A Neural Network Classifier for Predicting Diabetes

Dae-Gyue Han

May 2023

0 Abstract

Diabetes mellitus is a type endocrinal disease that results high blood sugar levels if left unmanaged, which can lead to diabetic ketoacidosis in the short term. Left chronically untreated, diabetes can result in many adverse medical complications. These can include, but are not limited to, diabetic retinopathy, nephropathy, neuropathy and heightened risk of heart attacks and stroke.

Type 1 diabetes is an autoimmune disorder caused by the destruction of beta cells responsible for the production of insulin. Type 2 diabetes, also known as adult-onset diabetes, results from insulin insensitivity and/or insufficient production of insulin often brought about by a sedentary lifestyle and obesity. Finally, gestational diabetes occurs in pregnant women and can be reversible, though it may lead to the development of type 2 diabetes.

The main objective of this project is to optimize a neural network classification model. This model will be making predictions on whether individuals have diabetes using the results of survey conducted by the CDC. More specifically, the 2015 Behavioral Risk Factor Surveillance System (BRFS). This data-set is a pre-cleaned and processed data set drawn from Kaggle, as linked in the sources at the end of this document.

However, in the course of running experiments, it is determined that no model produces much more than 75% validation accuracy. It can be concluded that this is owing to the selection of the survey questions, the answers to which are used as data set features. Upon examining the BRFS codebook, other potentially relevant surveys, including those specific to diabetes, were found.

1 Data Analysis and Preparation

The data set contain 21 features and 70,691 samples. The types of features and their corresponding index used for this project is described by the table below. The attached codebook pdf describing the survey questions they were taken is sourced at the end of this document.

Table 1: Model Features by Index

Index	Feature
0	DiabetesBinary
1	HighBP
2	HighChol
3	CholCheck
4	BMI
5	Smoker
6	Stroke
7	HeartDiseaseorAttack
8	PhysActivity
9	Fruits
10	Veggies
11	HvyAlcoholConsump
12	AnyHealthcare
13	NoDocbcCost
14	GenHlth
15	MentHlth
16	PhysHlth
17	DiffWalk
18	Sex
19	Age
20	Education

The following is a histogram of each feature in the data set.

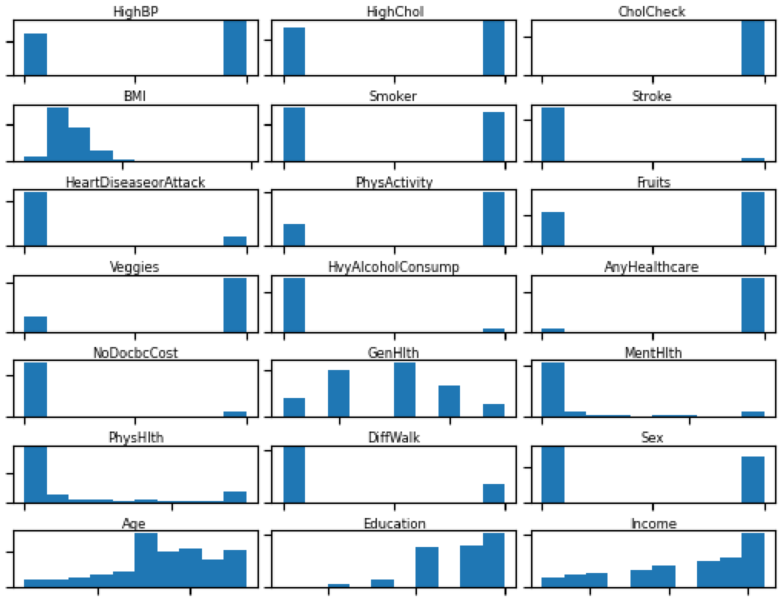

Figure 1.1: Histogram of the Features of the Data Set

The proportion of positive cases of diabetes versus false cases is determined to see whather the data is unbalanced. It was found that 35,345 of 70,691 samples are positive cases of diabetes. This is very near to a 50 percent split between true and false cases of diabetes. This is demonstrated in the histogram below.

