Predicting the Presence of Diabetes with Binary Classification

Phase 3: Model Selection and Evaluation

Jesse Gempel
Introduction to Artificial Intelligence

November 21, 2024

Preface

My Google Colab notebook for this report can be located at this link:
https://colab.research.google.com/drive/12jXgEM1_QjnQAs97DlCXwOsPqnMgcmS6?usp=sharing.

1 Abstract

Diabetes is a common disease that affects roughly 12 percent of the American population[1]. It occurs in two different ways: Type I diabetes attacks the immune system, preventing the body from producing insulin; type II diabetes causes the body to intercept any amount of insulin, each though it may produce enough of it[2]. In either case, the human body's inability to utilize insulin is caused by one's excessive consumption of glucose. It can be easy for an individual to consume too many sugary foods and beverages. This issue is particularly noteworthy for Americans because a concerning amount of food relies on high fructose corn syrup and excess sugar. A simple task that we do everyday can cause a person to end up like the 1.6 million people who passed away from diabetes in 2021[3]. Many factors, such as exercise, drug use, and cholesterol can help infer the presence of diabetes. For this reason, an artificial intelligence project based on predictive neural networks can help one survey the correlation and importance of these factors.

The intent of this project is to predict the presence of diabetes among hospital patients. The process of making these predictions can be defined as a binary classification problem, with a **0** for *non-diabetic* or a **1** for *diabetic/pre-diabetic*. Real-life patient data will be analyzed with logistic regression-based neural networks to make such predictions. An overfitting model will be created first to memorize the patient data, which consists of 21 questionnaire-based features. Several other models will be created to determine the optimal configuration for predicting the presence of diabetes. Once the ideal model is determined, each individual feature will be trained on the model, and features that appear unimportant will be incrementally removed from the prediction process.

2 Data Exploration

The "Diabetes Health Indicators Dataset" was gathered from a data collection website called Kaggle[4]. It was also archived from the UC Irvine Machine Learning Repository[5]. The dataset contains 70,692 samples: exactly 50.0% of the dataset's samples are classified as *non-diabetic*, while the remaining 50.0% are labelled as *diabetic/pre-diabetic*.

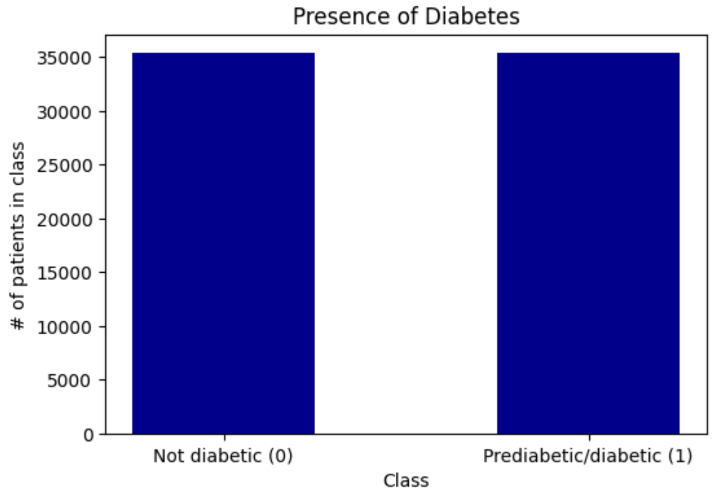
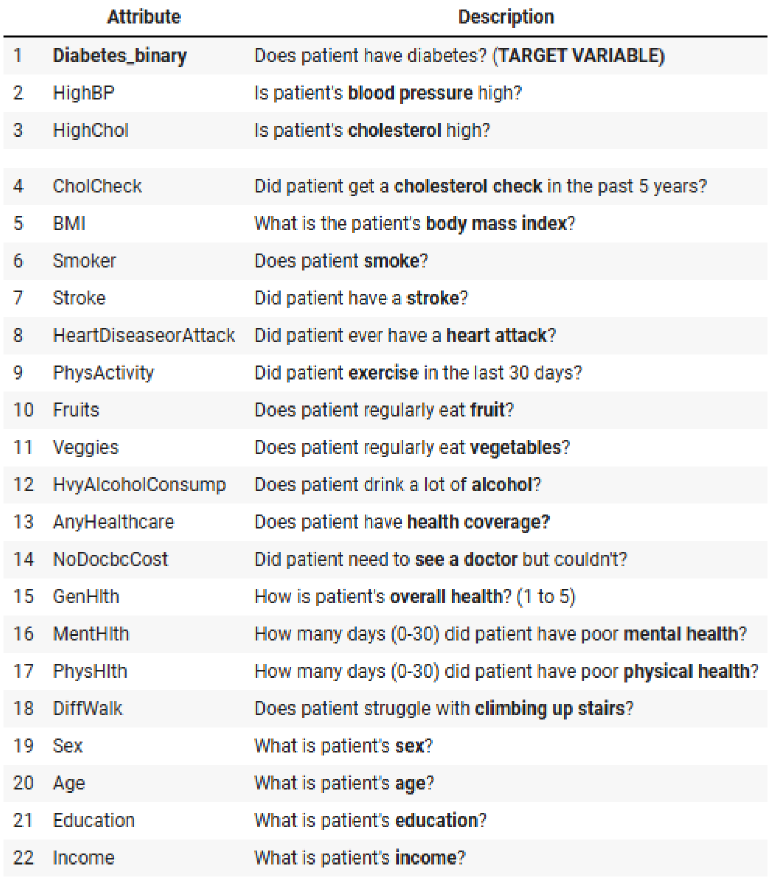


