A Comparative Analysis of Machine Learning Models for Predicting Diabetes Risk

Allen Chang
Dornsife College of Letters Arts and Sciences
University of Southern California
Los Angeles, California
achang30@usc.edu

Abstract

In this study, I identify the most effective machine learning approach for diabetes prediction. Here, I compare three models: a logistic regression, a simple fully connected neural network, and a decision tree classifier. Features for diabetes prediction are age, hypertension, heart disease, body mass index (BMI), HbA1c levels, and blood glucose levels. My findings reveal that the fully connected neural network outperforms other models, and that HbA1c levels have the most significant impact on diabetes prediction.

Introduction

Diabetes is a disease characterized by a constant state of hyperglycemia resulting in issues related to both insulin secretion and action [1]. In fact, 0.5% of US adults are diagnosed with type 1 diabetes while 8.5% of US adults are diagnosed with type 2 diabetes [2]. Globally, about 45.8% of diabetes cases are estimated to be undiagnosed [3]. Thus, there arises a need for an accurate method to diagnose or predict an individual's risk for diabetes. Here, I compare the performances of three distinct models including a logistic regression, a neural network, and a decision tree classifier. Using weights of the logistic regression model, I also present insights on the significance of each diabetes predictor.

	T-statistic	**P-value**
Age	119.6	0.0
BMI	60.3	0.0
Blood Glucose Level	94.8	0.0
HbA1c Level	127.0	0.0

Table 1 *Differences in the features of people with and without diabetes are shown.*

