**Diabetes Classification Using a Neural Network Model**

**Project Report**

CS 422 1002

Members:

David Lam

John Knapp

Kaitlynn Pham

**Member Contributions:**

* Kaitlynn: Dataset Selection, Report Writing, Graphs, Correlation, Interpretations
* David: Driver Code Creation, Optimization, Report Writing, Interpretations.
* John: Optimization, Secondary Model Creation, Testing, Report Writing, Interpretations.

**Dataset:**

* <https://www.archive.ics.uci.edu/dataset/891/cdc+diabetes+health+indicators>

**Libraries:**

* Pandas
* Torch
* Sklearn
* Matplotlib

**Project Overview**

For this project, we decided to tackle a major problem in healthcare. Is it possible to predict whether a given person has diabetes based on a number of health, lifestyle, and socioeconomic factors? By utilizing machine learning techniques and building multiple models, we have concluded that we are able to reliably predict whether an individual has diabetes based on the aforementioned factors. The dataset that we used classifies individuals into three categories, diabetes, pre-diabetes, or healthy, and it also provides a number of features for these individuals. By providing accurate predictions, individuals will be able to get early treatment and help prevent the disease from progressing any further; additionally, it can provide insight into what factors may affect how likely someone is to get diabetes.

**Computational Challenges**

The main computational challenges for this project involved creating the model architecture and tuning hyperparameters. Our dataset is relatively large, but since it does not include images, training was relatively quick, and processing power was not a challenge for us. For the architecture, it was difficult to determine how many hidden layers and nodes to use because it required a lot of trial and error to find the optimal values. We settled on three layers with a gradual reduction in the number of nodes per layer after experimenting with other configurations. This design was found to result in the greatest performance. Additionally, we had to choose which activation function and cost function should be used. We tried quite a few activation functions and found that RELU provided the greatest performance overall. For cost functions, we designed the NN to use the Cross Entropy Loss function. We tried a handful of other, simpler cost functions and observed inferior performance. Regarding the values in the dataset, the range varies from feature to feature with many that are binary values and a couple that range from 0 to 40. We applied normalization to the dataset, to address this issue and expected an increase in accuracy. However, the opposite was true and a drop of 2% was observed. Regardless, we believe that normalization is important for the validity of the results, so the normalization was left in.

**Dataset**

The dataset that we chose for our project came from the CDC (Center of Disease Control) from a study on the relationship between diabetes and lifestyle. It has 21 features and over 250k samples. Most of the features are either binary values or integers, and there are no missing values from the dataset. The features broadly fit into one of three categories: health characteristics, lifestyle choices, and socioeconomic status. The general health features include factors such as level of activity, sex, BMI, general health, and mental health. Some examples from the lifestyle features are diet(fruits and vegetables), alcohol consumption, and smoking. Finally, some socioeconomic status features are education, income, and no doctor because of cost. This dataset was created by the CDC via survey. It is worth noting that in regards to the target value (diabetes) there is an overwhelming proportion of healthy participants which resulted in a biased dataset.

Our dataset contains 3 different data files and we created a menu for the user to select which dataset they would like to use. Each file contains the same 21 features. The “Diabetes Health Data'' has 253,680 points. The target variable for this one is 0 for no diabetes and 1 for pre-diabetes or diabetes. The second file, “Diabetes 50/50 Split'' contains 70,692 data points, and as the name suggests, it has an even distribution of healthy and diabetic respondents. It also uses 0 or 1 for the target variable. The final file, “Diabetes, Yes, No, or Pre-Diabetes” contains 253,680 data points like the first file. However, this dataset differentiates between diabetes and pre-diabetes. Accordingly, it uses 0 (no diabetes), 1 (pre-diabetes) , and 2 (diabetes) as values for the target variable.