Figure 1: A breakdown of the quantity of features in each class. The **Not diabetic (0)** class contains 35,346 samples, while the **Prediabetic/diabetic (1)** class contains 35,346 samples as well.

Within each sample is a series of 21 features and 1 target variable. Each feature represents a piece of information gathered from a patient by a medical professional at a hospital. The target variable represents the presence of diabetes; it is the job of neural network models to correctly predict whether a patient has diabetes based on the information given from the 21 features.



	Attribute	Description
1	**Diabetes_binary**	Does patient have diabetes? (**TARGET VARIABLE**)
2	HighBP	Is patient's **blood pressure** high?
3	HighChol	Is patient's **cholesterol** high?
4	CholCheck	Did patient get a **cholesterol check** in the past 5 years?
5	BMI	What is the patient's **body mass index**?
6	Smoker	Does patient **smoke**?
7	Stroke	Did patient have a **stroke**?
8	HeartDiseaseorAttack	Did patient ever have a **heart attack**?
9	PhysActivity	Did patient **exercise** in the last 30 days?
10	Fruits	Does patient regularly eat **fruit**?
11	Veggies	Does patient regularly eat **vegetables**?
12	HvyAlcoholConsump	Does patient drink a lot of **alcohol**?
13	AnyHealthcare	Does patient have **health coverage?**
14	NoDocbcCost	Did patient need to **see a doctor** but couldn't?
15	GenHlth	How is patient's **overall health**? (1 to 5)
16	MentHlth	How many days (0-30) did patient have poor **mental health**?
17	PhysHlth	How many days (0-30) did patient have poor **physical health**?
18	DiffWalk	Does patient struggle with **climbing up stairs**?
19	Sex	What is patient's **sex**?
20	Age	What is patient's **age**?
21	Education	What is patient's **education**?
22	Income	What is patient's **income**?

Figure 2: A complete list of features from the ”Diabetes Health Indicators Dataset,” each with descriptions[6]. **Diabetes_binary** serves as the target variable that a model needs to predict.

Out of the 21 features, 14 of them contain binary data, while 7 of them contain continuous data. The features' distributions are visualized in Figure 3 and Figure 4 below:

