Analyzing the Probability of Diabetes through Machine Learning Models

Anish Subramanian

ABSTRACT:
Numerous diseases impact people around the world on a daily basis, worrying many about their physical health and well-being. Throughout the world, diabetes is one of the most widespread diseases, affecting hundreds of millions of individuals. Machine learning may be able to diagnose patients with diabetes based on their medical records. Using different models such as Logistic Regression, Random Forests, and Neural Networks, we have found that it is possible to predict the probability of having diabetes. The neural network had the highest AUROC (Area Under the Receiver Operating Characteristic curve), AUPRC (Area Under the Precision-Recall Curve), and Accuracy of 0.8330, 0.4316, and 0.8597, respectively, making it the best-performing model out of the three. The results of this paper suggest that machine learning models, specifically neural networks, may be useful in diabetes diagnosis.

1. INTRODUCTION:

Around the world, millions of people living with diabetes must remain constantly alert in managing their physical health, blood sugar levels, and diet. Diabetes is a chronic disease that occurs when the human body does not produce as much insulin as it requires. Insulin is a hormone that regulates blood sugar levels, and when there is not enough insulin being produced, blood sugar levels may increase, leading to diabetes. The prevention of diabetes is imperative as the disease affects nearly 830 million people worldwide, and the number is expected to increase. Diabetes is becoming more prevalent in today's society due to the increase in age and obesity rates [1]. In addition, diabetes can cause numerous other health problems, such as cardiovascular and renal disease [2].

However, in recent years, machine learning and artificial intelligence have had a significant impact on many industries, including healthcare. For example, through the usage of machine learning and artificial intelligence, doctors are able to efficiently complete tasks such as machine-assisted surgery, identify trends in diseases, or create medicines such as vaccines [2]. In addition to aiding with treatment, Artificial Intelligence and Machine Learning models can be used to predict and detect certain illnesses, such as diabetes.

There are many potential factors that could contribute to diabetes, including BMI (Body Mass Index), age, physical health, cholesterol level, etc. However, it is unclear from the data which of these features, among many others, are the most impactful towards diagnosing diabetes. Therefore, determining the factors that are most significantly associated with diabetes is an urgent health need that can be explored through machine learning.

In this study, we apply various machine learning models to determine which features are the most predictive of whether a person has diabetes. We consider a simple logistic regression as our baseline model, and compare its performance to more complex models, specifically

random forests and neural networks. We measure the performance of these models using metrics including accuracy, precision, recall, area under the receiver operating characteristic curve (AUROC), and Area Under the Precision-Recall Curve (AUPRC). Our findings indicate that the neural network performed the most consistently among the three models. It was also found that general health is the most influential feature for diagnosing a patient with diabetes.

2. MATERIALS AND METHODS:

Dataset:

The dataset we use in this paper was downloaded from Kaggle [4]. This diabetes dataset contains medical information on a group of individuals, including Body Mass Index, age, blood pressure, cholesterol levels, smoking status, physical health, and more. The dataset comprises approximately 253,680 samples with 21 features, and a binary target variable labeling whether a patient has diabetes.

Models:

The baseline model we used was logistic regression. Logistic regression is a machine learning algorithm that is commonly used for binary classification [5]. Binary classification predicts one of two outcomes based on the features of the data. In the case of diabetes, the binary classification task is to predict whether an individual has diabetes or not, based on their features. The equation for Logistic regression is:

$$\hat{y} = \sigma\left(\sum_{i=1}^{n} w_i x_i + b\right)$$

In the equation, x represents the features that are inputs to the model, such as age or high blood pressure. The index i in i =1 represents the first feature among n features (total amount). The weights, w, are the coefficients for each feature; b is the bias; sigma (σ) represents the sigmoid function, which will be addressed later; and $\hat{y}$ is the predicted value. The bias is basically the baseline value, and it is the output for the model if every feature value was 0. The bias usually remains the same in the equation. The coefficients of x, (w), represent the correlation strengths of the features. If a w value is positive, that means when the feature value increases, the higher the chance of the output being 1. On the other hand, when there is a negative coefficient, as the feature value increases, the probability of the output being 1 decreases. There also does not have to be only one w value. There are as many w values as there are features in the dataset, from w_1 to w_n. Ultimately, the magnitude of the coefficient shows how strongly the feature influences the prediction. The output $\hat{y}$ is always between zero and one in a binary classification. This is because zero represents the prediction being false and one represents the prediction being true. In this paper, one represents having diabetes and zero

represents not having diabetes, so the closer the output is to one, the more likely one is to have diabetes, and vice versa. Logistic regression is often used as a baseline model in many prediction problems because of its simplicity and interpretability.

