Machine Learning, I DATS 6202

Project Proposal

Stroke Prediction

Instructor - Amir Jafari

Presented by Group 2:

Shikha Sharma

Aron Rock

Sanjana

Data- 03/29/2023

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/
https://towardsdatascience.com/5-smote-techniques-for-oversampling-your-imbalance-data-b8155bdbe2b5
https://towardsdatascience.com/understanding-binary-cross-entropy-log-loss-a-visual-explanation-a3ac6025181a
https://nicolo-albanese.medium.com/tips-on-principal-component-analysis-7116265971ad
https://medium.com/dataman-in-ai/sampling-techniques-for-extremely-imbalanced-data-part-i-under-sampling-a8dbc3d8d6d8

Performance Evaluation:

The performance of the MLP Classifier algorithm will be evaluated using metrics such as accuracy, precision, recall, F1 score, receiver operating characteristic (ROC) curve, and the area under the ROC curve (AUC). The F1 score is a measure of the balance between precision and recall and provides a single score that summarizes the performance of the algorithm. The ROC curve and AUC are measures of the ability of the classifier to distinguish between positive and negative cases and provide a single score that summarizes the performance of the algorithm. The metrics will be used to evaluate the performance of the algorithm on the validation datasets.

Schedule:

The project is expected to take approximately 4-5 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: Training and evaluation of MLP classifier network and other algorithms
Week 4: Comparison of performance of different algorithms
Week 5: Presentation of results and conclusion.