**Machine Learning Techniques**

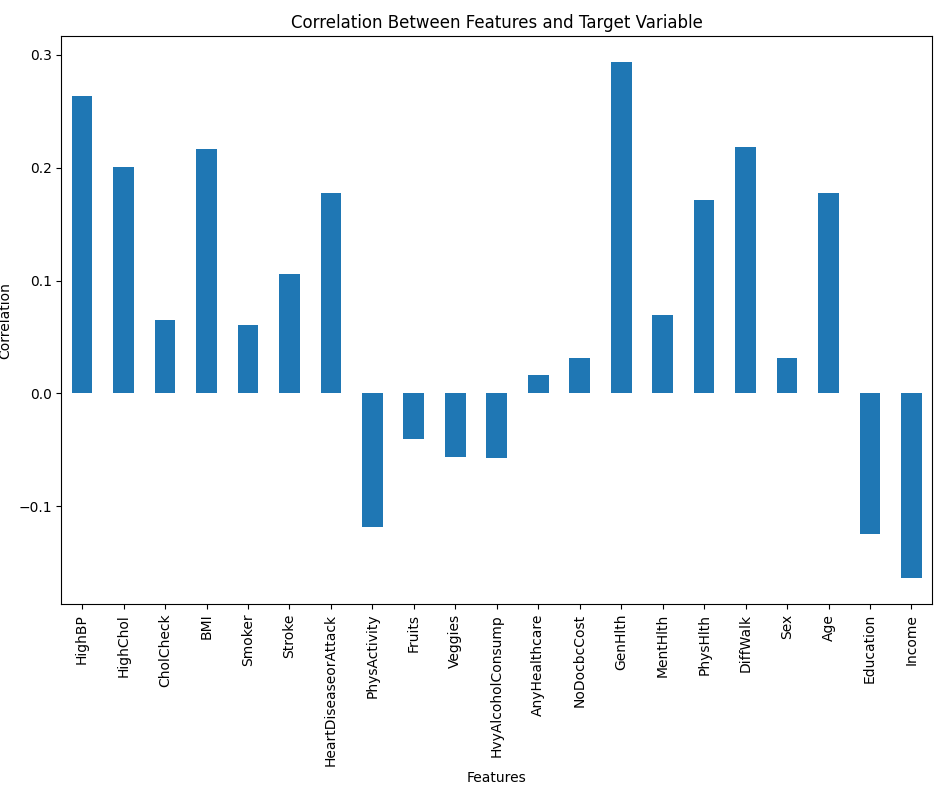
For our model we considered several techniques, but we ultimately decided on a neural network for the primary model. The dataset that we are working with is very large and the features have many complicated relationships to and with each other. Accordingly, we decided that the best way to recognize the many hidden patterns from these datasets was to utilize a Neural Network with a handful of hidden layers. The large dataset would mean the training process would take a little longer but would hopefully provide a better model compared to other potential algorithms. We also decided to implement a linear regression model to see how well a secondary approach would compare to the neural network.