The second machine learning model we used was Random Forest. Random Forest is a machine learning algorithm that uses labeled data to build an ensemble of decision trees for classification. A decision tree is a model that splits the data into branches from the feature values, and when graphing it, a structure that looks like a tree appears, hence the name decision tree. In the decision tree, there are boxes called internal nodes, which represent the features of the dataset. From the internal nodes, there are branches that lead to the two decisions of the decision tree, which lead to the outcome of that decision. Decision trees are usually weak and are not often used as predictors because of their performance. Decision trees can often overfit, meaning they learn the patterns of the training data so much that they cannot generalize and perform well on new test data [6]. This is where random forests come into play. Random forests use a collection of weaker individual decision trees, combine them, and make a stronger model. By combining many decision trees, random forests are able to develop a model that generalizes better on new data, making it more reasonable to use than decision trees. Each decision tree in the random forest makes its own prediction based on a subset of the data and features, and the random forest uses all the predictions to make one final prediction, which is the output. Random forest models are great for improving predictive performance while also reducing overfitting.

Finally, we have neural networks. Neural networks are made up of a collection of nodes, also called neurons. During this algorithm, each neuron takes in the input, multiplies it by the weights, adds the bias, passes it through an activation function, and returns an output to the next layer of neurons. An activation function is a mathematical function applied to the input to introduce non-linearity so the model is not just a linear function, and instead it can learn complex patterns in the data and essentially learn from it.

There are many different types of activation functions, and in this study, we used ReLU (Rectified Linear Unit) and sigmoid. Firstly, ReLU is a very common activation function that most people use as a default. The equation for ReLU is:

$$f(x) = \max(0,x)$$

For any positive value input into the ReLU function, ReLU returns it, however, if a negative value is input, the output is just 0. The other function, sigmoid, converts a value into a number in the range between 0 and 1, so it is ideal for binary outputs. The equation of the sigmoid function is as follows:

$$\sigma(x)= 1/ 1+e^{-}x.$$

As the input becomes more negative, the function approaches 0, and as the input is more positive, the output is closer to 1. If the input is 0, then the output would be 0.5. Moving forward, neural networks can change the values of the bias and weights during a process called backpropagation. Backpropagation is repeated multiple times to reduce the loss, or errors, in the model and ultimately make it stronger and more accurate.

Metrics:

To study the performance of each model, we used several classification metrics, including accuracy, precision, recall, AUROC (Area Under the Receiver Operating Characteristic Curve), and AUPRC (Area Under the Precision-Recall Curve). These metrics are defined in terms of the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN).

The formula for accuracy is:

$$\text{Accuracy} = TP + TN / TP + FP + TN + FN.$$

Next, precision is used to measure how many of the predicted positive outputs are actually positive using the equation:

$$\text{Precision} = TP / TP + FP.$$

Recall measures how many of the actual positive outputs of the model are correctly identified. To calculate recall, the formula is:

$$\text{Recall} = TP / TP + FN.$$

Precision and recall are important in medical prediction tasks where false positives and false negatives carry different consequences. If a model labeled someone as non-diabetic, but in reality they do have diabetes, that is an example of a false negative, and this is especially dangerous because if the model predicts the patient wrong, then the patient's diabetes would not be treated. A high recall ensures most true diabetes cases are caught, while high precision ensures that positive predictions are reliable.

After these three performance metrics, we used AUROC and AUPRC. To understand AUROC, we must first learn what ROC is. The ROC (receiver operating characteristic curve), is a useful graph that helps us visualize the sensitivity and specificity of the model for each different value of the threshold. Sensitivity is the true positive rate; specificity is the false positive rate. The graph also helps us evaluate the model's performance, using the area under the curve, or AUC. The closer the AUC of the ROC graph (AUROC) is to 1, the better the model. The closer the AUROC is to 0.5, the worse the model and the more it is just randomly guessing rather than distinguishing between positive and negative classes.

To conclude, the last metric we used was AUPRC. We use AUPRC because it can be more effective than AUROC when assessing models on imbalanced datasets, where false positives are important [7]. The PR (Precision-Recall) graph shows the relationship between the precision and recall at different threshold settings. The PR graph shows the performance of the model in terms of the positive class. AUPRC uses the PR graph to get a single value. The closer the AUPRC value is to 1, the stronger the model is at predicting positive cases without producing too many false positives. However, the closer the AUPRC value is to 0, the worse the model performed. Using these metrics, we can see whether our model is performing as it should or if it is just randomly guessing instead of predicting.

Interpretation methods:

To interpret the factors driving model performance, we utilize feature importance scores for the random forest and Shapley scores for the neural network. Feature importance in the random forest model shows which features the model uses the most to make its prediction. Features that are more helpful for the model have higher importance, which can then be used to see which factors correlate the most with diabetes (Figure 5). For the neural network, the SHAP value similarly checks how much each feature contributes to the prediction of each individual in the dataset. Thus, by averaging these values, we can clearly see which specific qualities increase or decrease the likelihood of diabetes (Figure 6). Therefore, these methods ultimately enable us to see which features have the most effect on the likelihood of developing diabetes.

3. RESULTS AND DISCUSSION:

Model performance:

Model	Mean AUROC	Standard Deviation AUROC	Mean AUPRC	Standard Deviation AUPRC	Mean Accuracy	Standard Deviation Accuracy
Logistic Regression	0.822000	0.003020	0.404875	0.004542	0.863521	0.001365
Random Forest	0.786381	0.002008	0.353479	0.005502	0.857904	0.001004
Neural Network	0.823317	0.002217	0.416635	0.006753	0.865035	0.001070

Table 1: Mean and Standard Deviations for each model's AUROC, AUPRC, and Accuracy score. This table shows the results of the mean AUROC, AUPRC, and Accuracy scores after performing k-fold cross validation. The standard deviations for each score are also listed.

