Machine Learning, I DATS 6202

Group Proposal

Group 2

Stroke Prediction

Instructor - Amir Jafari

Authors:

Shikha Sharma

Aron Rock

Sanjana

Data- 04/02/2023

**Problem** Statement:

Stroke is a leading cause of death and disability worldwide and early detection and prevention can greatly improve the outcome of a stroke. The goal of this project is to predict the likelihood of a person having a stroke in the future based on various risk factors such as age, gender, medical history, lifestyle habits, etc.

**Data Source:**

The data source for this project is the "Stroke Prediction Dataset" available on Kaggle (<https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset>). The dataset contains various features such as age, gender, hypertension, heart disease, and others that can be used as input features to predict the likelihood of a person having a stroke. The dataset is large enough to train a machine-learning model. The Stroke Prediction Dataset available on Kaggle will be used, Dataset has 5110 instances and 12 variables.

**Machine Learning Algorithms:**

For this project, we plan to use a Multi-layer Perceptron (MLP) Classifier algorithm. MLP is a powerful algorithm for multi-class classification problems and has been used successfully in various domains, including healthcare.

**Software:**

We will be using the scikit-learn and Keras library in Python to implement the MLP Classifier algorithm. Keras is an open-source software library for building and training deep learning models and provides a high-level interface for implementing neural networks.

**Reference Material:**

We will be using various research papers, books, and online tutorials to obtain sufficient background on applying the MLP Classifier algorithm to the specific problem of stroke prediction. Some of the reference materials that we will be using include.

Scikit-learn - <https://scikit-learn.org/stable/>

**Potential Smart Questions:**

1. gender: Is there a significant difference in the likelihood of stroke between genders?
2. age: Is there a relationship between age and the likelihood of stroke? If so, is it linear or nonlinear?
3. hypertension: How strongly is hypertension correlated with stroke, and how does this correlation vary with other variables in the dataset?
4. heart\_disease: How strongly is heart disease correlated with stroke, and how does this correlation vary with other variables in the dataset?
5. ever\_married: Is there a significant difference in the likelihood of stroke between married and unmarried individuals?
6. work\_type: Does the type of work a person does affect their likelihood of stroke?
7. Residence\_type: Does a person's place of residence (urban vs. rural) affect their likelihood of stroke?
8. avg\_glucose\_level: Is there a relationship between glucose level and the likelihood of stroke? If so, is it linear or nonlinear?
9. bmi: Is there a relationship between BMI and the likelihood of stroke? If so, is it linear or nonlinear?
10. smoking\_status: Is there a significant difference in the likelihood of stroke between smokers and non-smokers? If so, how does this relationship vary with other variables in the dataset?

**Performance Evaluation:**

The performance of the MLP Classifier algorithm will be evaluated using metrics such as accuracy, precision, recall, F1 score, and others. The metrics will be used to evaluate the performance of the algorithm on the validation datasets.

**Schedule:**

The project is expected to take approximately 4-6 weeks to complete, including data pre-processing, model training and evaluation, and presentation of results. The schedule for completing the project is as follows:

Week 1-2: Data pre-processing and exploration

Week 3-4: Training and evaluation of MLP classifier network and other algorithms

Week 5: Comparison of performance of different algorithms

Week 6: Presentation of results and conclusion.