As there is a nearly even split in the data, a simple validation accuracy metric can be used to evaluate models.

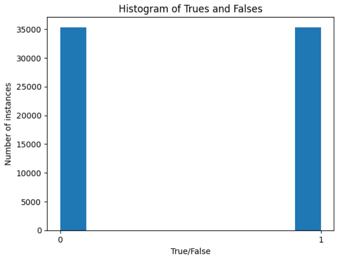

Figure 1.2: Histogram of True and False Cases of Diabetes

2 Model Selection And Evaluation

The models discussed are sequential neural networks. After some preliminary tests for potential architectures, a more thorough testing was conducted on them. Below are tables which displays the results of those trails. The Random baseline classifier is assumed to be 50% to match the split between true and false diabetes cases in the data. As the resulting model can vary between different training sessions, the trails were run on each model type. Their average accuracies are likewise displayed below.

Table 2.1: Validation Accuracies of Evaluated Models

Model	Trial 1	Trial 2	Trial 3	Average
Random baseline classifier	50.00%	50.00%	50.00%	50.00%
Logistic regression model	74.64%	74.77%	74.76%	74.72%
Neural network model (1-1)	74.90%	75.07%	74.61%	74.86%
Neural network model (2-1)	75.06%	74.88%	74.68%	74.87%
Neural network model (4-2-1)	75.12%	75.05%	75.01%	75.06%
Neural network model (512-1)	74.81%	76.33%	76.26%	75.80%

Table 2.2: Validation Accuracies of Evaluated Models

Model	Trial 1	Trial 2	Trial 3	Average
Random baseline classifier	50.00%	50.00%	50.00%	50.00%
Logistic regression model	74.65%	74.84%	74.82%	74.77%
Neural network model (1-1)	74.88%	74.91%	74.89%	74.86%
Neural network model (2-1)	74.92%	74.91%	74.65%	74.83%
Neural network model (4-2-1)	74.87 %	74.99%	74.86%	74.80%
Neural network model (512-1)	73.68%	74.62%	74.38%	74.23%

Displayed below are the accuracy curves over the lifespan of the training of different architectural models. Ultimately, I settled upon the 1-1 model as the best compromise of fit and training speed. While the validation accuracy for 1-1 models are better, the differences between it and the other models are generally marginal. Training speed was also a significant factor in my decision.

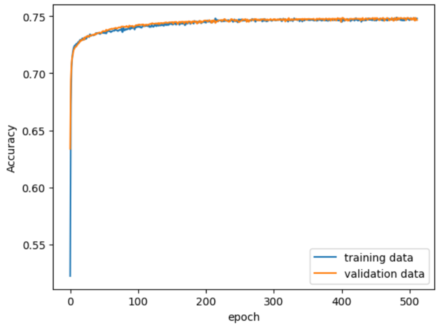

Figure 2.1: Logistic regression model

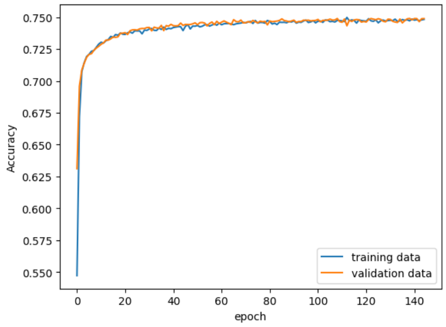

Figure 2.2: Neural network model (1-1)

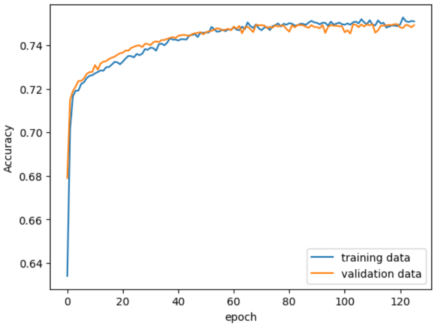

Figure 2.3: Neural network model (2-1)