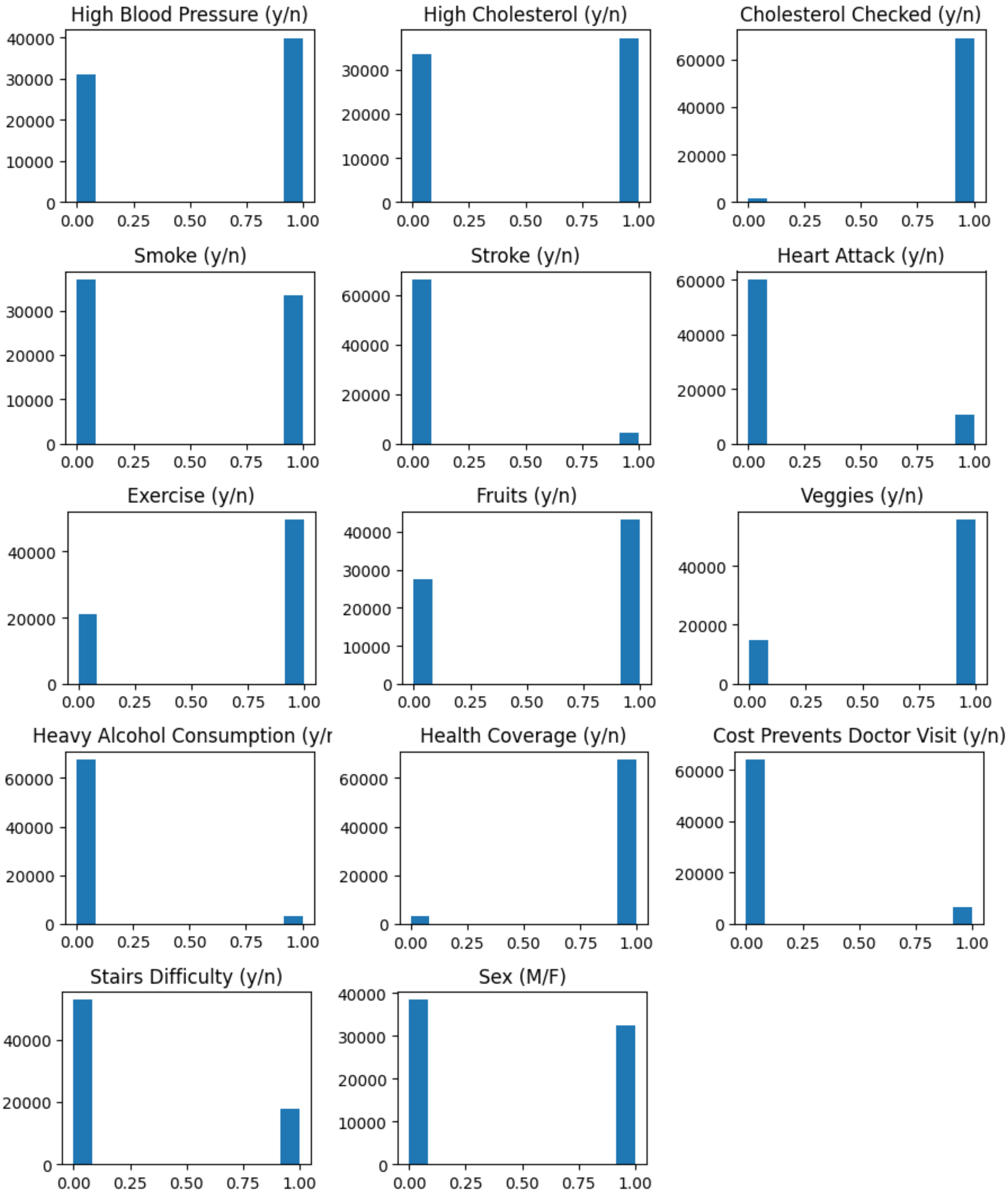


Figure 3: Histograms visualizing the distribution of the 14 features from the "Diabetes Health Indicators Dataset." These particular features are *binary*, meaning that **0** resembles *no* and **1** resembles *yes*. For the **Sex (M/F)** feature, 0 represents "female" while 1 represents "male." The features' names were changed slightly from Figure 2's names to provide a cleaner data visualization.