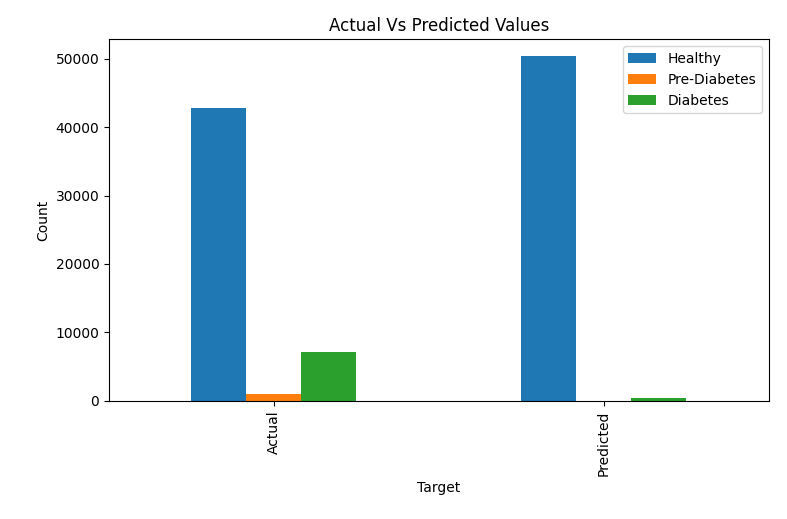
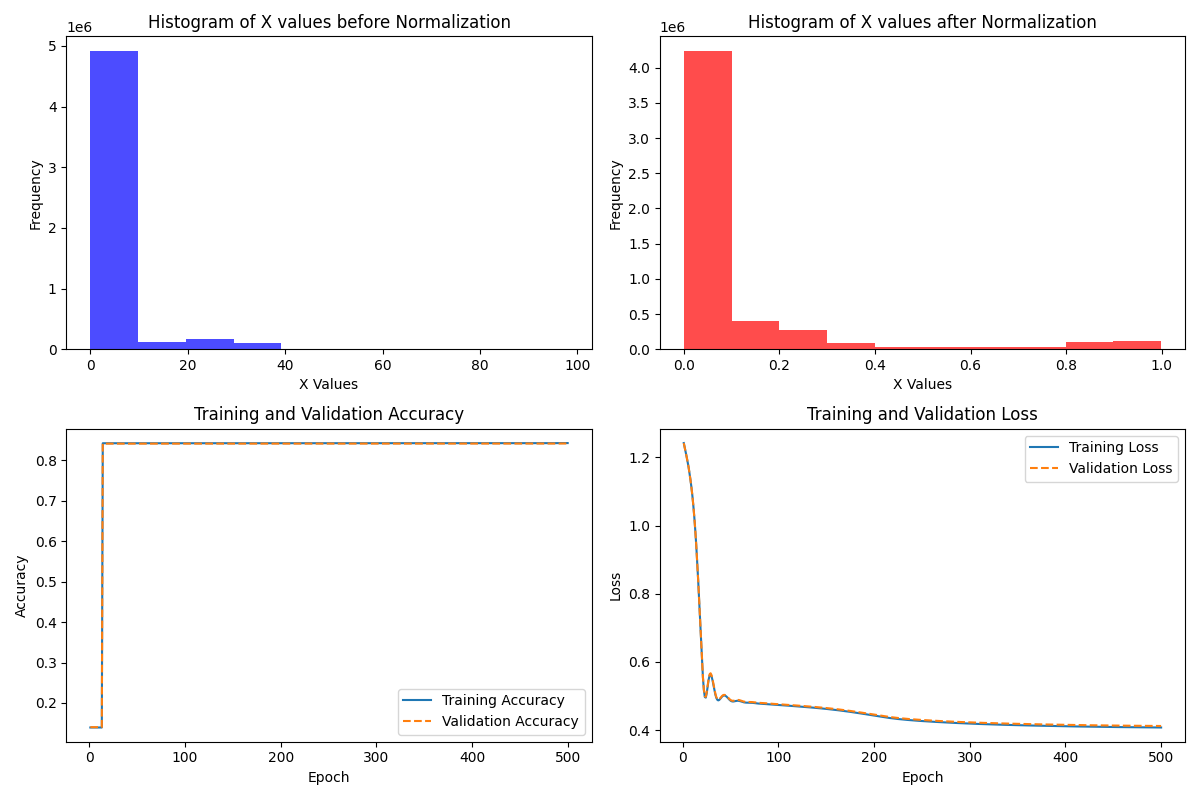
**Experiment Design**

For our primary model, we created a Neural Network which consists of 3 hidden layers utilizing a RELU activation function for each layer. It takes 21 features as input and gives out 2 outputs. We chose to use RELU because we found through testing that other methods have worse or similar performance. We take the data, normalize all features except the target variable and then split the dataset into a training and testing set with 20% of the data reserved for testing. Then, we train the model over 500 epochs. We decided on this value after extensive testing where we found that larger values such as 1,000 or even 10,000 greatly increased computation time with relatively minor performance improvements of around 0.20% on average. For our secondary model which performs linear regression, the model performs some basic data processing after loading in the data. Then it performs the linear regression with 100 fold cross validation before outputting the RMSE every tenth fold. After calculating RMSE, the predicted values are rounded to the nearest integer so that accuracy can be approximately measured.

