Predicting Diabetes in Patients

Dallin Moore - DATA 5610 - Project #1

Overview

Ultimately, using machine learning to predict diabetes in patients is promising but neural networks may not be the best approach with the current data available. The RandomForest Classification model slightly outperformed the Feedforward Neural Network, with accuracy scores of 0.970 and 0.964, respectively.

The Problem

We are concerned with using Data Science to provide a suite of tools to support clinicians in the diagnosis and analysis of patients in their care. Specifically, we are concerned with predicting diabetes in patients before a doctor even reviews the patients charts. This is one of many ways Data Science can be used to increase the efficiency of doctors.

The Project

We compare a simple RandomForest Classification model with a Feedforward Neural Network (2 hidden layers) to assess the best method for classifying patients positively or negatively for diabetes. Operating under compute and time/cost restraints

The Dataset

The dataset used for the models was acquired through Kaggle here. The only preprocessing steps for both models were dummy coding the categorical variables. The data was split into a training set, a test set, and an evaluation set. The test set was only used for accuracy scores after the model had been trained.

The Models

RandomForest Classification

A RandomForest Classification was applied to the data as a baseline to test against the neural network. The following metrics were observed, RMSE: 0.1715, R^2 Score: 0.6176, and Accuracy (test set): 0.9696.

Feedforward Neural Network

The Feedforward Neural Network classification was originally trained with 2 hidden layers containing 32 nodes in the first layer and 16 the second, a batch size of 16, 20 epochs, adam as

the optimizer, binary cross entropy as the loss function, and accuracy as the metric. After the original model performed well, the model was tuned with the following values being tested: number of nodes in the first and second hidden layer (16, 32, 64) and number of epochs (5,10,20). Ultimately, the best of the 27 variations tried contained 32 nodes in the first hidden layer, 64 in the second, and 20 epochs. This variation was able to achieve an accuracy of 0.966. Overall, the variations did not make a huge difference with the worst accuracy reaching a low of 0.951.

Conclusion

This study demonstrates that predicting diabetes in patients using machine learning is highly feasible, with both the RandomForest Classifier and Feedforward Neural Network achieving high accuracy scores (0.970 and 0.964, respectively). The slight edge of the RandomForest model suggests that for structured, tabular data like this, tree-based methods may be more efficient than deep learning approaches due to their ability to handle feature importance and interactions without extensive tuning.

Beyond diabetes prediction, these results suggest that similar approaches could be applied to other binary or multi-class medical classification tasks, such as predicting heart disease, hypertension, or even identifying patients at risk for complications based on medical history. Given the structured nature of the dataset, diagnoses with clear numerical indicators (e.g., blood pressure for hypertension, cholesterol levels for cardiovascular disease) are particularly well-suited for this type of modeling.

The best model identified in this study was the RandomForest Classifier, which achieved the highest accuracy (0.970). While the Feedforward Neural Network performed well, it required hyperparameter tuning and computational resources without significantly outperforming the simpler model. The definition of "best" in this case was based on accuracy, but in a clinical setting, other metrics such as precision, recall, and interpretability might be more relevant.

Ultimately, this project supports the idea that machine learning can enhance clinical decision-making by preemptively flagging patients at risk for diabetes, thereby improving efficiency and patient outcomes. Future work could explore additional data sources, feature engineering techniques, and ensemble methods to further refine predictive accuracy.