CS 4710 – Artificial Intelligence
Predictive Classification Model for Diabetes Identification
Author(s): Ishan Koroth (ik4vrb@virginia.edu)

Introduction

When it comes to health, people tend to raise their concern for such a topic and prioritize it if they happen to possess certain symptoms that lead to a particular health risk. In this case, diabetes is a disease that has been considered a major concern, with the number of diabetes patients going from 108 million in 1980 to 422 million in 2014 (WHO, 2023). Diabetes is also a cause for multiple different health risks, such as heart attacks and kidney failure, which shows that this disease is something that needs to be addressed.

Fortunately, people have been developing ways to detect signs of diabetes based off different characteristics a person may experience. To state an example, a few Korean researchers have developed a predictive model with accessible health screening test parameters, which would then be used in a random forest algorithm to make predictions regarding someone having signs of diabetes. In terms of social good, this project seeks to contribute to the ongoing effort in analyzing signs of diabetes to ensure that the client takes preventive measures to avoid or delay diabetes.

With regards to work being done in this project, a neural network model had been developed to make predictions, and an accuracy within 70-75% had been achieved upon refining the model. Such a result may not be desirable, but with more thorough data analysis/features and better ways for hyper parameter tuning, significant improvements can be made to further expand upon this project.

Methods

To solve this problem, I decided to develop a neural network model that takes in multiple features and create potential new ones within the network. In doing so, the model would make predictions based off the features and activated nodes to calculate the results. This design was chosen because the neural network's ability to iteratively detect patterns via input and output through artificial neurons would generate thorough results for analysis. Even though logistic regression is the model generally used for classification problems, I felt that the neural network model fit better with the high variance in the sample data. Before developing the neural network, I first decided to use a logical regression model as a baseline for the classification model.

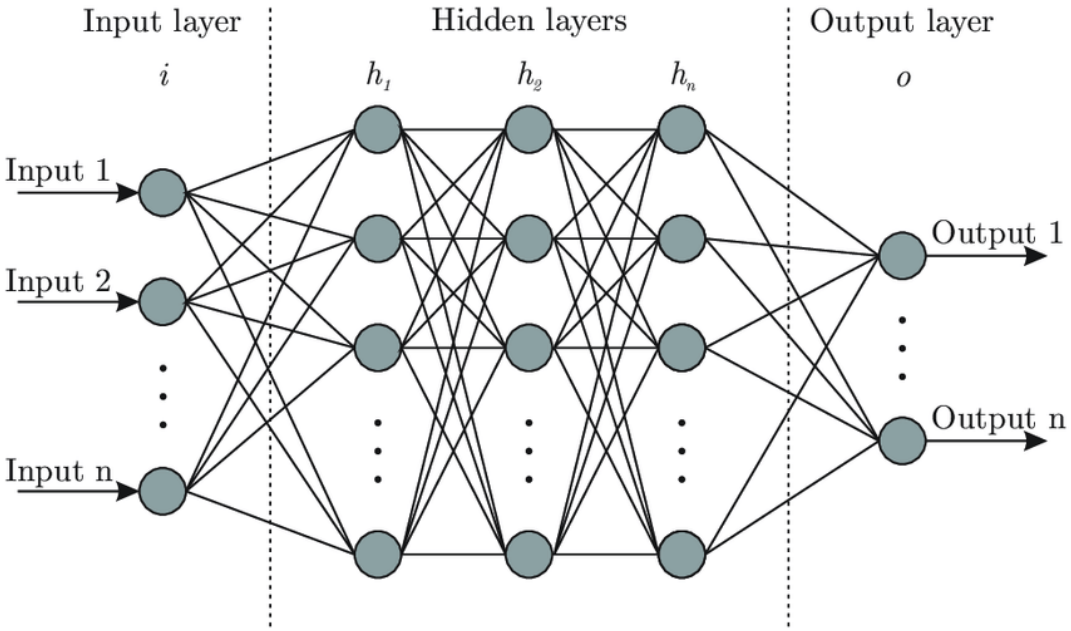

Figure 1 Artificial Neural Network Model Diagram (Facundo Bre, 2017)

Using the sklearn.linear_model library, I imported the LogisticRegression library that was used to train and test the dataset for analysis. Since the model was being used as a baseline, I did not emphasize much hyperparameter tuning. This model uses the limited-memory BFGS algorithm to compute parameter estimation for solving the classification problem.