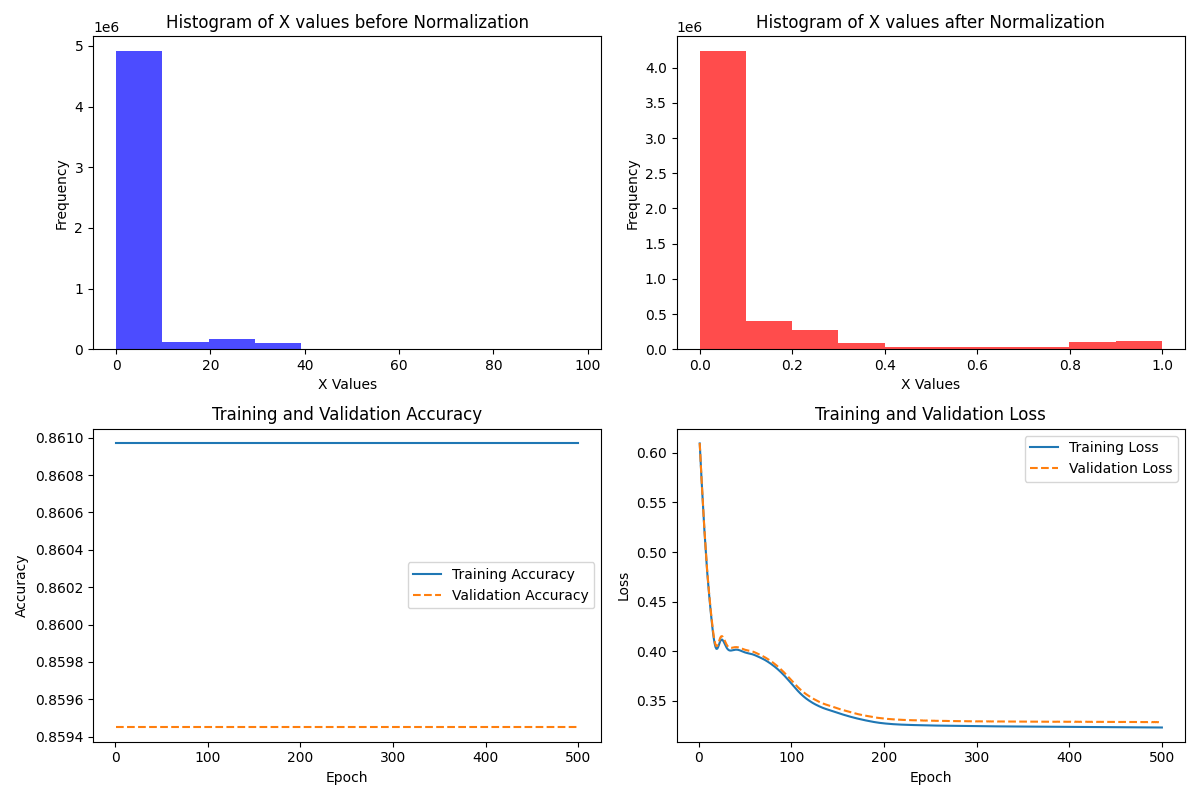
**Results**

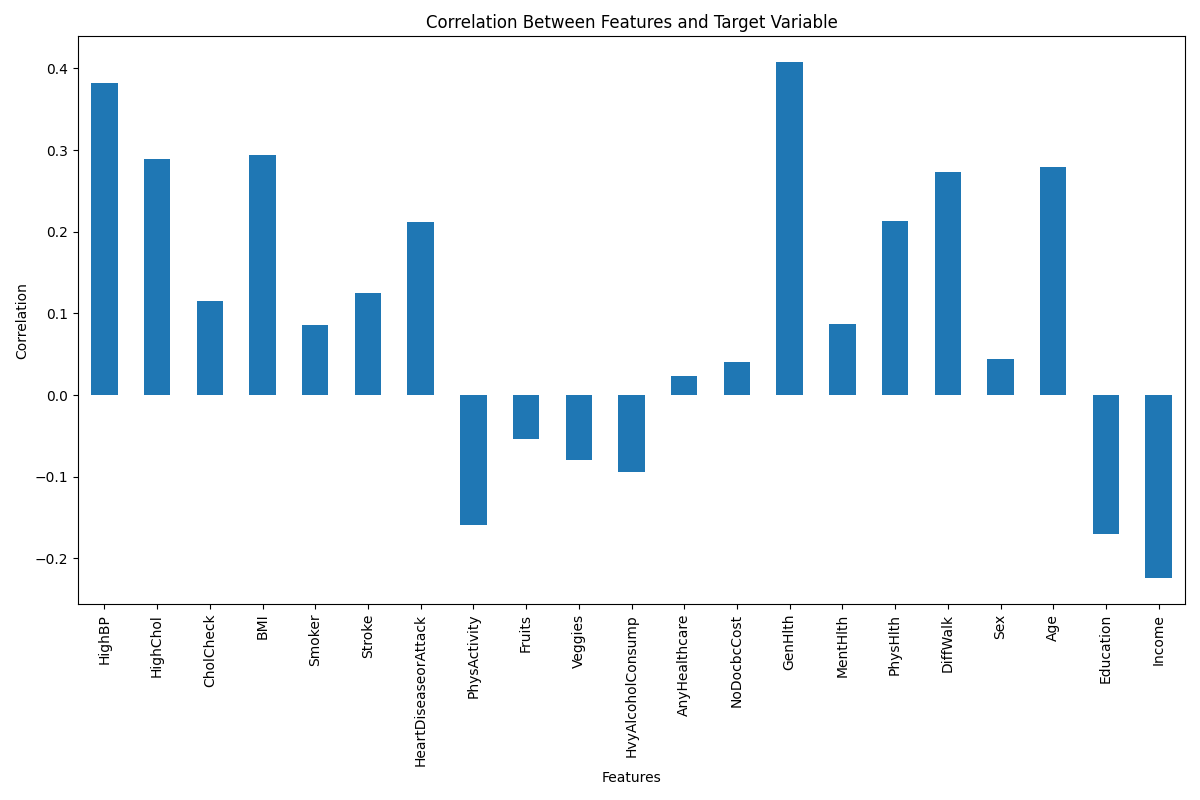
Our neural network model is able to output an average of 85% for predicting diabetes; however, depending on the dataset, these results may vary. The “50/50 Split” dataset had a lower average of 74%. Looking at how the model performed, we see that it would have a loss of around 0.70 at the beginning of the epochs and this value would slowly drop to around 0.40 by the time it reached 500 epochs. Of course, our model improved as the loss value dropped, however, we noticed that this number does not get much further than 0.40 even with more