STROKE PREDICTION

Gazi University
Faculty of Engineering

Gamze AKSU
171180005

Cansu AYTEN
171180010

Ceren Umay Özten
181180060

Instructor: Dr. Duygu Sarıkaya
TA: Aybike Şimşek Dilbaz

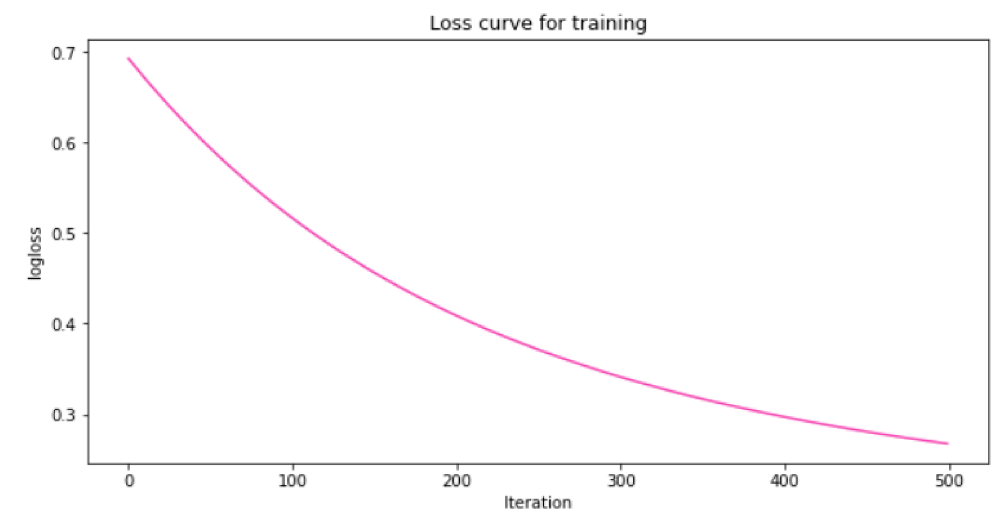
Group Jigglypuff

INTRODUCTION

In project stroke prediction dataset credentials are prepared for deep learning layered structure and we get the critic raw data for realizing train process of layered network nodes that changes weight and bias parameters. First we analyse the column information of stroke prediction dataset -gender, age, hypertension, heart disease, ever married, work type, residence type, average glucose level, BMI, smoking status and lastly stroke condition- we set apart minority stroke positive conditioned patients to analyse columns qualifications critically identifying them. And we also set apart outliers of stroke prediction dataset that can cause wrong data depictions connected to training of neural data signals. Then we structure the our from scratch deep learning network layers and we assign own weight and bias parameters across backpropagation and forward propagation functioning mechanisms. Our deep learning network structure has two hidden layer and one input and output layers. And this deep learning network structure has training data that is consist of just train part, just 0.15 validation part and just 0.15 percentage test part. We train our neural network structure with training data and we experience parameter changings during data training in neural nodes. Our neural model structure can be controlled over inner workings and we can change parameters and biases.