In **Table 1**, we performed k-fold cross validation with a k value of 5 to further evaluate the performance of the models. The method of k-fold cross validation basically trains and tests the model k amount of times (we set k = 5) and provides the evaluation scores (AUROC, AUPRC, and Accuracy scores). We then found the mean and standard deviation of the scores, which can describe which model had the best average performance score. From **Table 1**, we can see that neural networks had the highest mean scores compared to the other 2 models with a mean AUROC score of 0.823317, mean AUPRC score of 0.416635, and a mean Accuracy score of 0.865. Although the difference in effect sizes between the models are modest, the narrow standard deviations suggest that these differences are consistent and robust.

Model Evaluation Summary:
Logistic Regression - AUROC: 0.8264
Random Forest - AUROC: 0.7357
Neural Network - AUROC: 0.8330

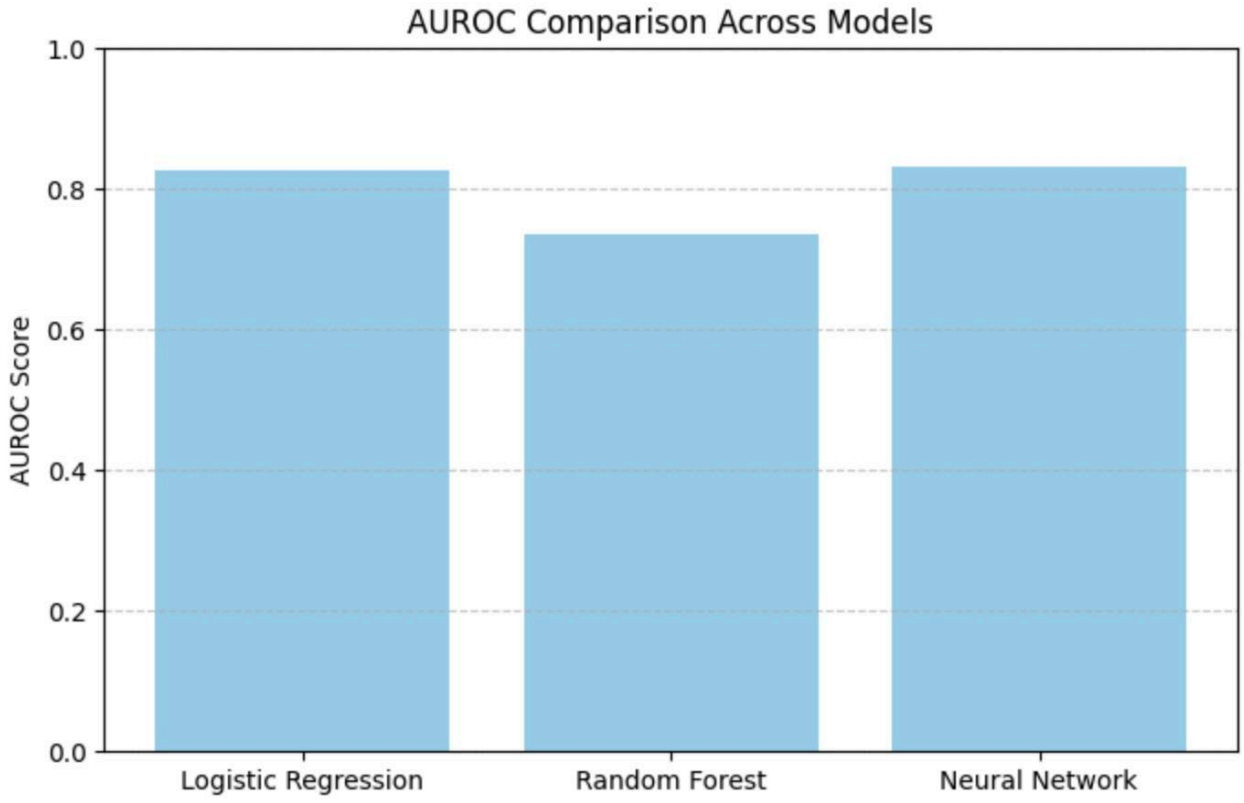

Figure 1: **AUROC Score comparison across all three models.** This bar graph shows the AUROC scores for each of the three models. The table above shows the actual score, rounded to the fourth digit.

Out of the three models we used, we found that Neural Network was the best performing model based on the AUROC score. In **Figure 1**, we can see that Logistic Regression and Neural Network performed similarly, while Random Forest performed comparatively worse. The graph shows that the Neural Network had a AUROC score of approximately 0.833, the Logistic Regression model had a slightly lower AUROC score of 0.826, and the Random Forest model had a score of 0.736.