epochs. In addition, the accuracy scores became stagnant after around 85%. The secondary model was configured to run the “50/50 Split” dataset, and its results were similar to the neural network, but with much larger variance. Since the learning restarts every fold instead of being refined epoch by epoch, this model struggles to compound its learning. This makes it ultimately less useful. Additionally, this model would likely struggle even more with the other two datasets due to their imbalanced nature.

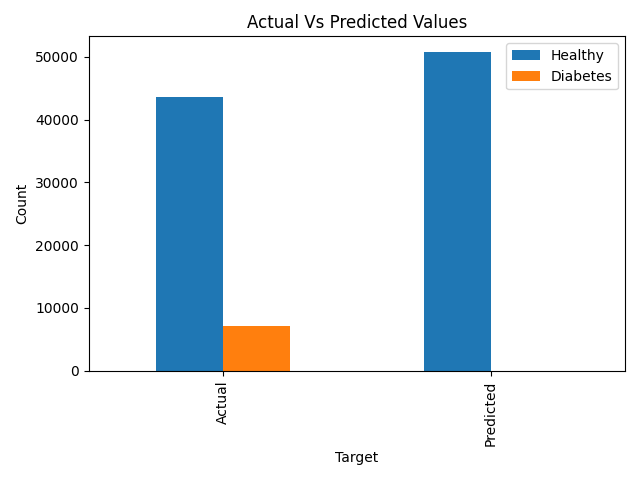
**Dataset : Diabetes, Yes, No, or Pre-Diabetes (diabetes\_012\_health\_indicators.csv)**