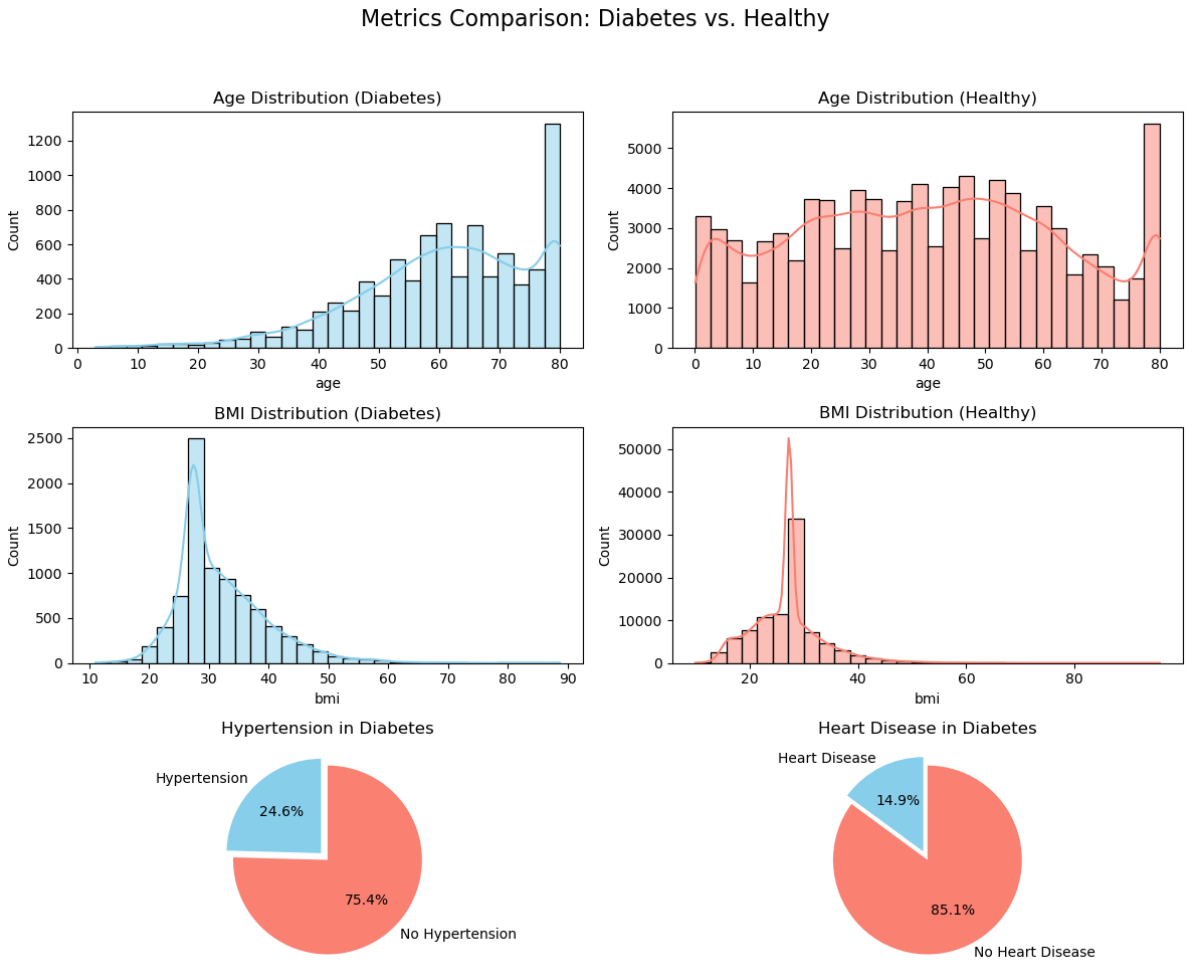

Figure 1 *Feature information and distributions for both diabetic and healthy patients are shown.*

Methods

The cohort includes 8,500 diabetic patients and 91,500 healthy patients. Further information about the dataset can be found here. To make the models gender agnostic, gender was not used as a training feature. Smoking history was also excluded due to an excess amount of missing information.

The dataset was divided using an 80:20 split for training and test sets for the logistic regression and decision tree classifier. The training set was further divided into an 80:20 split for the training and validation sets for the neural network.

For the logistic regression model, Lasso (L1) regularization was used with a regularization parameter of 10.0. The feature weights were retrieved for further analysis. For the neural network model, four dense layers were used in total. The first three dense layers had 256, 128, and 64 units, respectively and employed ReLU activation. The last dense layer had 1 unit and employed a sigmoid activation function. The Adam optimizer with a learning rate of 0.001 minimizes the binary cross-entropy loss function. The model was run for 100 epochs using a batch size of 150, and the epoch with the highest accuracy was saved. Information about model

training is shown in **Figure 2**. The decision tree classifier was trained using default parameters. To assess model performance, AUROC (Area under Receiver Operator Characteristics) score, f1-score, precision, recall, and accuracy were calculated.