Model Evaluation Summary:
Logistic Regression - AUPRC: 0.4069
Random Forest - AUPRC: 0.2908
Neural Network - AUPRC: 0.4316

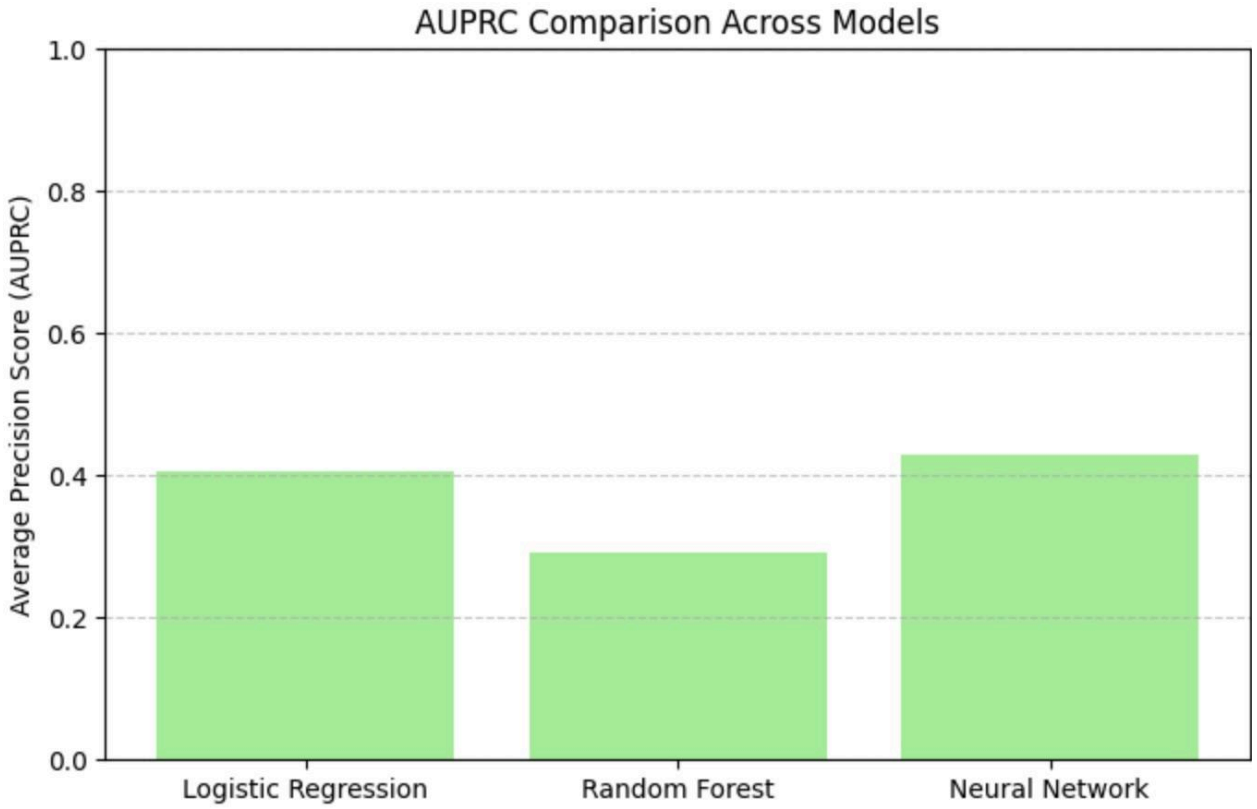

Figure 2: AUPRC Score comparison across all three models. This bar graph shows the AUPRC scores of each of the three models. The table above shows the actual score, rounded to the fourth digit.

To further evaluate the performance of the models, we plotted the AUPRC scores, as shown in **Figure 2**. We can see in **Figure 2** that random forest has a score below both logistic regression and neural network, making it the least accurate model. Additionally, neural networks once again had a higher score than logistic regression. The graph shows that the Neural Network had an AUPRC score of approximately 0.432, the Logistic Regression model had an AUPRC score

of 0.407, and the Random Forest model had a score of 0.291. Although all three scores are relatively low, suggesting that the models are not very strong in correctly identifying positive diabetes cases, these scores still show the difference between the three models. It can clearly be seen that even with a low score, the neural network still performs the most consistently. Although the AUPRC scores are notably low for all three models, the result is expected. The low AUPRC scores come from the dataset imbalance. In the dataset, the number of positive diabetes cases is smaller than the number of those without diabetes. Despite the relatively lower AUPRC score, the neural network still achieved the highest AUPRC score, proving it to be a stronger model at identifying positive diabetes cases than the other two models.

Model Evaluation Summary:
Logistic Regression - Accuracy: 0.8659
Random Forest - Accuracy: 0.8621
Neural Network - Accuracy: 0.8597