As for how the neural network model was developed, I designed two separate models based off different Python libraries to verify and validate my results. I chose to do this as the results I got from the first model was not what I expected. The first model, in this case, was the MLPClassifier library imported from sklearn.neural_network. The model contains 5 hidden layers, with the number of nodes for each one being set as follows: 40, 120, 80, 120, 40. The rectifier, or ReLU, was used for the activation function, and the Adam optimization algorithm was used to execute stochastic gradient descent for the machine learning model.

The second model makes use of Tensorflow's Sequential model, which allows the user to add layers of different kinds, with the Dense layer being commonly used for hidden layers. The node values for the hidden layers were 75, 50, 25, and 1. All the Dense layers use the ReLU activation function, with the last layer being an exception since it uses a sigmoid activation function. Along with the Dense layers are Dropout layers, which are used to help prevent overfitting by setting certain outputs to 0 with a frequency rate. The rate used in this case was 0.01. Lastly, to compile the model and its results, a binary cross entropy loss is executed, and the Adam optimization algorithm is also used for this case.

Data

The dataset used was pulled from Kaggle, and the title of the dataset is called Diabetes Health Indicators Dataset (Teboul, 2022). As the title states, the features

include health-related factors like blood pressure and cholesterol, and habitual factors like physical activity and smoking. Upon understanding what the dataset contained, I saw promise in having multiple factors to analyze the classification problem. The website contained 3 cleaned datasets. The first one classified the diabetes target variable with 3 classes: non-diabetic, pre-diabetic, and diabetic. This dataset is followed by an identical data set that used a binary classification. Due to an imbalance where non-diabetics made up the majority, a third dataset was created to resample the data. This involved a 50-50 split between non-diabetic and diabetic binaries. For this project, the latter-most dataset was used for the experimentation phase.

Since the dataset was already cleaned, there was no NA data to be found. The columns all contained numerical data, so there was no need to encode anything that was not numeric. When it came to splitting the data into the target variable and features, the binary column used to identify the status for diabetes was used for the target variable. The data would then be split into X (feature data) and Y (target variable data). Following this step, the X data set is then fit transformed to scale each column to an appropriate value mapping scheme. Upon doing this, each of the data sets split into train-test, with the split being 75% and 25% respectively. From here, this process will be iterated two more times for the sake of further experimentation. The second iteration will only contain health condition related features, and the third iteration will only contain habitual related features.

Experiments

For the experiment, I focused on analyzing the data with the two neural networks in three different cases:

- all features
- only health condition related features
- only habit related features

The baseline logistic regression model is used to evaluate the first case for testing purposes only. The two neural network models are trained with the train data and tested for validation accuracy with the test data. For the evaluation metrics, a confusion matrix and classification report are used for the sklearn neural network model to evaluate the accuracy, precision, recall, f1-score, and other metrics. Since the TensorFlow model was used for validating the results from the sklearn neural network model, only the accuracy metrics was used to evaluate and compare with the Sklearn results, which will be represented with a confusion matrix and classification report.

Results

After running the models on the datasets and analyzing the results, the highest accuracy achieved on the neural network model was 74.76% when computing all features. Compared to 74.57% with health condition related features only and 68.33% with habitual related features only, it seems that health conditions have more significance in accuracy. What makes this more interesting is the fact that the baseline logistic regression model achieved around 74.62% when computing all the features into the input. This reveals that the neural network model did not make much improvement in comparison. Even though the methodology worked to an extent, the accuracy of the classification could be much better. When it comes to detecting health conditions, a high accuracy is needed for the model to be considered reliable for proper use.