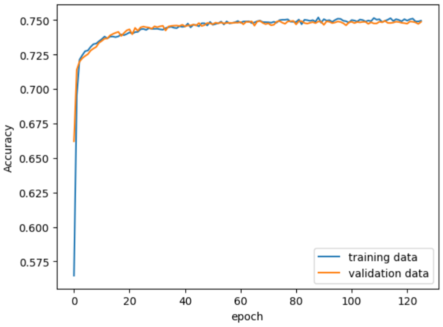

Figure 2.4: Neural network model (4-2-1)

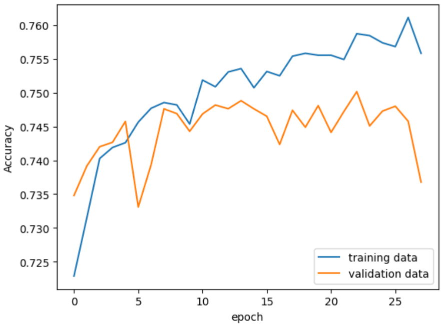

Figure 2.5: Neural network model (512-1)

As many of these models are quite similar to each other, it cannot be entirely certain that they were constructed correctly or are learning. In order to verify this, the output (i.e. target feature) is included as an additional input and a 1-1 model is trained. The accuracy curve produced by this is plotted below. Figure 2.6 shows that the model does indeed a 100% prediction accuracy, which would be expected, and confirms that these models are indeed learning as intended.

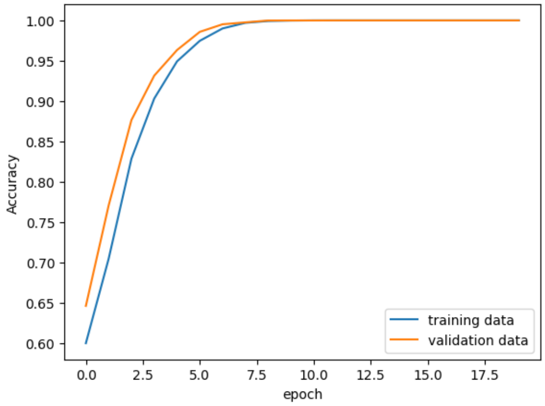

Figure 2.6: Feeding Output as an Input Feature

Finally, we would expect the 1-1 models to produce predictions congruent with the following equation:

$$prediction_i = sigmoid(relu(\sum_{i=0}^{20} x_i * w_i + b_i)w_2 + b_2) \tag{1}$$

Where, ‘i’ corresponds to each feature in each data set which feed into the input layer and ‘w’ and ‘b’ corresponding to weigh and bias. The weights and biases with the subscript ‘2’ are associated with the output layer.

A duplicate 1-1 neural network was trained and its weights and bias were extracted and new predictions recalculated using the equation above. The predictions from the model and those from the calculation are plotted on Figure 2.7. The plot confirms that our expected values matches the predictions made by the model.