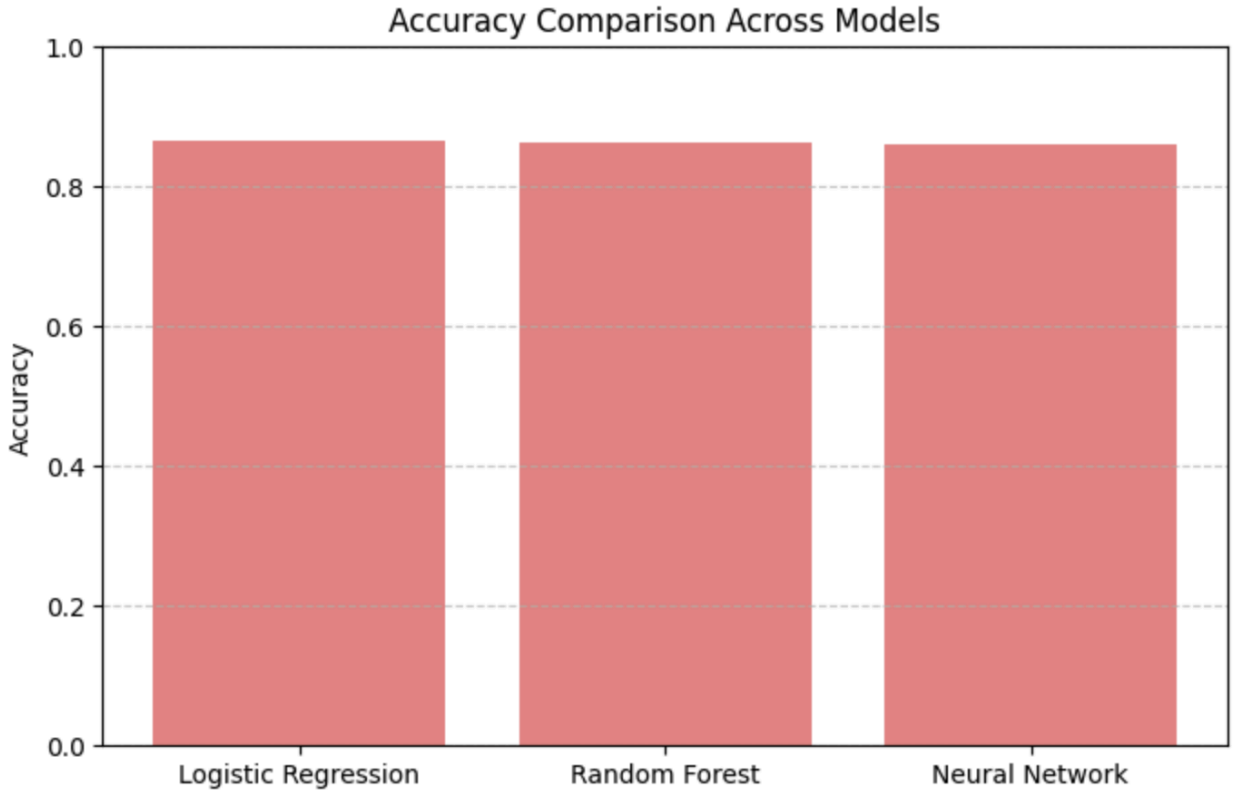

Figure 3: Accuracy Score comparison across all three models. This bar graph shows the Accuracy scores for each model. The table above shows the actual score, rounded to the fourth digit.

Finally, we used Accuracy scores as another method of comparison for the models. As seen in **Figure 3**, the scores among the models are all pretty high, with all being above 0.8, and having very similar scores. According to the table and graph, Logistic Regression had the highest score

of about 0.866; then Random Forest, with a slightly lower score at around 0.862; and finally, the Neural Network had a score of close to 0.86. The differences between each score are notably small, indicating that the models performed similarly in terms of accuracy. As mentioned earlier, due to dataset imbalance, the scores can be misleading. That is why logistic regression has a high accuracy score, but a lower AUPRC and AUROC score compared to the Neural Network. The Neural Network most likely achieved the lowest accuracy score because it is more sensitive in finding positive diabetes cases, which causes a higher recall, but also more false positives. This, as a result, reduces the overall accuracy score. This behavior of the Neural Network is what leads to its higher AUROC and AUPRC scores since the model is stronger at identifying positive cases at different thresholds. Thus, a slightly lower score from Neural Network does not necessarily mean a weaker performance.

