Predicting Diabetes with Deep Learning

STAT 479 Midterm Group Project

Luis Labra, Shane Grothe, Sonny, Marian Lu

1 Dataset

The dataset we explore can be found in Kaggle. The data set is a selection of survey responses from The 2015 Behavioral Risk Factor Surveillance System (BRFSS), a collaborative project between all of the states in the United States and several territories, surveys the adult population (age 18 and older) and is designed to measure behavioral risk factors and preventative health practices. (“Behavioral Risk Factor Surveillance System Home”, n.d.)

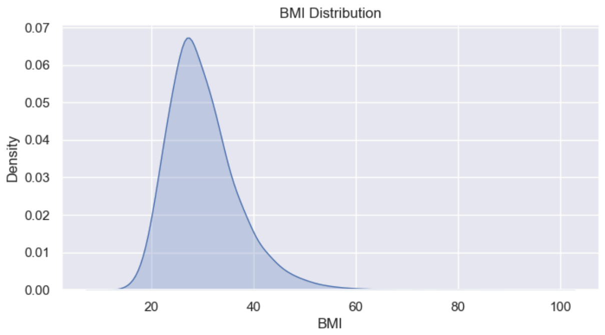
Our data set consists of 70692 survey responses from a cleaned BRFSS 2015. There are 17 feature variables and 1 target variable. The target variable is “Diabetes,” the indicator variable for whether a patient has diabetes (0 = no diabetes, 1 = diabetes) with feature variables described below.

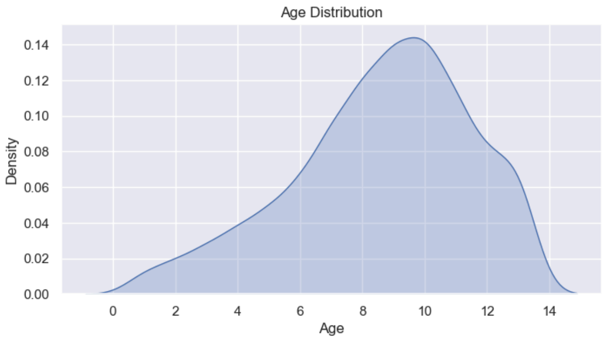
- Age: 13-level age category (_AGEG5YR see codebook) 1 = 18-24 9 = 60-64 13 = 80 or older
- Sex: patient's gender (1: male; 0: female).
- HighChol: 0 = no high cholesterol 1 = high cholesterol
- CholCheck: 0 = no cholesterol check in 5 years 1 = yes cholesterol check in 5 years
- BMI: 0 = no cholesterol check in 5 years 1 = yes cholesterol check in 5 years
- Smoker: Have you smoked at least 100 cigarettes in your entire life? [Note: 5 packs = 100 cigarettes] 0 = no 1 = yes
- HeartDiseasorAttack: coronary heart disease (CHD) or myocardial infarction (MI) 0 = no 1 = yes
- PhysActivity: physical acti
- vity in past 30 days - not including job 0 = no 1 = yes
- Fruits: Consume Fruit 1 or more times per day 0 = no 1 = yes
- Veggies: Consume Vegetables 1 or more times per day 0 = no 1 = yes
- HvyAlcoholConsump: (adult men >=14 drinks per week and adult women>=7 drinks per week) 0 = no 1 = yes
- GenHlth: Would you say that in general your health is: scale 1-5 1 = excellent 2 = very good 3 = good 4 = fair 5 = poor
- MentHlth: days of poor mental health scale 1-30 days
- PhysHlth: physical illness or injury days in past 30 days scale 1-30
- DiffWalk: Do you have serious difficulty walking or climbing stairs? 0 = no 1 = yes
- Stroke: you ever had a stroke. 0 = no, 1 = yes
- HighBP: 0 = no high, BP 1 = high BP