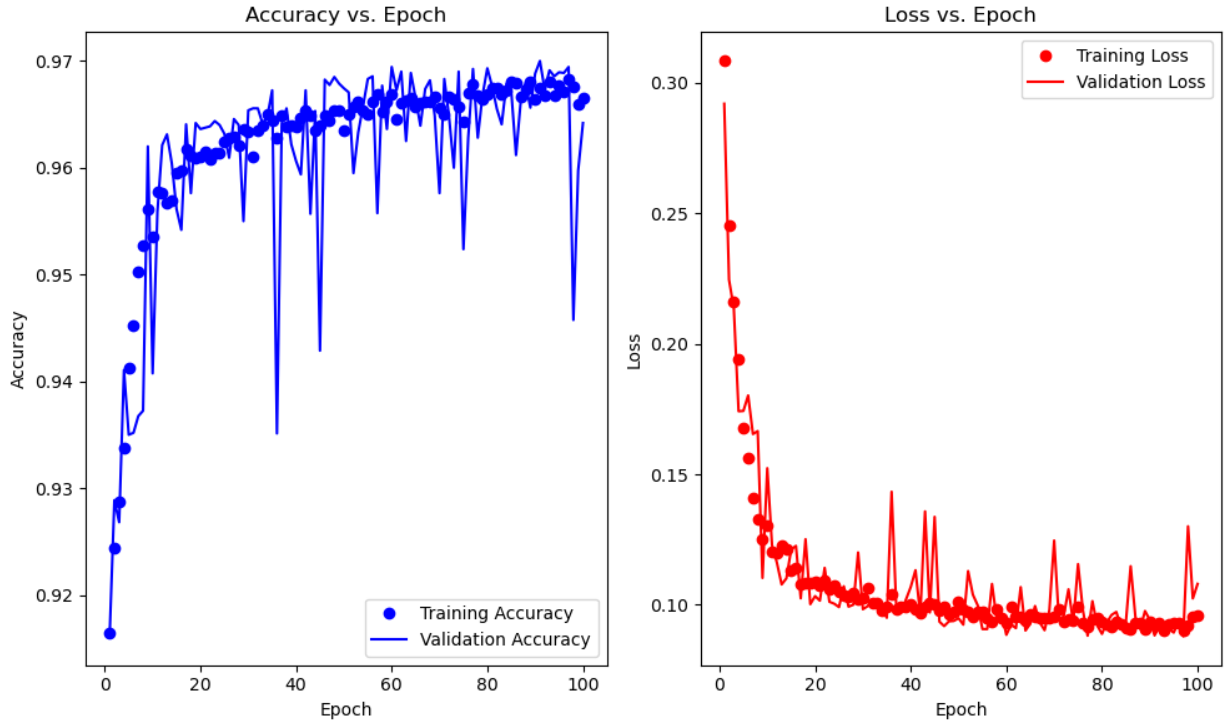

Figure 2 *Training and validation accuracy and loss for the neural network are shown.*

Results and Discussion

As seen in **Table 2** and **Figure 2**, the neural network has the highest AUROC score, f1-score, accuracy, and recall. Nevertheless, its performance for precision was worse than that of the decision tree but also better than the logistic regression model. The logistic regression had the second highest AUROC score, recall, and accuracy. Nevertheless, its f1-score was lower than that of the decision tree classifier. The decision tree classifier had the worst AUROC score, second best f1-score, worst recall, and worst accuracy despite have a precision much higher than the other two models. However, it is essential to recognize that the dataset's imbalance and inherent classification nature of decision trees can lead to suboptimal insights from the AUROC. In fact, the imbalance can clearly be visualized in **Figure 4**.

	AUROC	F1-score	Precision	Recall	Accuracy
Logistic Regression	0.96	0.72	0.61	0.87	0.96
Neural Network	0.97	0.80	0.68	0.97	0.97
Decision Tree Classifier	0.86	0.75	0.75	0.74	0.96