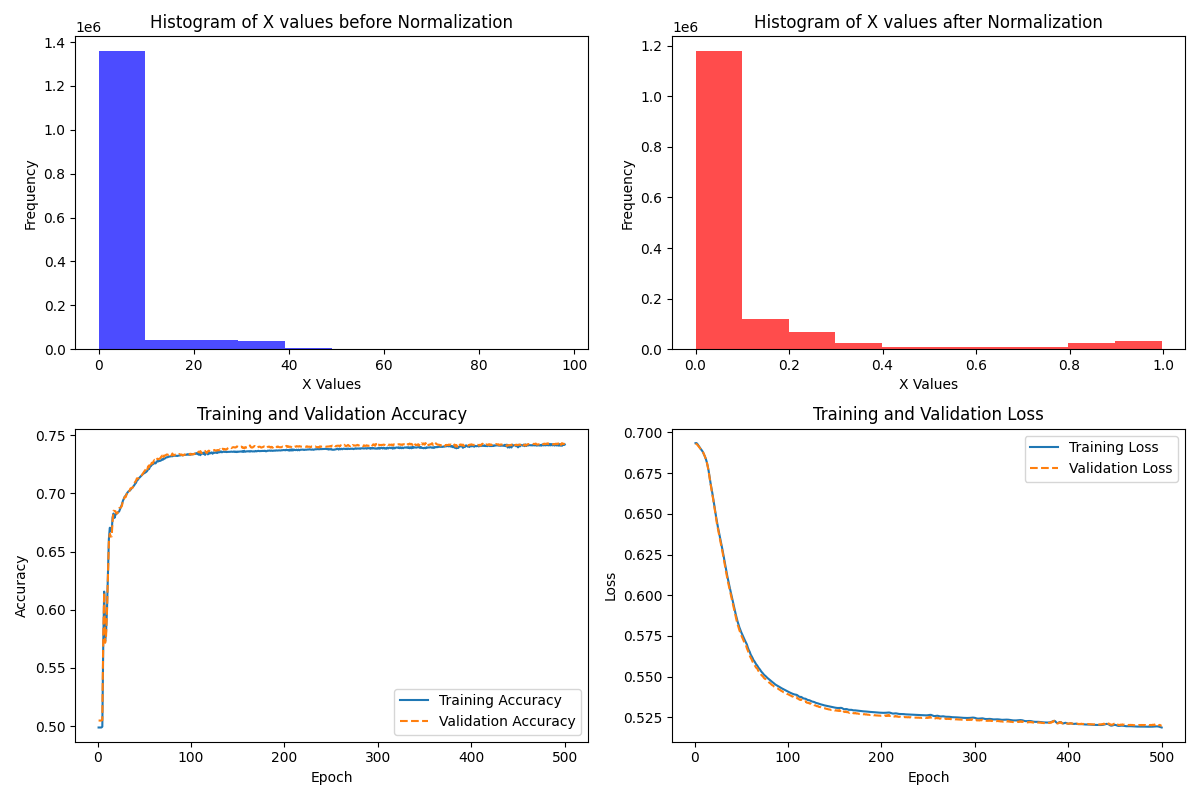
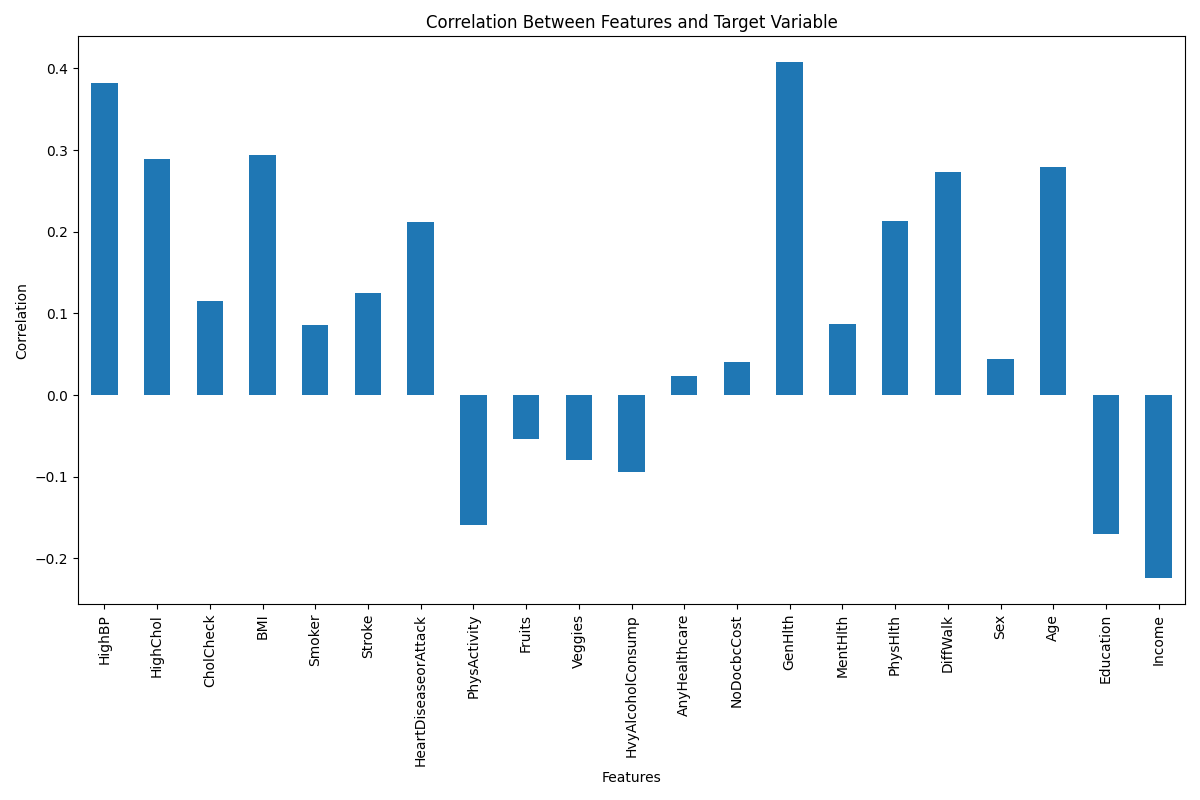
**Dataset : Diabetes Health Data (diabetes\_binary\_health\_indicators.csv)**