RESULTS

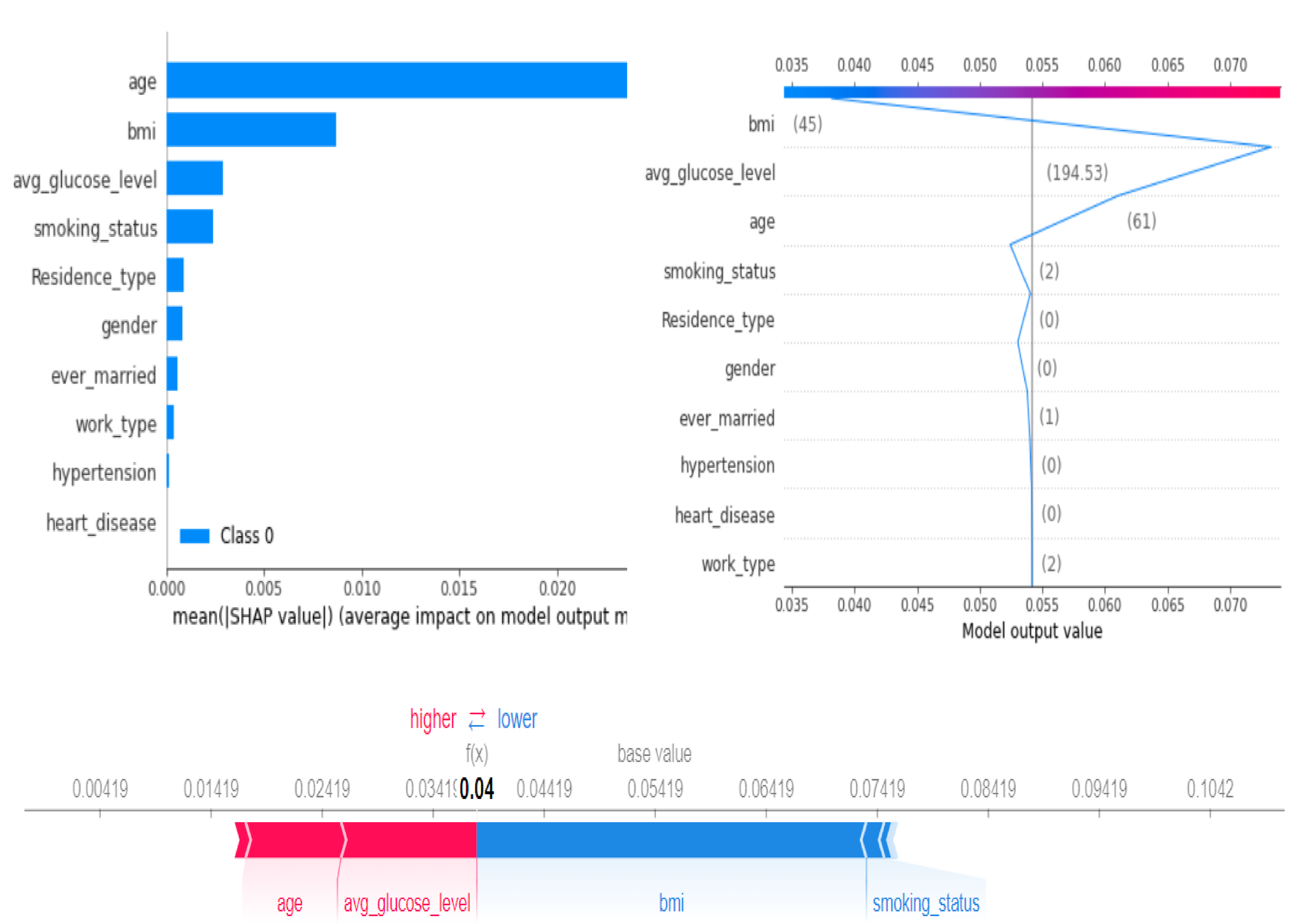
We obtained 0.94 accuracy for both models created. Experiments were carried out by changing the learning rate, epoch and initial weight values in the model. After adding interpretability to the model, Age feature affects the model a lot as it has much more contribution than others. Therefore, the age feature was removed and the effect of other features on the model was examined. After removing the age feature, it is seen that it is the average glucose level feature that contributes the most to the model. It is seen that the feature with the least contribution in the summary plot of the model is the heart disease feature. Therefore, the heart disease feature was removed and the effect of other features on the model was observed. After removing the heart disease feature, the age feature again contributes by far the most. The loss graph is as follows.



MATERIAL AND METHOD

DATASET

We used the Stroke Prediction Dataset, which we have accessed from the Kaggle platform, for the stroke prediction project. This dataset includes 11 features. These are id, gender, age, hypertension, heart disease, marriage status, work type, residence type, average glucose level, BMI (body mass index), and smoking status. id feature was removed from the dataset. Using these 10 features and using deep learning techniques, the probability of a person having a stroke will be estimated. There are 5110 records in the dataset. There are 201 null values in the BMI feature. These null values have been changed to be the median of the BMI feature. Then outlier values were found. Outlier values are also replaced with the median of the feature.