Table 2 *Model performance is shown.*

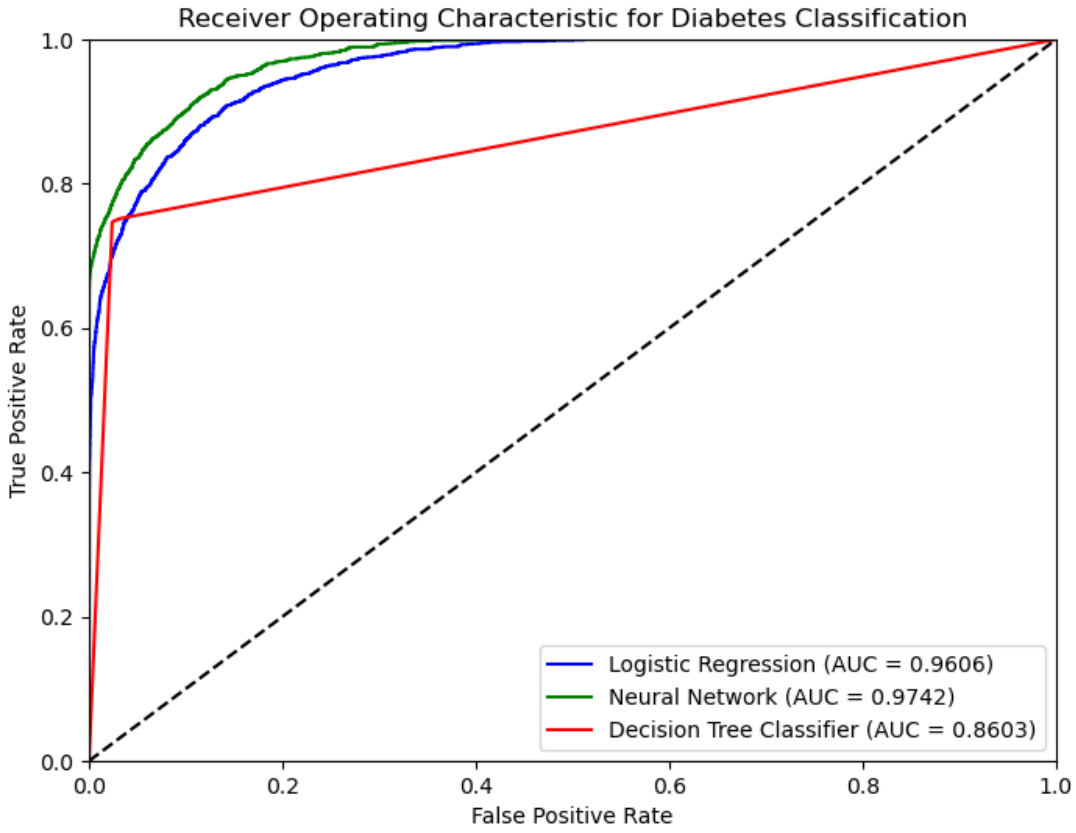

Figure 3 *Area under the ROC curve is shown.*

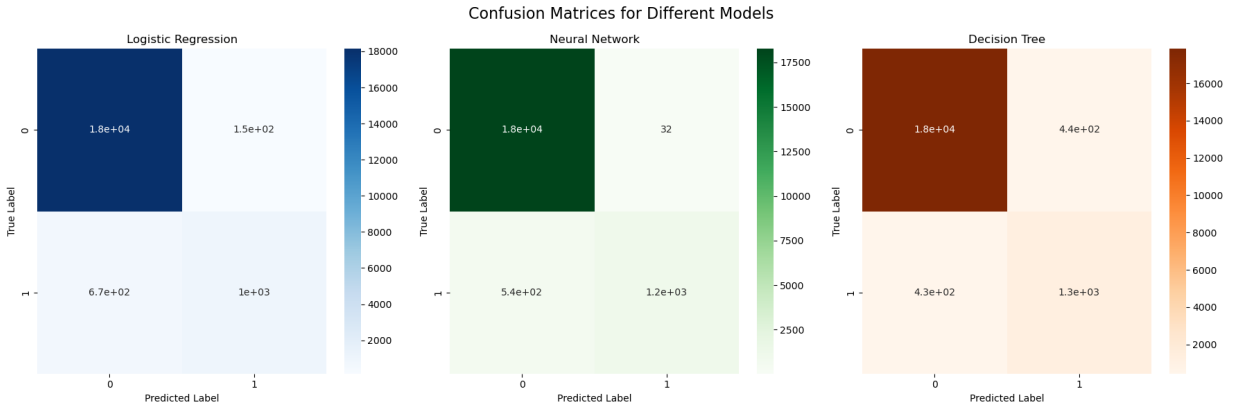

Figure 4 *The confusion matrix for the three models is shown.*

Examining the weights from logistic regression reveals features importance for diabetes prediction. As seen in **Figure 5**, HbA1c levels are by far the most important feature for a positive diabetes prediction. A similar finding is seen in **Table 1** where there is a significant difference between the HbA1c levels of healthy and diabetic patients. HbA1c is glucose-bound hemoglobin, and it is directly proportional to blood glucose levels [4]. In fact, HbA1c is considered the test of choice for monitoring and managing chronic diabetes [4]. Hypertension and heart disease are also important predictors for diabetes. This validates previous findings that patients with diabetes are 2 to 8 times more likely to have heart issues than non-diabetic individuals [5].