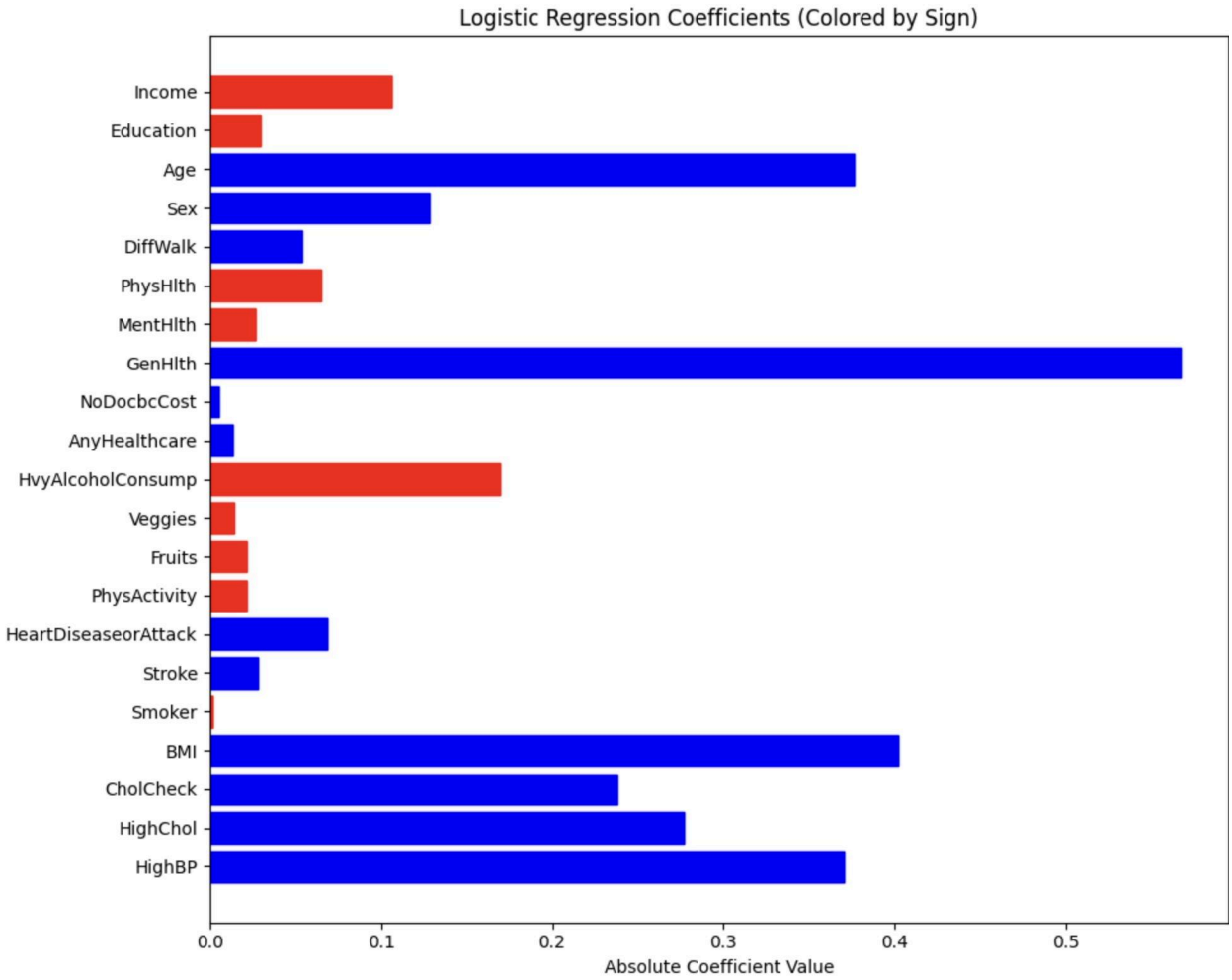

Figure 4: Logistic Regression Coefficients for each feature. This graph shows the absolute value of the coefficients from the logistic regression model. Red means the feature originally had a negative value, and blue means the feature has a positive value.

To better understand the features influencing each model's performance, we applied a range of interpretation techniques. For the logistic regression model, we examine the feature coefficients. For the random forest model, we use feature importance scores. Lastly, for the neural network, we employ Shapley scores.

In **Figure 4,** we can see the results of the logistic regression model. The graph shows all the features and their significance. The negative coefficients (red) show all the features in the dataset that do not contribute to developing diabetes. For example, as seen in **Figure 4** fruits and vegetables had a negative coefficient. This makes sense since eating more fruits and vegetables and having a healthy diet will not lead to developing diabetes. On the other hand, from this model, we can see that factors such as general health and BMI significantly contribute to having a higher chance of developing diabetes, as shown by their large positive coefficient values.