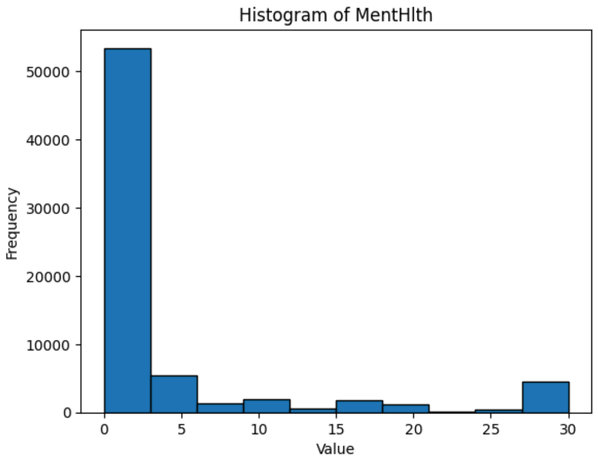
It is important to note that even variables which are quantitative, such as blood pressure or cholesterol level, are categorized as high vs. low. This is certainly a limitation of the data collection process which was via interviews and did not constitute a genuine health check with precise readings.Even age of each participant is recorded as a category rather than actual age.

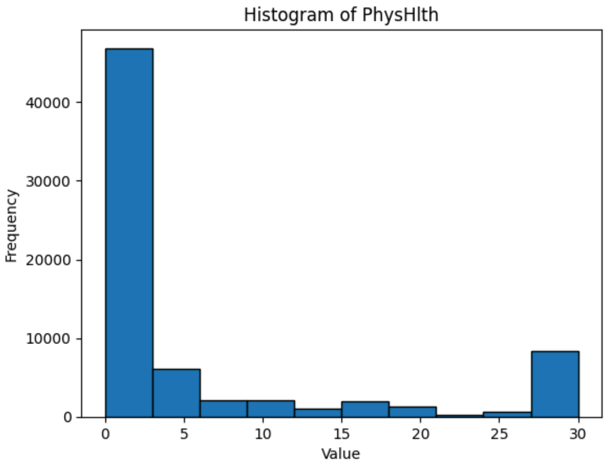
2 Exploratory Data Analysis and Preprocessing

Of important note is the large number of categorical variables since the data only contains the quantitative variables ‘BMI,’ ‘MentHlth,’ and ‘PhysHlth,’ the latter quantifying the number of poor mental and physical health, respectively. BMI, shown below, exhibits a slightly right skewed set of values.