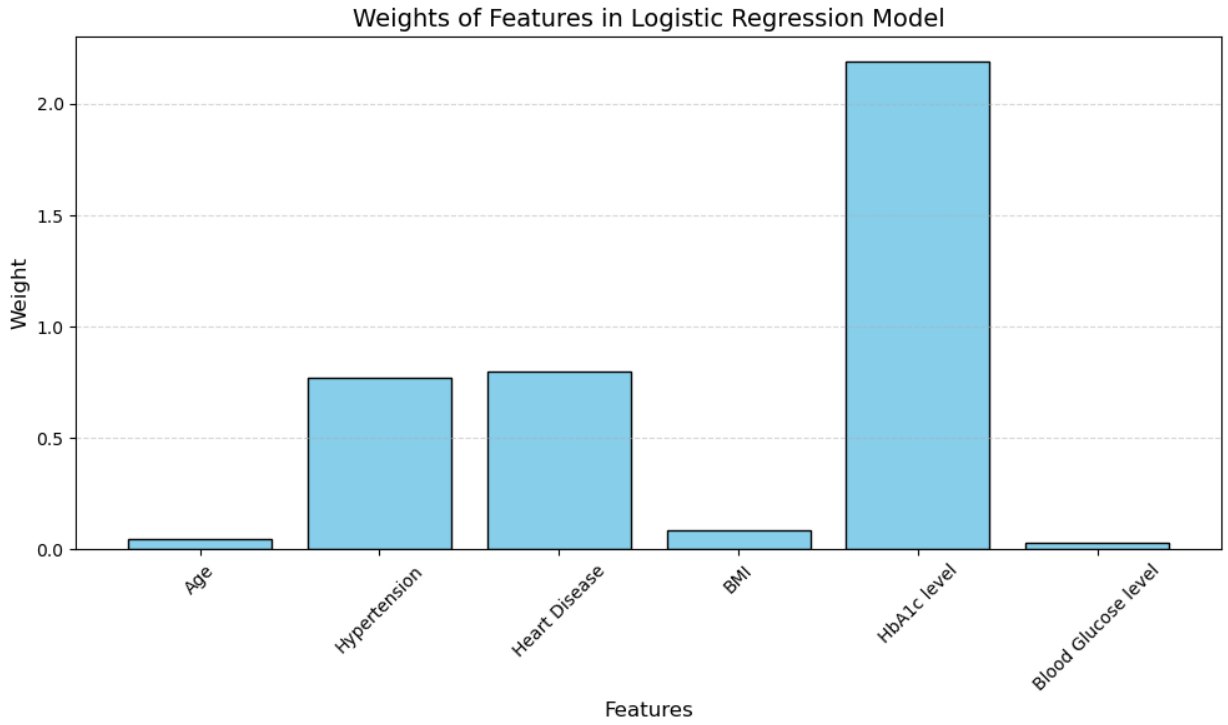

Figure 5 *Feature weights for the logistic regression model are shown.*

Conclusion

Neural networks outperform the logistic regression and decision tree classifier models in all tested metrics besides precision. Thus, it is most robust for most real-world applications. Furthermore, examining weights retrieved from the logistic regression shows that HbA1c level and cardiovascular health are among the most important predictors of diabetes.

More research can be done on a wider variety of models and hyperparameters. Furthermore, more potential features can also be investigated. The dataset used may not be representative of the real-world population.

References

[1] A. T. Kharroubi, "Diabetes mellitus: The epidemic of the century," *World J. Diabetes*, vol. 6, no. 6, p. 850, 2015, doi: 10.4239/wjd.v6.i6.850.

[2] G. Xu *et al.*, "Prevalence of diagnosed type 1 and type 2 diabetes among US adults in 2016 and 2017: population based study," *BMJ*, p. k1497, Sep. 2018, doi: 10.1136/bmj.k1497.

[3] J. Beagley, L. Guariguata, C. Weil, and A. A. Motala, "Global estimates of undiagnosed diabetes in adults," *Diabetes Res. Clin. Pract.*, vol. 103, no. 2, pp. 150–160, Feb. 2014, doi: 10.1016/j.diabres.2013.11.001.

[4] S. I. Sherwani, H. A. Khan, A. Ekhzaimy, A. Masood, and M. K. Sakharkar, "Significance of HbA1c Test in Diagnosis and Prognosis of Diabetic Patients," *Biomark. Insights*, vol. 11, p. BMI.S38440, Jan. 2016, doi: 10.4137/BMI.S38440.

[5] M. Moslemirad4*, "DIABETES AND THE RISK OF SUFFERING CARDIOVASCULAR DISEASES: A TWO-YEAR RETROSPECTIVE STUDY," *Int. J. Ecosyst. Ecol. Sci.*, vol. 8, no. 3, pp. 649–656, Jun. 2018, doi: 10.31407/ijees8328.

Dataset: https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset/data