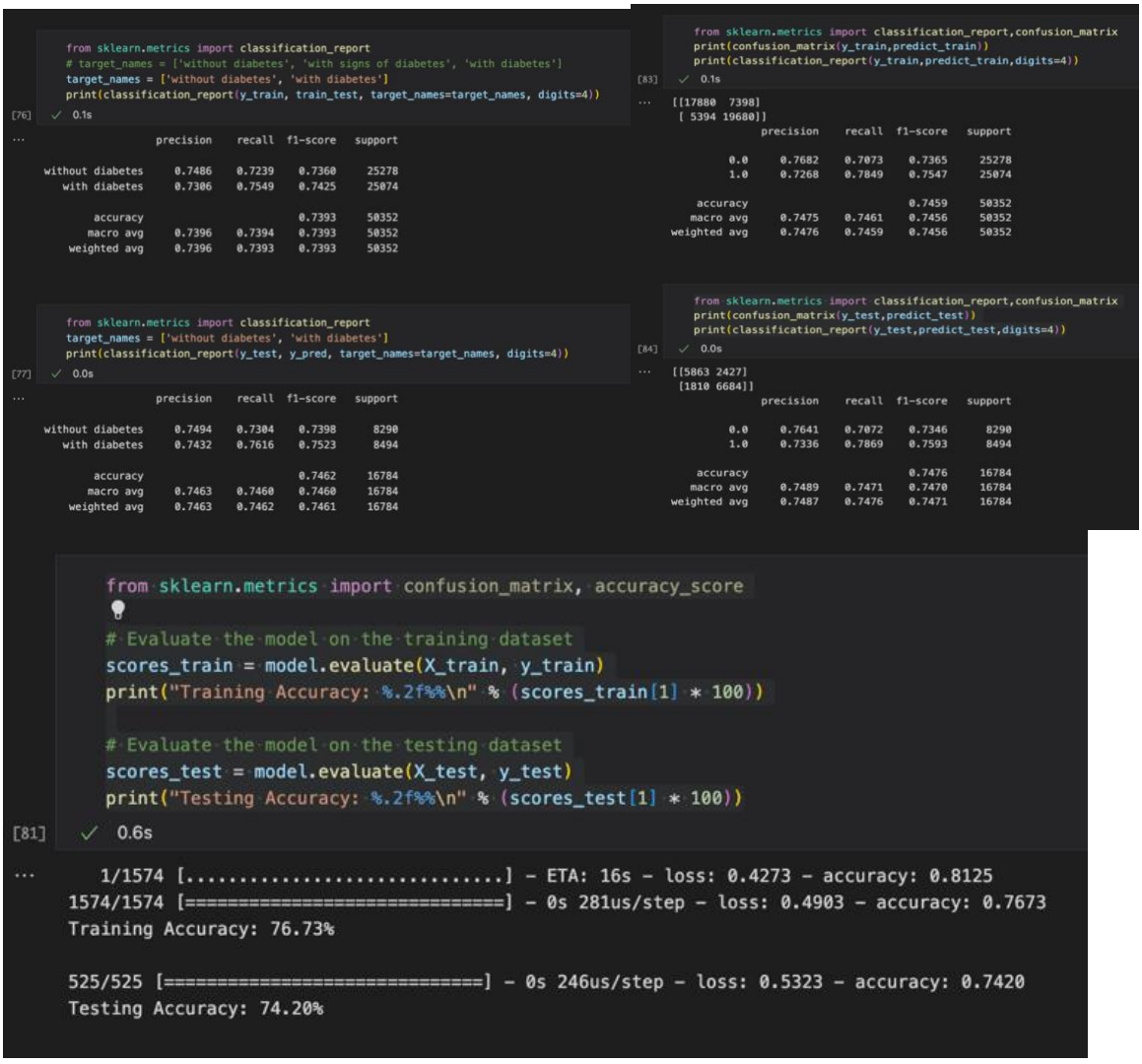

Figure 2 Results when considering all features. Top-Left: Logistic Regression Results; Top-Right: Sklearn MLPClassifier Results; Bottom-Left: Tensorflow Sequential Model Results (0 = non-diabetic, 1 = diabetic)

Efforts in modifying to improve the neural network model had been done to improve the accuracy of the classification, such as increasing the learning rate,

alpha, and hidden layer node values. However, even though the training accuracy increased significantly, the test/validation accuracy would either remain stagnant or decrease, which indicates signs of overfitting in the model. As a result of this, the model had to be modified such that the model would stop running when it no longer sees an improvement in the test/validation accuracy.