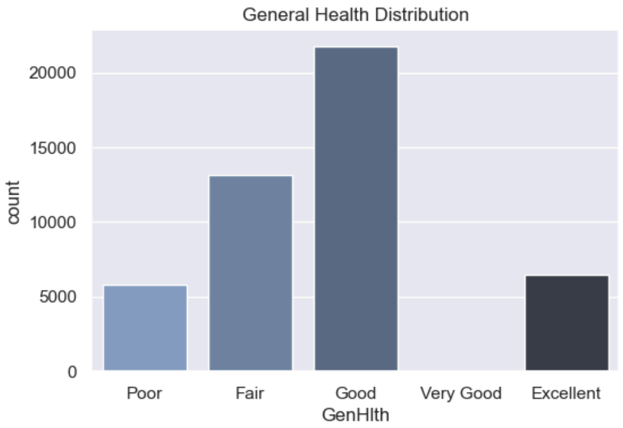






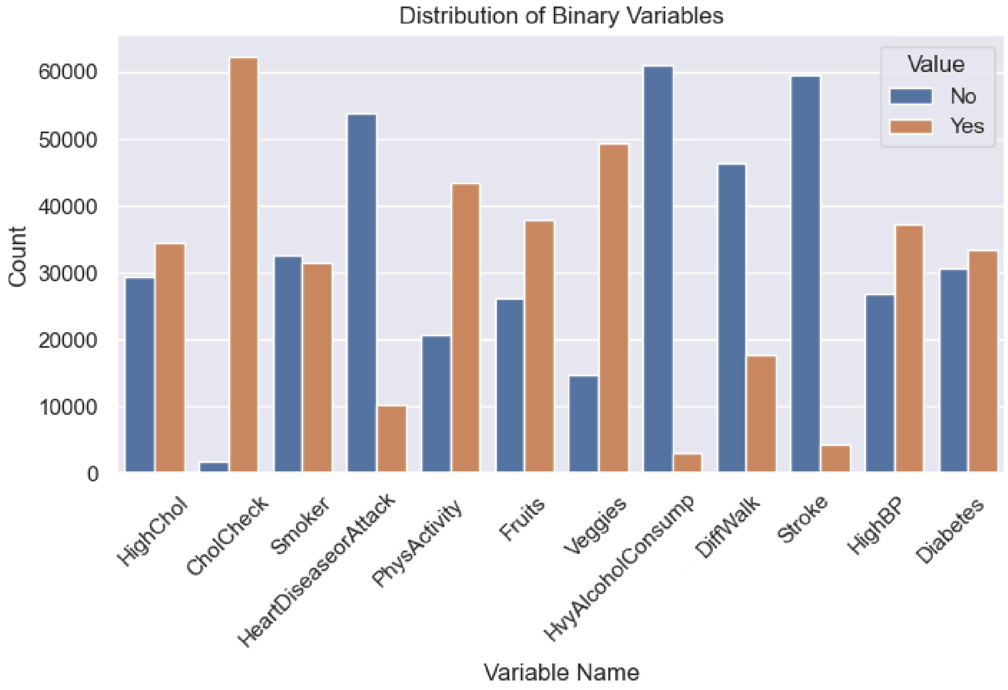


Age is traditionally a genuine meaningful quantitative variable. However, our data set has stratified ages into ranges and is reported in our data set as ordinal data. Since the data is ordinal and ranges are fairly consistent, we will treat the age variable as a genuine quantitative variable. The variables 'MentHlth,' and 'PhysHlth,' show an irregular distribution highly right-skewed as expected with few large responses and many smaller responses. Each of these variables are normalized in the preprocessing step to ensure all features are in similar ranges and no one variable is artificially impacting our Neural Network.



Above we observe one of the only categorical variables encoded via One Hot Encoding.

The binary variables have more variation, with most of them being cases of imbalanced data. However, the target variable, diabetes, is balanced, so it should provide the model with adequate cases of each class for it to make accurate predictions.



The data was checked for and removed of any missing observations and duplicate rows. Since the numerical features, 'BMI' and 'Age,' generally follow a symmetric distribution, they were standardized to ensure that they contribute equally to the data. The categorical variables required one-hot-encoding in order to prepare the data for the model. After preprocessing, the dataset contains 64020 observations.

The data was then split into training and testing sets, with 80% (51216) allocated for training and 20% (6402) allocated for testing.

3 Model Architecture

Model Building

The model for predicting diabetes is a Feedforward Neural Network (FNN). Feedforward neural networks operate on a unidirectional flow, meaning that the information flows in only one direction. Data from an external environment, in this case our data set, enters the network through the input layer, gets passed through the hidden layers, and is sent as a final processed value by the output layer. The output layer will predict the probability that an observed sample is an individual with diabetes.

The initial model has the following architecture and parameters:

Input Layer	Hidden Layer	Output Layer	Parameters
Neurons: 64 Activation: ReLu	Neurons: 64 Activation: ReLu	Neurons: 1 Activation: Sigmoid	Learning rate: 0.001 Batch size: 32 Epochs: 100

Because we are dealing with a binary target variable, the activation function of the output layer is a sigmoid function, given by this equation:

$$f(x) = \frac{1}{1+e^{-x}}$$

The input layer and hidden layers use a Rectified Linear Unit, or ReLu, activation function. It outputs the input directly if it is positive, and outputs zero if otherwise. This is appropriate for our data since we are working with positive values.