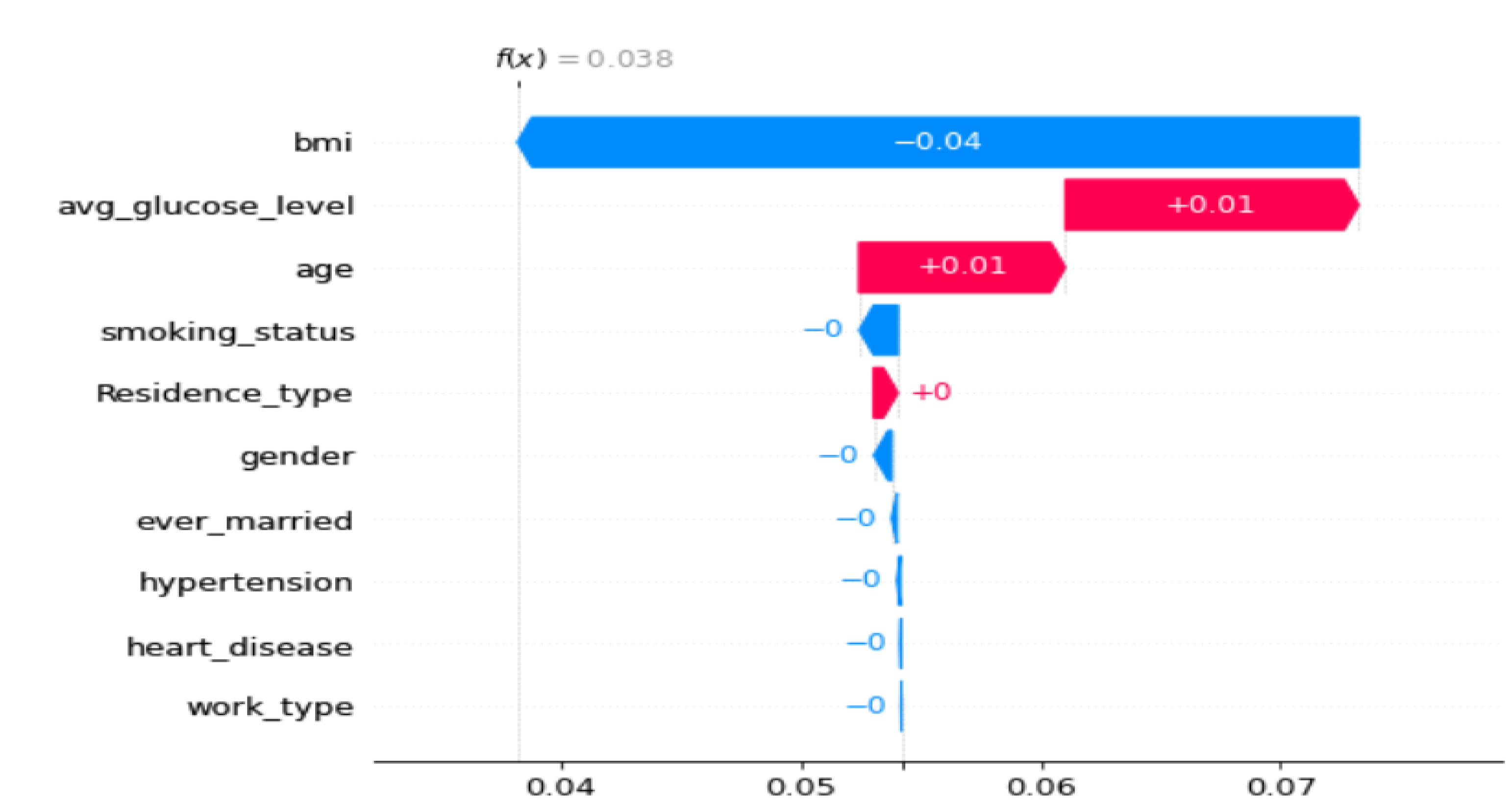


MODEL

In our ANN model, we implement Dense layers that have sequentially 64-32-16-1 number of neural nodes from input layer -hidden layer- output layer. We assign "relu" activation function for each of hidden layers and assign "sigmoid" activation function as sigmoid probability distributions over binary classification of stroke condition. Then we compile our model with "binary_crossentropy" loss and fit our model with 50 epochs for loop cycle.

Then, we added our model to interpretability. Interpretability is used to get the models created with Deep Learning out of the Black Box state. Thus, the model's decision-making style can be easily understood by associating it with its outputs. Model interpretability makes deep learning or machine learning processes more effective in many fields such as medicine and finance.

Interpretability can be examined under two headings, a global method and a local method. A global method is used to understand the decision-making process according to the general structure of the model. A local method is used to understand the decision-making process for a single instance of the model.



RECOMMENDATION

After the outliers were removed in the dataset used within the scope of the project, 5109 records remained. This number is not sufficient for a complex model. Therefore, retraining can be done with more records. Similarly, the dataset used is unbalanced. Negative stroke values in the dataset are much higher than positive ones. Therefore, retraining can be done with a balanced dataset or the dataset can be balanced by using data augmentation methods in the preprocessing stage. In this way, better results can be obtained.

REFERENCES

1. C. -C. Peng, S. -H. Wang, S. -J. Liu, Y. -K. Yang and B. -H. Liao, "Artificial Neural Network Application to the Stroke Prediction," *2020 IEEE 2nd Eurasia Conference on Biomedical Engineering, Healthcare and Sustainability (ECBIOS)*, 2020, pp. 130-133.

2. Chantamit-o-pas, P., Goyal, M. (2017). Prediction of Stroke Using Deep Learning Model. In: Liu, D., Xie, S., Li, Y., Zhao, D., El-Alfy, ES. (eds) Neural Information Processing. ICONIP 2017. Lecture Notes in Computer Science(), vol 10638. Springer, Cham.