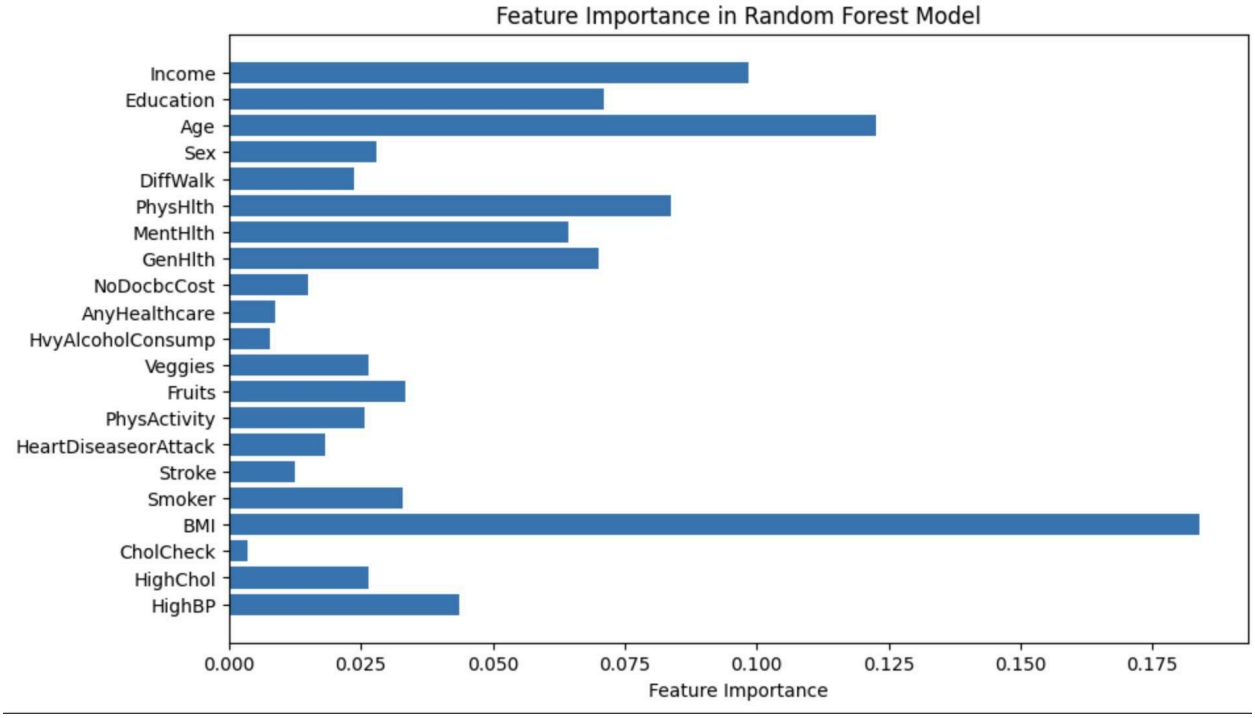

Figure 5: Feature importance bar graph from the Random Forest Model. This graph shows the feature importance values from the random forest model.

Figure 5 shows the feature importance from the random forest model. From the bar graph, we can see that in the random forest model, BMI was the most important feature in terms of predicting whether one will develop diabetes. Not only does the model predict that BMI is the most important feature, but it also shows that BMI is the most significant feature as compared to other features with a feature value of over 0.175. The second highest feature, age, only has a

feature importance value of around 0.125. Income also has a high feature value from the random forest model.

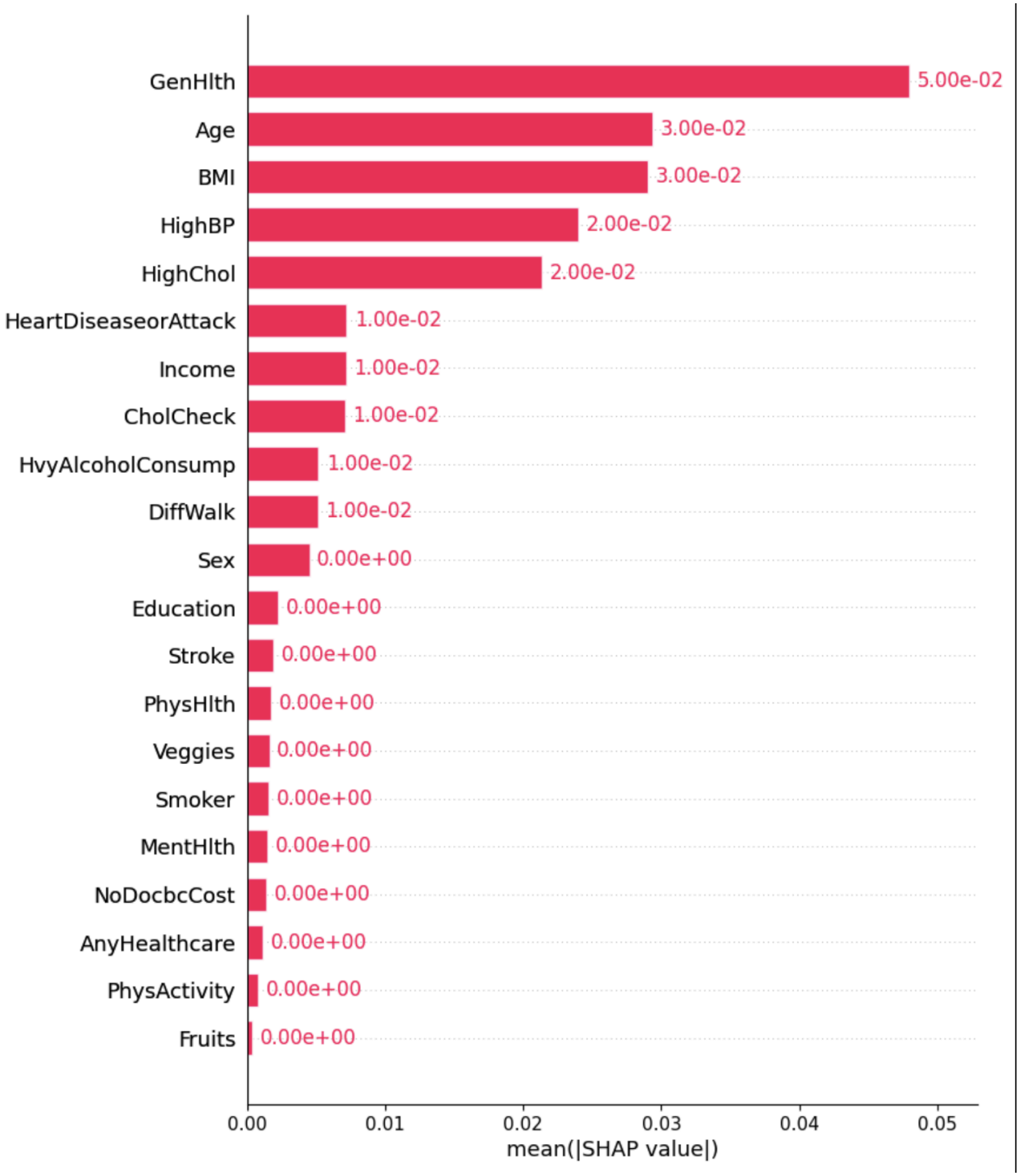

Figure 5: Shapley graph summary plot from the neural network model. This is the Shapley graph, showcasing the results of Neural networks. The features at the top have the highest SHAP value.