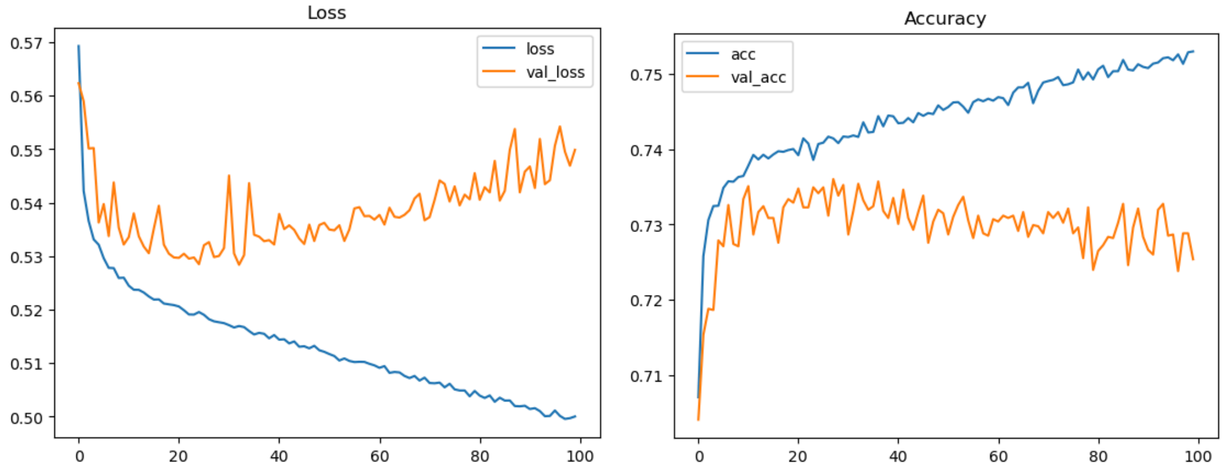
$$\varphi(a(\boldsymbol{x})) = \max(0, a(\boldsymbol{x})) \rightarrow \varphi'(a(\boldsymbol{x})) = \begin{cases} 1, & a(\boldsymbol{x}) > 0 \\ 0, & a(\boldsymbol{x}) \leq 0 \end{cases}$$

Loss during training will be calculated using binary cross-entropy. Binary cross-entropy returns the log probability of the prediction being either of the cases, given by this equation:

$$L =- \frac{1}{N} \sum_{i=1}^{N} [y_i log(\widehat{y_i}) + (1 - y_i) log(1 - \widehat{y_i})]$$

Model Fitting and Evaluation

The loss and accuracy plots after fitting the model:



For the loss plot, there is a large gap between training and validation loss, showing signs of overfitting that the model may not be able to generalize on new data. The accuracy oscillates back and forth in a range of (0.71, 0.75). This sets the starting point for the model's performance prior to hyper parameter tuning.