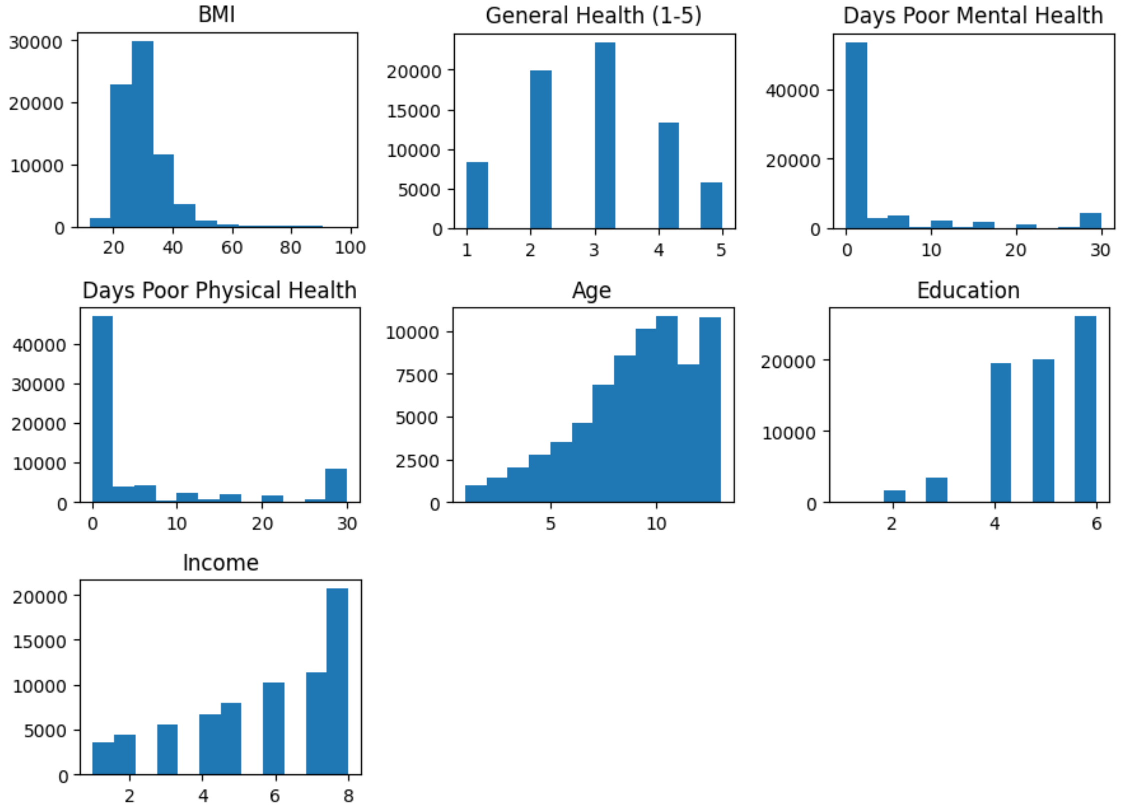


Figure 4: Histograms visualizing the distribution of 7 more features from the "Diabetes Health Indicators Dataset." Unlike the 14 features in Figure 3, these histograms consist of more than two values.

The last three features from Figure 4–Age, Education, and Income–contain some seemingly nonsensical values. However, the values of Age (1 through 13), Education (1 through 6), and Income (1 through 8) are categorical. The specific categorizations of these feature quantities can be observed by navigating to these websites: *ResearchGate*[7] for the **Age** categorization, the *National Addiction & HIV Archive Program*[8] website for the **Education** classification, and the *Resource Center for Minority Data*[9] website for the **Income** categorization. These websites are located in the References section under Numbers 7, 8, and 9.

Data Normalization

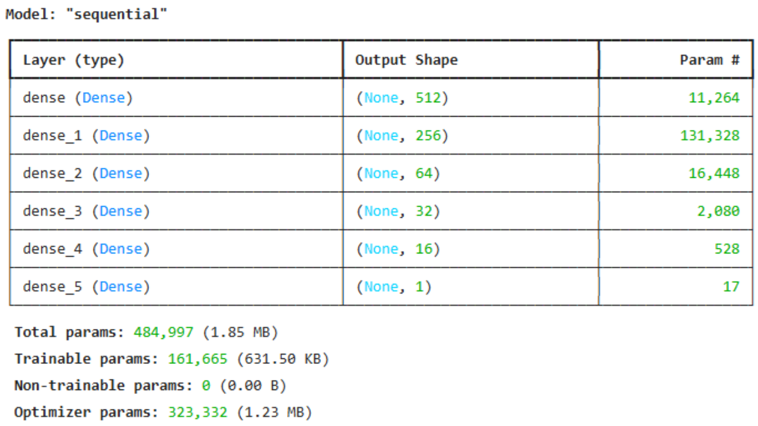
The 21 features underwent a normalization calculation to facilitate the performance for all potential models. The target variable **Diabetes_binary** was not included in the calculation due to its variables already containing binary values. The calculation for each feature's value was conducted by using a *z-score* calculation, which is represented like this:

$$z - score = \frac{x - \mu}{\sigma}$$

where x is a feature's current value inside a row, μ is the standard mean of the feature, and σ is the feature's standard deviation. The *z-score* serves as the result that replaces the raw value inside the feature's instance. This calculation is performed iteratively for every value inside of each feature. This process reduces the minimum-maximum range for each of the 21 features, which will tentatively allow each model to perform more quickly with more robust metrics.

3 Overfitting the Data

The data overfitting process is crucial for ensuring a neural network model can effectively memorize the data. If a model can accomplish this task, then it should eventually perform reasonably well with a training/validation split of the data. A large neural network with a **512-256-64-32-16-1 architecture** served as the ideal model for overfitting the data. The network was implemented over the course of 400 epochs due to the large dataset size.



```
Model: "sequential"
┌─────────────────────────────────┬────────────────────────┬───────────────┐
│ Layer (type)                    │ Output Shape           │       Param # │
├─────────────────────────────────┼────────────────────────┼───────────────┤
│ dense (Dense)                   │ (None, 512)            │        11,264 │
├─────────────────────────────────┼────────────────────────┼───────────────┤
│ dense_1 (Dense)                 │ (None, 256)            │       131,328 │
├─────────────────────────────────┼────────────────────────┼───────────────┤
│ dense_2 (Dense)                 │ (None, 64)             │        16,448 │
├─────────────────────────────────┼────────────────────────┼───────────────┤
│ dense_3 (Dense)                 │ (None, 32)             │         2,080 │
├─────────────────────────────────┼────────────────────────┼───────────────┤
│ dense_4 (Dense)                 │ (None, 16)             │           528 │
├─────────────────────────────────┼────────────────────────┼───────────────┤
│ dense_5 (Dense)                 │ (None, 1)              │            17 │
└─────────────────────────────────┴────────────────────────┴───────────────┘
 Total params: 484,997 (1.85 MB)
 Trainable params: 161,665 (631.50 KB)
 Non-trainable params: 0 (0.00 B)
 Optimizer params: 323,332 (1.23 MB)
```

Figure 5: A summary of the **512-256-64-32-16-1** model architecture. This model served as an ideal candidate for overfitting the ”Diabetes Health Indicators Dataset.”

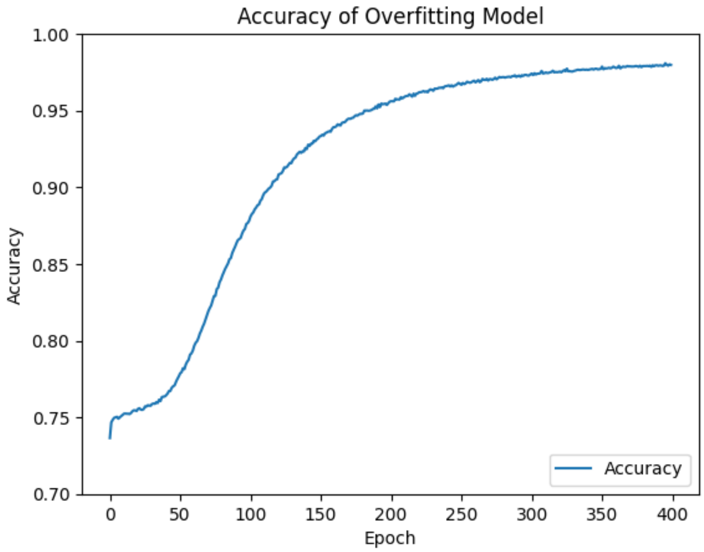


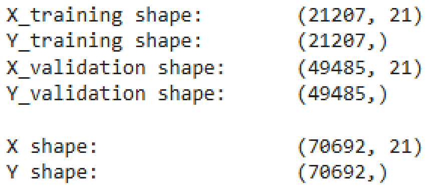
Figure 6: A graph illustrating the accuracy of the overfitting model over 400 epochs.