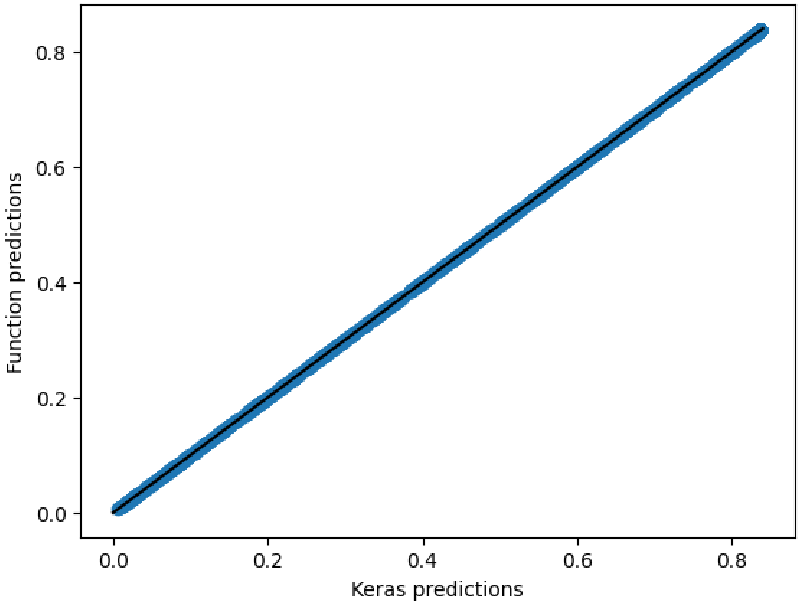

Figure 2.7: Feeding Output as an Input Feature

3 Feature Importance and Reduction

Finally, single-feature models were built to quickly assess feature relevance. While Figure 3.1 shows a bar plot of these data for easier visualization, Table 3.1 is also provided to give the exact figures used to derive the plot.

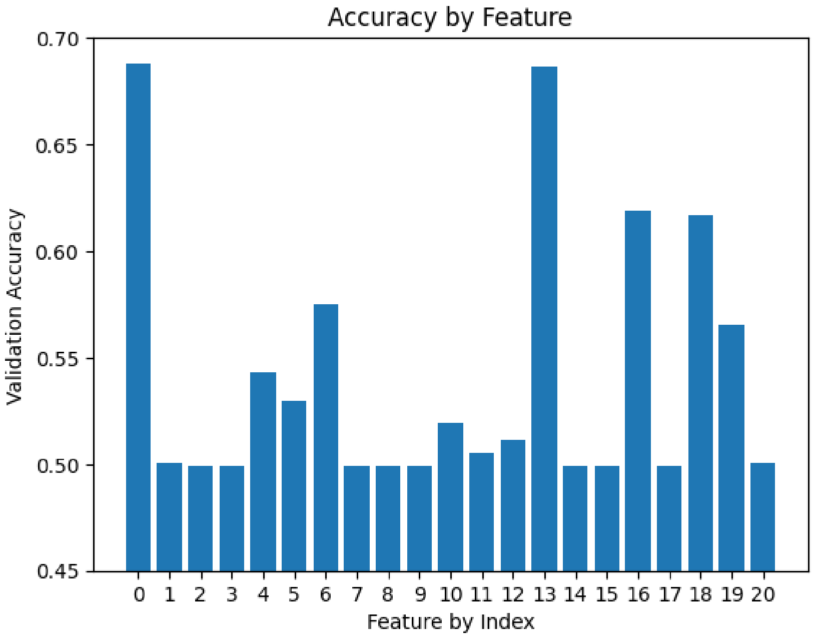

Figure 3.1: Feature Importance by Index

Table 3.1: Feature Indices and Importance

Feature	Index	Validation Accuracy as Single Feature Model
DiabetesBinary	0	68.82%
HighBP	1	50.05%
HighChol	2	49.95%
CholCheck	3	49.95%
BMI	4	54.32%
Smoker	5	52.98%
Stroke	6	57.48%
HeartDiseaseorAttack	7	49.95%
PhysActivity	8	49.95%
Fruits	9	49.95%
Veggies	10	51.98%
HvyAlcoholConsump	11	50.52%
AnyHealthcare	12	51.16%
NoDocbcCost	13	68.66%
GenHlth	14	49.95%
MentHlth	15	49.95%
PhysHlth	16	61.9%
DiffWalk	17	49.95%
Sex	18	61.7%
Age	19	56.56%
Education	20	50.05%

I used this data to systematically remove features one-by-one, by using low batch models to guide the decision which features (and in what order), I should drop features. This information was then used to train higher batch models. This produced six potential feature-reduced models. The table below describes features which were dropped for each model.

Table 3.2: Feature Indices and Importance

Model Number	Features Dropped
0	None (Baseline 1-1 model)
1	MentHlth
2	MentHlth, Income
3	MentHlth, Income, Veggies
4	MentHlth, Income, Veggies, PhysHlth
5	MentHlth, Income, Veggies, PhysHlth, High Chol
6	MentHlth, Income, Veggies, PhysHlth, High Chol, Fruits

Table 3.3 compiles the data from the training of these models, whilee Figure 3.2 plots the final validation accuracies of these reduced models. Note that the Model Number corresponds to those of Table 3.2.