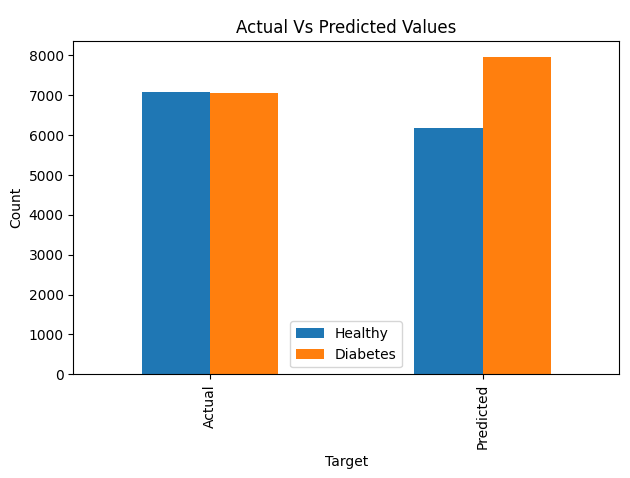






**Dataset : Diabetes 50/50 Split (diabetes\_binary\_5050split\_health\_indicators\_.csv)**



****

**Interpretations of Graphs**

According to the correlation graphs, we can see several factors that contribute the most to the risk of diabetes. These include strong positive correlations for general health, high blood pressure, BMI, and whether the individual has difficulty walking or climbing stairs. There were also strong negative correlations with education, income, and physical activity. Factors with the least correlation and accordingly, the smallest predictive power include whether or not the individual has healthcare, whether they could not see a doctor recently because of cost, and what their sex is. Additionally, according to the graph, consumption of fruits, vegetables, and alcohol do not appear to play a big role compared to the other features. This is an interesting observation since we notice that socioeconomic factors such as income and education play a larger role in diabetes prediction than do some traditional health values such as eating healthy.

**Conclusion and Applications**

Overall, we have found that it is possible to predict whether an individual has diabetes to a reasonable degree of accuracy with some knowledge of their health, lifestyle, and socioeconomic status. While the accuracy scores from our model are not sufficient for medical diagnosis and would need further improvements, our model can still be used as a supportive tool for preliminary screenings in combination with assessments from healthcare professionals. An initial evaluation can be provided by this model which may give the healthcare profession some more insight into their patient. It can also be used as a research tool for identifying patterns and risk factors that possibly contribute to the disease.