From **Figure 5**, we can see the bar graph summary plot from the neural network model. The graph shows the SHAP value in order from the highest value, being at the top of the graph, to the lowest value, at the bottom. The SHAP value is similar to the feature importance values in the previous models. The higher the SHAP value, the more likely that feature is correlated to developing diabetes. In **Figure 5**, we can see that general health has the highest SHAP value of 5.00e^-2, followed by BMI and age. The features with the lowest SHAP value, or impact on developing diabetes, from the neural network model are things such as education, vegetables, mental health, etc. These features, although shown as 0.00e^00 in **Figure 5**, have a low feature importance value.

4. CONCLUSION:

The outcomes of this study show that it is possible to use machine learning models to predict whether a patient is likely to develop diabetes based on different features. Among the three models we used: Logistic Regression (the baseline), Random Forest, and Neural Networks, neural networks proved to have the highest performance, outperforming the other two models with the AUROC score of 0.8283 and an AUPRC score of 0.4276. For accuracy, all three models had similar and high scores above 0.8. Although the neural network performed the best, overall, the other two models proved to be effective models as well, based on their AUROC and AUPRC scores. Even after performing k-fold cross-validation with a k value of 5, all three models yielded similar AUROC and AUPRC scores, which shows that even with unseen data, these models can still be reliable.

The neural network model predicted that general health was the feature with the most impact on being diagnosed with diabetes. Following general health was BMI. The three models had similar results, especially logistic regression and the neural network, but the random forest had slightly different results. Features such as fruits, vegetables, education, mental health, etc., were considered insignificant across all models, but unlike neural networks and logistic regression, random forests still considered those features to be predictive of diabetes. Both logistic regression and neural networks, which had similar and high AUROC and AUPRC scores, had an output showing that the three main factors we can use to determine whether a patient has diabetes or not are general health, BMI, and age. Additionally, logistic regression and neural networks had similar accuracy scores and produced highly consistent feature outputs, proving the similarity of the models. Thus, if one reports poor general health, they are the most likely to have diabetes.

These models (especially logistic regression and neural network) have shown promise for use in healthcare settings. These models can determine significant predictors that lead to diabetes, and these models can be trained and used to determine the likelihood of a patient being diagnosed with a disease or illness.

Although the overall performance of the models was reassuring, our study still has limitations. The models we used were only trained on one dataset. Furthermore, the dataset contained bias, or imbalances that potentially affected the models' accuracy. Because of dataset imbalance, looking at accuracy alone can be misleading, as it may not fully reflect performance on positive diabetes cases. In the future, this research could use more diverse datasets to further improve the model's accuracy and better understand which factors determine whether someone has diabetes.

To conclude, this research showcases the importance and feasibility of using machine learning models to detect and potentially prevent diseases such as diabetes. Ultimately, as diabetes continues to be one of the leading causes of death around the world, the usage of machine learning models must be used to provide earlier treatment to patients and prevent them from developing the disease.

References:

1. Bloomgarden, Zachary T. "What Will We See in Diabetes in the next 10 Years?" *Journal of Diabetes*, vol. 16, no. 6, June 2024, p. e13594. *DOI.org (Crossref)*, https://doi.org/10.1111/1753-0407.13594.

2. Qin, Yifan, et al. "Machine Learning Models for Data-Driven Prediction of Diabetes by Lifestyle Type." *International Journal of Environmental Research and Public Health*, vol. 19, no. 22, Nov. 2022, p. 15027. *Crossref*, https://doi.org/10.3390/ijerph192215027.

3. Habehh, Hafsa, and Suril Gohel. "Machine Learning in Healthcare." *Current Genomics*, vol. 22, no. 4, Dec. 2021, pp. 291–300. *Crossref*, https://doi.org/10.2174/1389202922666210705124359.

4. Khare, Akshay Dattatray . "Diabetes Dataset." *Www.kaggle.com*, 2022, www.kaggle.com/datasets/akshaydattatraykhare/diabetes-dataset.

5. Serrano, Luis. *Grokking Machine Learning*. Manning Publications Co. LLC, 2021.

6. Schonlau, Matthias, and Rosie Yuyan Zou. "The Random Forest Algorithm for Statistical Learning." *The Stata Journal: Promoting Communications on Statistics and Stata*, vol. 20, no. 1, Mar. 2020, pp. 3–29. *DOI.org (Crossref)*, https://doi.org/10.1177/1536867X20909688.

7. Hancock, John T., et al. “Evaluating Classifier Performance with Highly Imbalanced Big Data.” *Journal of Big Data*, vol. 10, no. 1, Apr. 2023, p. 42. *DOI.org (Crossref)*, https://doi.org/10.1186/s40537-023-00724-5.