Table 3.3: Validation Accuracies of Pruned Models

Model Number	Validation Accuracy
0	74.88%
1	74.85%
2	74.81%
3	74.71%
4	74.85%
5	74.46%
6	74.44%

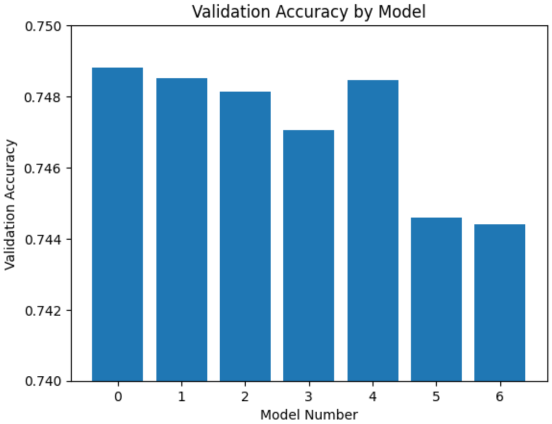

Figure 3.2: Validation Accuracies of Pruned Models

Overall, it appears that ’Model 4’, (which drops considerations of mental health, income, consumption of vegetables and physical health) produces the result closest to the original 1-1 model trained on the entire data set.

4 Conclusion

Examining the relevant survey questions leads me to conclude that many of the survey questions are not good indicators of diabetes. This is either because the features are weakly correlated to diabetes or because the questions do not yield particularly useful information. Likewise, the data set does not distinguish between different types of diabetes, such as between Type 1, adult-onset and gestational diabetes. This might explain why the results of the selected 1-1 model was not notably better than any of its alternatives. The roughly 75% prediction accuracy may well be the realistic limit of what can be achieved on the data set.

For example, the survey includes questions about whether fruit or vegetables are consumed more than one times per day. The question about physical activity simply asks whether the adult has done any physical activity or exercise in the past 30 days. The question about smoking merely asks whether you have smoked more than 100 cigarettes in your entire life. The question about high cholesterol asks if you had a check in the past five years.

Naturally, this information is incredibly limiting, because there are no information such as the frequency of exercise, how recently or how often a person smokes, precise measurements of how much fruit or vegetables they eat, or how often they check their cholesterol. The lack of continuous data may well be a hindrance to the training of a model.

Worse still, examining the 2015 BRFS Codebook reveals that there are other possible features that could have been used. As mentioned in the abstract, untreated diabetes is associated with elevated risk of stroke and heart attack. Both of these have relevant questions in the Codebook. There is also a question asking the age you were told you were diabetic, which may be helpful for characterizing the different types of diabetes individuals. Likewise, there is also a question that directly polls whether female patients only had diabetes during pregnancy, which we would expect from gestational diabetes.

It would be natural to think that different kinds of diabetes are associated with different risk factors. For example, patients born with Type 1 diabetes could not reasonably be associated with their lifestyle nor the quality nor availability of their healthcare, unless one cared to see how well they managed their condition. As such, I believe it would be better to narrow down the scope of the classification to a particular type of diabetes and/or select for features more directly related to the medical causes of diabetes. This can be accomplished by more in-depth research and/or consultation with domain experts.

5 Sources and Links

Github Hosting:
https://github.com/ChristopherDGHan/Final_Project_CS5300

Overleaf Link:
https://www.overleaf.com/project/64559e6c213b6223c0077b64

Kaggle Dataset Link:
https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset?resource=download

Data Cleaning Notebook:
https://www.kaggle.com/alexteboul/diabetes-health-indicators-dataset

CDC Behavioral Risk Factor Surveillance System 2015 Codebook Report:
https://www.cdc.gov/brfss/annual_data/2015/pdf/codebook15_llcp.pdf