools:**
    - Initialize the GitHub project repository.
      URL: `https://github.com/hchen794/CS539-Fall2023-group13`
    - Familiar with relevant models and required libraries and frameworks.
    - Prepare the collaborative environment on Google Colab for IPython-Notebooks (ipynb).
2. **Initial Project Proposal:**
    - Finalize and upload the project proposal to GitHub and Canvas.

Week 2

1. **Data Analyze and Pre-processing:**
    - Gather the dataset and analyze its properties.
    - Conduct data processing if necessary, for instance, we may select and combine different datasets together, data scaling and encoding are also possible.
    - Split the dataset into training, validation, and test sets.
2. **Begin Baseline Model Development:**
    - Analyze the existing models and re-construct for validation.
    - Initialize multiple models for comparison.

Week 3

1. **Baseline Model Training and Verification:**
    - Implement and train the baseline models using the training dataset.
    - Analyze and verify the baseline model's accuracy and loss metrics.
2. **Research Optimizations:**
    - Review literature and existing projects for optimization techniques suitable for the classification models.

Week 4

1. **Apply Optimizations to the Model:**
    - Integrate identified optimizations into the current models.
    - Retrain the model with optimizations.
    - Validate and compare its performance against the baseline model.
2. **Project Progress Report:**
    - Draft and finalize a progress report detailing all activities, results, and challenges so far.
    - Upload the progress report to GitHub and Canvas.

Week 5

1. **Advanced Model Development:**
    - Initiate the development of a more advanced or specialized model based on current research on different models.
    - Train and validate this model.
2. **Test the Model:**
    - Evaluate the model's performance using the test dataset.
    - Document the model's accuracy, loss, and other performance.

Week 6

1. **Results Analysis and Visualization:**
    - Analyze the model's performance in depth.
    - Prepare visualizations such as confusion matrices, ROC curves, etc.
2. **Draft Final Report:**
    - Begin writing the final report detailing the project's entire process, methodologies, results, and conclusions.

Week 7

1. **Project Presentation Preparation:**
    - Prepare slides or other presentation materials.
    - Practice the presentation to ensure clarity and coherence.
2. **Finalize and Submit the Report:**
    - Finalize the draft report after reviews and corrections.
    - Upload the final report and all project files to GitHub and Canvas.

4.2 Gantt Chart

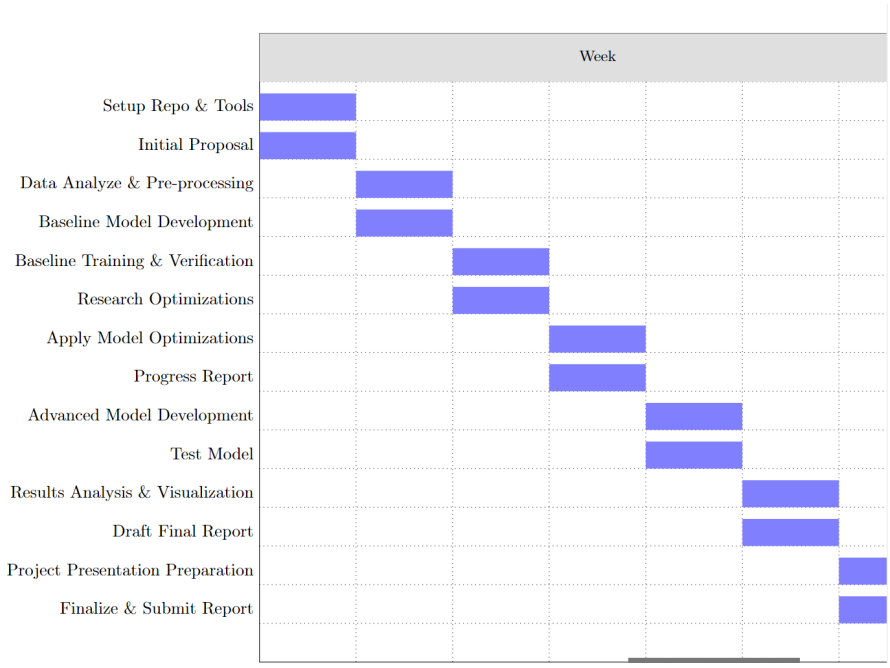

Figure 2.

5 References

[1]"The Top 10 Causes of Death." World Health Organization, World Health Organization, www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death. Accessed 21 Oct. 2023.

[2]"How Many People Are Affected by/at Risk for Stroke?" Eunice Kennedy Shriver National Institute of Child Health and Human Development, U.S. Department of Health and Human Services, www.nichd.nih.gov/health/topics/stroke/conditioninfo/risk. Accessed 21 Oct. 2023.

[3]MUSTAFA GÜRKAN ÇANAKÇI. (2023, March). Prediction with 7-Classification Models | ROC AUC. Kaggle. Retrieved October 19, 2023, from https://www.kaggle.com/code/mechatronixs/prediction-with-7-classification-models-roc-auc/notebook

[4] RACHIDYZ. (2020, May 22). EDA and modeling for predicting stroke. Kaggle. Retrieved October 19, 2023, from https://www.kaggle.com/code/rachidyz/eda-and-modeling-for-predicting-stroke