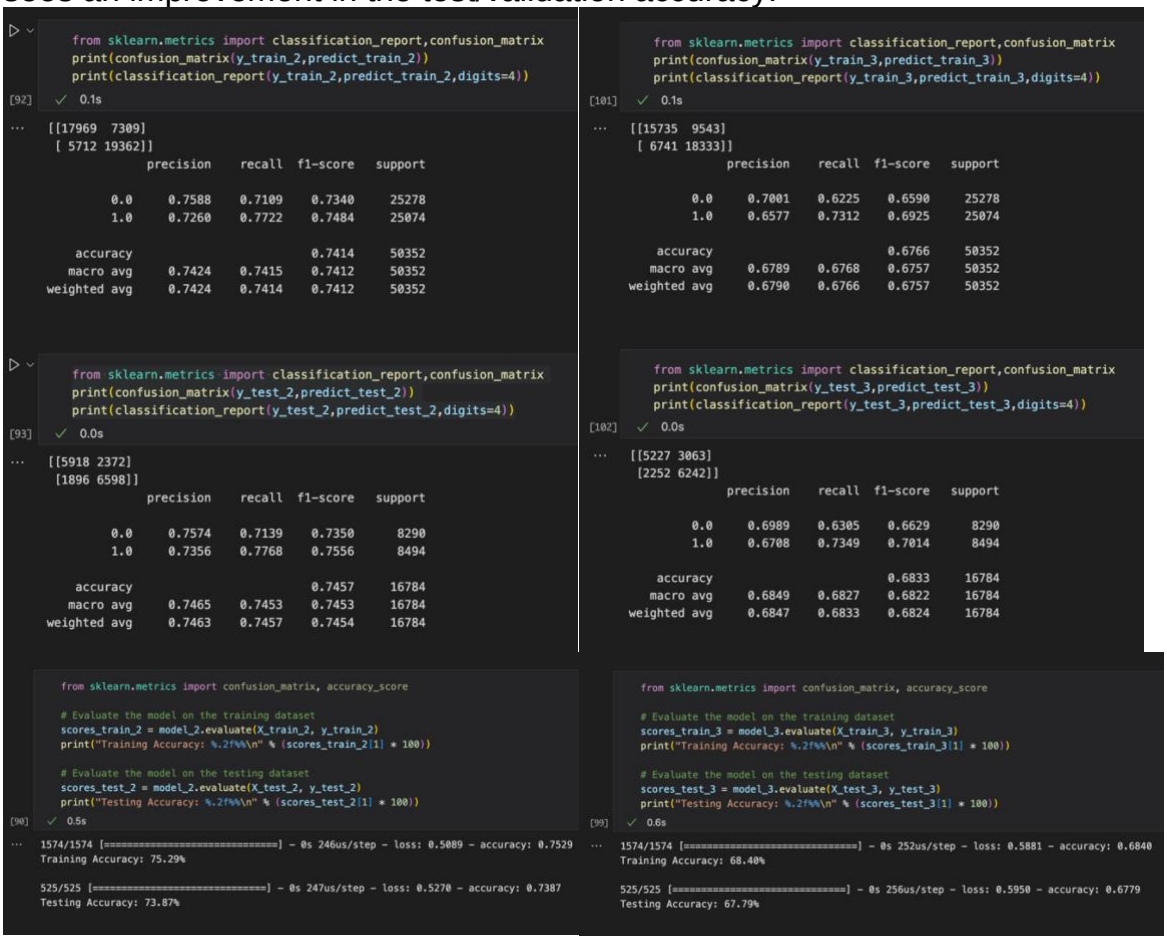

Figure 3 Results for next two iterations. Top-Left: Sklearn MLPClassifier Results for Health Condition Features; Bottom-Left: Tensorflow Model Accuracy Verification for Health Conditions; Top-Right: Sklearn MLPClassifier Results for Habitual Features; Bottom-Left: Tensorflow Model Accuracy Verification for Habitual Features;

Conclusion

Predictive classification models can be helpful in detecting signs of diseases arising, and diabetes is one of such diseases that people in general wish to either prevent or delay. A model in detecting signs of diabetes based off health indicator factors is desirable, and this project is an example of a model that shows some promise in detecting signs of diabetes with a fair percentage for an accuracy. Of course, the model can be improved upon with more concise and direct data, such as race, family history, and other specific biological information. Even though such data may be non-numeric, it can still be encoded using one-hot encoding to then have values that can be computed into the neural network. On top of that, having more historical data, like family history, could make a difference in outputting desirable

accuracies. However, the availability of such data is limited, possibly due to privacy-related reasons, which may fall under HIPAA regulations.

According to MAYO clinic, the main risk factors for type 2 diabetes are family history, environment factors, geography, race/ethnicity, and being overweight/obese. Although BMI is one of the features, it does not accurately depict if someone is overweight or obese, so these columns could be helpful as well for the predictive model. MAYO clinic also states that the exact cause of most diabetes types is unknown. By getting as much information as possible that correlates to having signs of diabetes or sugar build up in the blood stream, the model will have the potential to have better results. Perhaps for a future project or improvement, more historical and biological data could be collected if the patients have given their consent.

References

- https://www.who.int/news-room/fact-sheets/detail/diabetes#:~:text=Over%20time%2C%20diabetes%20can%20damage,blood%20vessels%20in%20the%20eyes.
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9353566/
- https://www.mayoclinic.org/diseases-conditions/diabetes/symptoms-causes/syc-20371444#:~:text=Heart%20and%20blood%20vessel%20(cardiovascular,have%20heart%20disease%20or%20stroke.
- https://www.researchgate.net/figure/Artificial-neural-network-architecture-ANN-i-h-1-h-2-h-n-o_fig1_321259051
- https://www.kaggle.com/datasets/julnazz/diabetes-health-indicators-dataset