AIN2002Report

Kerem Cantimur, Alper ..., Timuçin ...

April 2023

Abstract

1 Introduction

Stroke is a medical condition which happens when blood supply to some region in brain is cut-off or significantly reduced. According to the World Health Organization (WHO), there are near 15 million cases of stroke every year [1]. About third of this cases results in death, other third results in permanent disability. This study aims to create an MLP model that can predict the possibility of stroke for a patient for patient's given status.

2 Related Works

There is a 2022 paper [2] which collects and utilizes Electroencephalography (EEG) readings, which used Gradient Boosting algorithms. This study achieved 80% AUC with Adaptive Gradient Boosting, 77% AUC with XGBoost and 78% AUC with LightGBM.

Another study in 2022 [3] which implemented; Logistic Regression, Random Forest, XGBoost, K-Nearest Neighbors, Support Vector Machine and Multi-Layer Perceptron. Their LR model had the best accuracy (73.52%), specificity (73.43%), and AUC (83.30%) score, whereas the MLP model recorded the best sensitivity and G-Mean scores, 81.4% and 75.83%, respectively.

3 Materials and Methods

3.1 Data Description

There are two datasets, the original "Stroke Prediction Dataset" from Kaggle [4] which has 249 stroke patiens and 4861 healthy people for a total of 5110 samples. The second dataset is a synthetic one created from the original set using a deep learning model [5] which has 632 stroke patients and 14672 healthy people for a total of 15304 samples. Both datasets have 11 variables: gender, age, hypertension, heart_disease, ever_married, work_type, Residence_type, avg_glucose_level, bmi, smoking_status, stroke; with "stroke" being the target variable.

3.2 Algorithms Used

We used MLP as our main model and also experimented with auxiliary models to compare its relative performance. Models were selected based on their AUROC in validation set and they were assessed based on their Kaggle submission public score.

For the specifications of the architecture of said models, see Github page for the code.

4 Results

Table 1: Performance Overview

Model	AUROC Mean	AUROC Best	Kaggle Mean	Kaggle Best
MLP_{AllSet}	0.9244 ±0.005	0.9334	0.8708 ±0.008	0.8889
$MLP_{OriginalSet}$	0.8724 ±0.007	0.8910	0.8408 ±0.013	0.8609
$MLP_{SyntheticSet}$	0.8762 ±0.003	0.8801	0.8476 ±0.007	0.8608
$XGBoost$	0.8756 ±0.002	0.8812	0.8476 ±0.004	0.8672
$RandomForest$	0.8579 ±0.003	0.8601	0.5231 ±0.034	0.5549
$LogisticRegression$	0.8754 ±0.01	0.8803	0.7780 ±0.012	0.7797
SVR	0.6723 ±0.012	0.6815	0.6239 ±0.007	0.6310

Among the models tried, MLP had the best results on avarage. Using the merge of the two set resulted in higher accuracies as it decreased the the model's overfitness.

Using methods such as Integrated Gradients showed us that MLP model gives high importance to the patients age, bmi, gender and avarage glucose levels.

While the AUROC score ranking of the models is consistent with their Kaggle ranking, there isn't a clear correlation between the two values as the models who had similar AUROC scores had differed greatly in terms of Kaggle score. Among the auxiliary models, XGBoost had both a AUROC and Kaggle score comparable to MLP models. Although XGBoost and Random Forest both use desicion trees as their base, gradient boosting seem to increase XGBoost's predictive power significantly.

5 Discussion

In this project we tried to built a model to predict stroke risks for possible patients. We mainly used MLP models with 3 different sets and used other models to compare their performance to the MLP models.

In the best model, we have found that patient's stroke risk increases greatly even though they are in a healthy state. Having high body mass index or avarage glucose levels is also increases stroke risk. Additionally, the model gives considerably higher probabilities to male patients compared to female ones. Other features had smaller effects on stroke risk: rural communities have smaller chances compared to urban ones, having heart disease, hypertension or smoking increases the risk, while work type other than children seemed to

contribute little to nothing. Oddly enough, work type children raises the stroke risk, which is something we can't explain.

6 Conclusion

We experimented with some models on a dataset for stroke patients to see the difference between the predictive power of the models, as well as the difference caused by using slightly different distributions with the same features. Due to their inherent algorithmic superiority, neural networks have advantages over other models, an effect which appears more clearly as the data size grows. With the further advancement of training and testing methods, neural networks will become a valuable asset in making important decisions.

7 References

1. World Health Organization Eastern Mediterranean, https://www.emro.who.int/health-topics/stroke-cerebrovascular-accident/index.html

2. Islam, M.S.; Hussain, I.; Rahman, M.M.; Park, S.J.; Hossain, M.A. Explainable Artificial Intelligence Model for Stroke Prediction Using EEG Signal. Sensors 2022, 22, 9859. https://doi.org/10.3390/s22249859

3. Kokkotis, C.; Giarmatzis, G.; Giannakou, E.; Moustakidis, S.; Tsatalas, T.; Tsiptsios, D.; Vadikolias, K.; Aggelousis, N. An Explainable Machine Learning Pipeline for Stroke Prediction on Imbalanced Data. Diagnostics 2022, 12, 2392. https://doi.org/10.3390/diagnostics12102392

4. Fedesoriano, Stroke Prediction Dataset, https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

5. Binary Classification with a Tabular Stroke Prediction Dataset, https://www.kaggle.com/competitions/pl series-s3e2/data

References

[1] World health organization eastern mediterranean; stroke, cerebrovascular accident. *WHO*.

[2] Mohammed Saidul Islam, Iqram Hussain, Md Mezbaur Rahman, Se Jin Park, and Md Azam Hossain. Explainable artificial intelligence model for stroke prediction using eeg signal. *Sensors*, 22(24), 2022.

[3] Christos Kokkotis, Georgios Giarmatzis, Erasmia Giannakou, Serafeim Moustakidis, Themistoklis Tsatalas, Dimitrios Tsiptsios, Konstantinos Vadikolias, and Nikolaos Aggelousis. An explainable machine learning pipeline for stroke prediction on imbalanced data. *Diagnostics*, 12(10), 2022.

[4] Fedosoriano. Stroke prediction dataset. `https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset`, 2021.

[5] Binary classification with a tabular stroke prediction dataset. `https://www.kaggle.com/competitions/playground-series-s3e2/data`, 2023.