The model appeared to struggle with memorizing the dataset for the first 50 epochs. Its performance significantly improved between epochs 50 and 100 since Figure 6’s chart encountered an upward concavity. The rate in which accuracy improves started decreasing at around 100 epochs, then continued to gradually diminish until it successfully completed all 400 epochs. Overfitting did not begin to occur until around the last 50 epochs of training. Since more than 70,000 samples exist in the data, it is rather challenging to overfit the data regardless of the neural network’s configuration.

4 Data Separation into Training and Validation Sets

Separation of data served as the next phase of creating a model that can predict the presence of diabetes. The ”Diabetes Health Indicators Dataset” was divided into two additional sets for training and validation. Separating the dataset was essential towards ensuring the efficiency and accuracy of finding the ideal classification model.

The number of rows (or samples) in the training and validation sets are smaller than in the original set. This scenario occurs due to **70%** of the original data's placement into the *training set*, and the remaining **30%** into the *validation set*. That is why 49,485 rows of data are used for training and 21,207 rows are used for validation purposes. This separation process is illustrated graphically in Figure 7.



```
X_training shape:         (21207, 21)
Y_training shape:         (21207,)
X_validation shape:       (49485, 21)
Y_validation shape:       (49485,)

X shape:                  (70692, 21)
Y shape:                  (70692,)
```

Figure 7: A depiction of the shapes of the training and validation sets for the input features X and the output target Y. Each coordinate represents the number of rows and columns in a separated dataset respectively. NOTE: the target Y contains missing second values, meaning that only one column exists for each target variable.

5 Random Baseline Classifier

The random baseline classifier is a concept that involves comparing the effectiveness of neural network models. These models are compared to benchmark values that dictate the minimum performance that a model should achieve to be useful. A minimum accuracy metric can serve as an example of such a benchmark value. In this particular study, the random baseline classifier relies on the percent distribution of samples in each class. Since neural networks in this study attempt to solve a binary classification problem, the baseline relates to the percentage of samples in the *larger* class. For instance, if 70% of the data samples belong to the larger binary class, then the random baseline classifier accuracy would be 70%.

As mentioned in the Data Exploration section, the two classes–Diabetic/Pre-diabetic and Non-diabetic–both contain exactly 50.0% of the dataset's samples. Since no class is larger than the other class, it is safe to say that the **random baseline classifier** metric will be **50.0%**.

6 Model Selection and Evaluation

6.1 Simple Logistic Regression Model

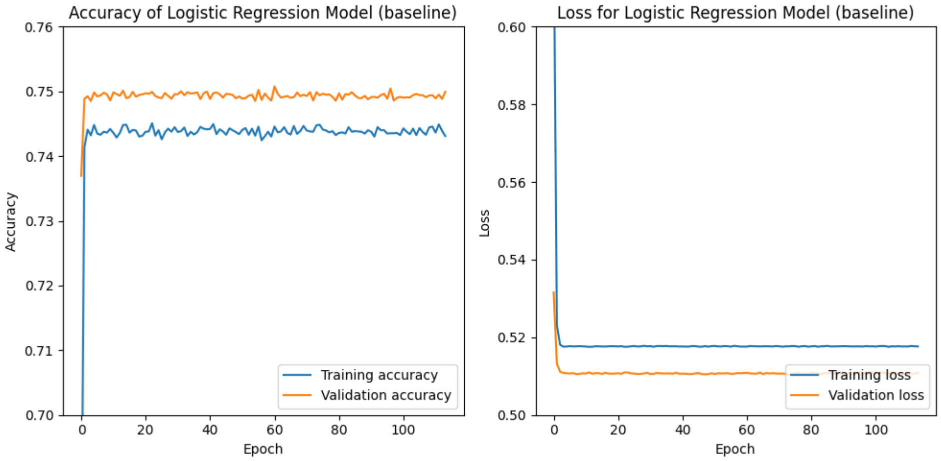


Figure 8:

6.2 5-layer Neural Network

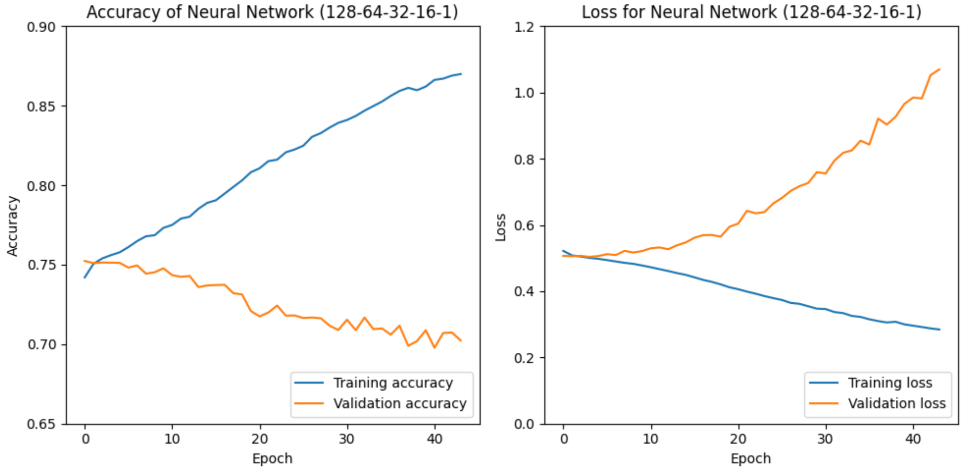


Figure 9:

6.3 4-layer Neural Network

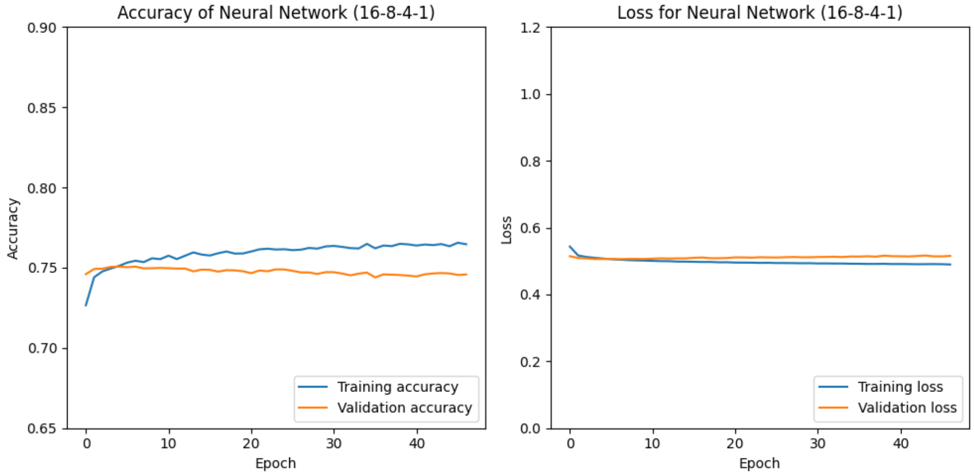


Figure 10:

6.4 3-layer Neural Network

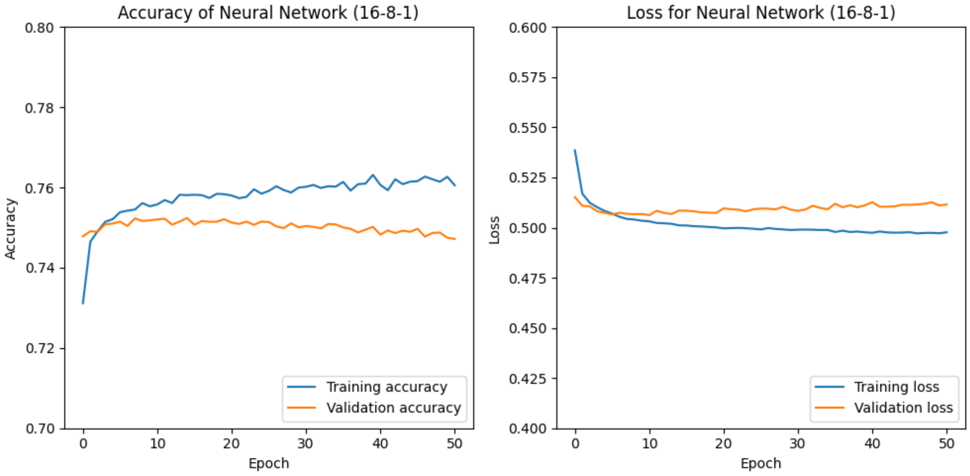


Figure 11:

6.5 2-layer Neural Network

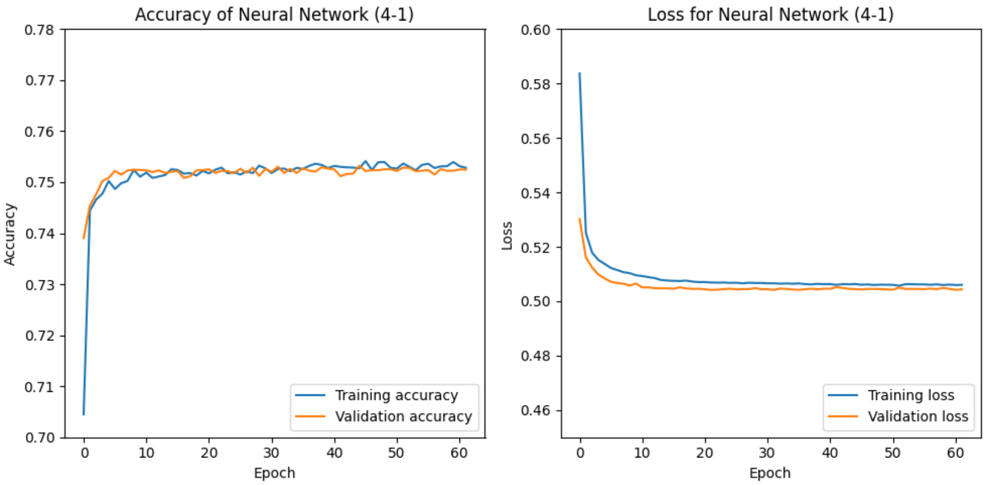


Figure 12:

7 References

[1] American Diabetes Association. (2024). *Statistics About Diabetes.* Retrieved November 18, 2024, from https://diabetes.org/about-diabetes/statistics/about-diabetes

[2] National Institute of Diabetes and Digestive and Kidney Diseases. (2023). *What is diabetes?.* U.S. Department of Health and Human Services. Retrieved November 18, 2024, from https://www.niddk.nih.gov/health-information/diabetes/overview/what-is-diabetes

[3] World Health Organization. (2024, November 14). *Diabetes.* U.S. Retrieved November 18, 2024, from https://www.who.int/news-room/fact-sheets/detail/diabetes#:~:text=Diabetes%20is%20a%20chronic%20disease,hormone%20that%20regulates%20blood%20glucose.

[4] Teboul, A. (2021, November 8). *Diabetes Health Indicators Dataset* [Data set]. Kaggle. Retrieved November 18, 2024, from https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset

[5] *CDC Diabetes Health Indicators.* (n.d.). UC Irvine Machine Learning Repository. Retrieved November 18, 2024, from https://archive.ics.uci.edu/dataset/891/cdc+diabetes+health+indicators

[6] Teboul, A. (2021, November 8). *Diabetes Health Indicators Dataset Notebook* [Jupyter notebook]. Kaggle. Retrieved November 18, 2024, from https://www.kaggle.com/code/alexteboul/diabetes-health-indicators-dataset-notebook

[7] Unknown author. (n.d.). *Variable AGE5YR: Fourteen-level age category (20)* [Image]. ResearchGate. Retrieved November 19, 2024, from https://www.researchgate.net/figure/Variable-AGE5YR-Fourteen-Level-Age-Category-20_tbl3_340098871

[8] The National Addiction & HIV Data Archive Program. (n.d.). *EDUCA (Education level completed)* [Variable]. Regents of the University of Michigan. Retrieved November 19, 2024, from https://www.icpsr.umich.edu/web/NAHDAP/studies/34085/datasets/0001/variables/EDUCA?archive=NAHDAP

[9] Resource Center for Minority Data (n.d.). *INCOME2: Income Level* [Variable]. Regents of the University of Michigan. Retrieved November 19, 2024, from https://www.icpsr.umich.edu/web/RCMD/studies/34085/datasets/0001/variables/INCOME2?archive=RCMD