4 Hyper Parameter Tuning

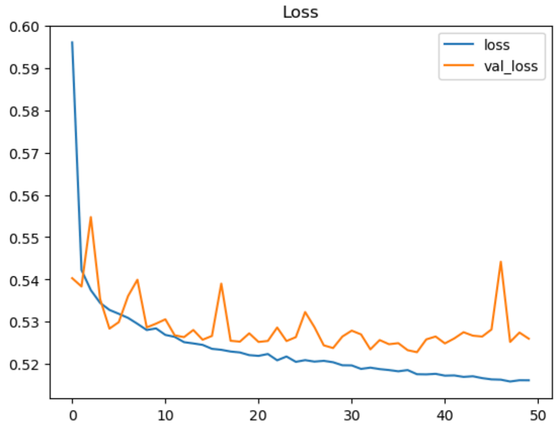
Parameter Tuning

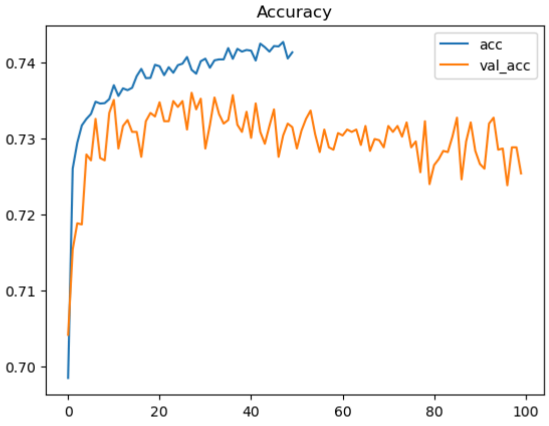
A model-building function was built utilizing the keras_tuner library to tune the hyper parameters. The parameters of interest were number of neurons per hidden layer and the learning rate.

The number of neurons were tested on a range of (32, 128). The learning rate was tested on three values: [1e-2, 1e-3, 1e-4]. The batch sizes were tested on the values [8, 16, 32]. These tests were performed over 10 epochs.

The model performed similarly in the range of 32 to 96 units for the input and hidden layer, so we selected 32 due to its efficiency. The optimal learning rate was 0.001 and the optimal batch size was 32. Since the initial model's performance stagnated before it reached the maximum number of epochs, we ran the adjusted model on 50 epochs. The adjusted model has the following architecture:

Input Layer	Hidden Layer	Output Layer	Parameters
Neurons: 32 Activation: ReLu	Neurons: 32 Activation: ReLu	Neurons: 1 Activation: Sigmoid	Learning rate: 0.001 Batch size: 32 Epochs: 50





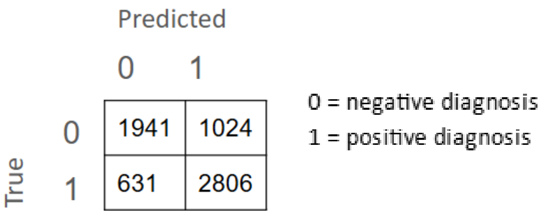
After re-fitting the model with these adjustments, the loss and accuracy plots showed favorable results in the model's performance. The divergence between the training and validation loss curves was reduced significantly, indicating that the model improved its ability to generalize on new data. The accuracy remained in the same range as the initial model, with the curves behaving similarly.

Additional Methods

In addition to optimizing the hyperparameters, we also experimented with different activation functions. As an alternative to the ReLu function, the Leaky ReLu function and Swish function were implemented to see if it had any impact on the model's performance. However, the model performed similarly despite these changes so the final model still utilizes the ReLu function.

5 Conclusion

The final model was able to predict if a patient had diabetes or not with an accuracy of around 74.3%. Its training loss ended up approaching 0.50. As a predictive classifier, its performance could leave more to be desired. While its training loss vs. validation loss curves show that it is able to decently generalize on new data, its accuracy is still quite low.



		Predicted	
		0	1
True	0	1941	1024
	1	631	2806

0 = negative diagnosis
1 = positive diagnosis

A confusion matrix reveals that the model's predictions skews towards a positive case of diabetes. In the context of this problem, this is not the worst case scenario. A disease prediction model that predicts more positives is more effective than one that predicts more false negatives because it reduces the risk of undiagnosed cases.

Some strategies that could have improved the model include feature engineering and experimenting with more layers. We used every single feature in the original dataset, which could have led to overfitting. Deeper understanding of how much each feature is actually contributing to the prediction could help us be more deliberate with our feature selection. We kept the model to three layers (one input, one hidden, and one output) because it only had to handle a simple classification task. However, while multiple hidden layers are usually more productive for models that have to make complex calculations, adding more hidden layers may change the way our model interacts with the data and thus help it make better predictions. Overall, the model's predictive ability is passable but can be further improved.

References

Centers for Disease Control and Prevention. (n.d.). *Behavioral Risk Factor Surveillance System (BRFSS)*. Retrieved November 8, 2024, from https://www.cdc.gov/brfss/index.html