Machine Learning to Predict Diabetes

Presented by

Puja Sonawane

TABLE OF CONTENTS

1. Abstract

In USA, one person in seven suffers from Diabetes and it will keep increasing according to the Centre of Disease Control and Prevention. To address the issue of increasing rate of patient suffering from Diabetes, Machine Learning was used in this project to predict the Diabetic Condition of a person. This study implements Deep Learning - the advance approach of machine learning to overcome the drawbacks of the traditional machine learning classification models.

In later part of the project, traditional classification models and deep learning classification model was evaluated using different evaluation matrices. It was found that deep learning model gave better performance with respect to other classification models.

2. Introduction

The rate of diabetic patients has been increasing in USA. In the next coming years, it is predicted that proportion of people suffering from diabetes can go up to one person in three by 2050 from one person in seven.

In order to address this issue before it gets more sever, machine learning was used in this project to predict the diabetes. The goal of this project is to build significantly efficient models to predict diabetes of an individual so that depending on the prediction individual one can take care of the lifestyle and diet in a healthy way which would not cause diabetes. In this study, it was also premeditated which factor contributes most towards making decision. This will help any individual to avoid the factors which are more likely to cause diabetes.

For the scope of this project, different traditional classification models are explored, later deep learning was used to build the better classification model. All technical details and methodology used are described further in the report.

3. Problem Statement

To build a binary classification model using machine learning to predict the diabetes so that precautions can be taken to control the rising number of patients of diabetes in the United States.

4. Overview of Proposed Approach

In this study, Deep learning was used to get better model for classification for predicting diabetes. As deep learning uses neural network which has unique ability to dynamically create complex prediction function and to emulate human thinking for better predictions in real world problems. Sequential neural network works better with tabular data and the dataset used in this project was a tabular dataset. Therefore, sequential neural network is used to build classification model.

To evaluate the model performance, several performance measures were used like accuracy and loss. All technical details are mentioned in next section of the report.

5. Technical Details

This section covers technical details of all models built during this study, their evaluation measures and the best selected model with accuracies achieved.

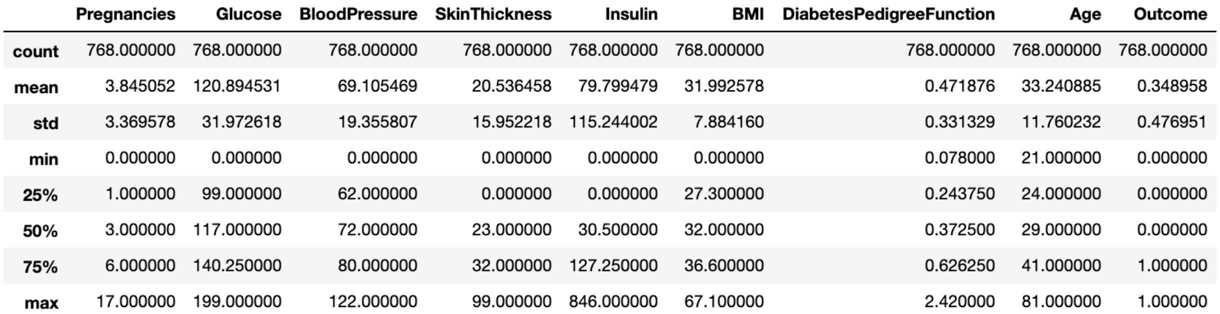
a. Data Description

The diabetes dataset is taken from UCI Machine Learning Repository. It consists of 768 datapoints and 9 features. Outcome was the dependent variable. Zero indicates an outcome means no diabetes whereas one value of diabetes means diabetes.

The nine features in dataset are as represented below:

Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, Age, Outcome.

Statistical description of the variables was the following:



	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
count	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000
mean	3.845052	120.894531	69.105469	20.536458	79.799479	31.992578	0.471876	33.240885	0.348958
std	3.369578	31.972618	19.355807	15.952218	115.244002	7.884160	0.331329	11.760232	0.476951
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.078000	21.000000	0.000000
25%	1.000000	99.000000	62.000000	0.000000	0.000000	27.300000	0.243750	24.000000	0.000000
50%	3.000000	117.000000	72.000000	23.000000	30.500000	32.000000	0.372500	29.000000	0.000000
75%	6.000000	140.250000	80.000000	32.000000	127.250000	36.600000	0.626250	41.000000	1.000000
max	17.000000	199.000000	122.000000	99.000000	846.000000	67.100000	2.420000	81.000000	1.000000

Data preprocessing

Scaling of data – Scaling of data used was done by using normalization means and obtained the means close to zero. Normalization of data speeds up the training of neural network and gives better model. To normalize the data, StandardScalar function was used from sklearn preprocessing library.

Following data was divided into training and testing dataset with 70:30 proportion respectively.

b. Baseline Methods

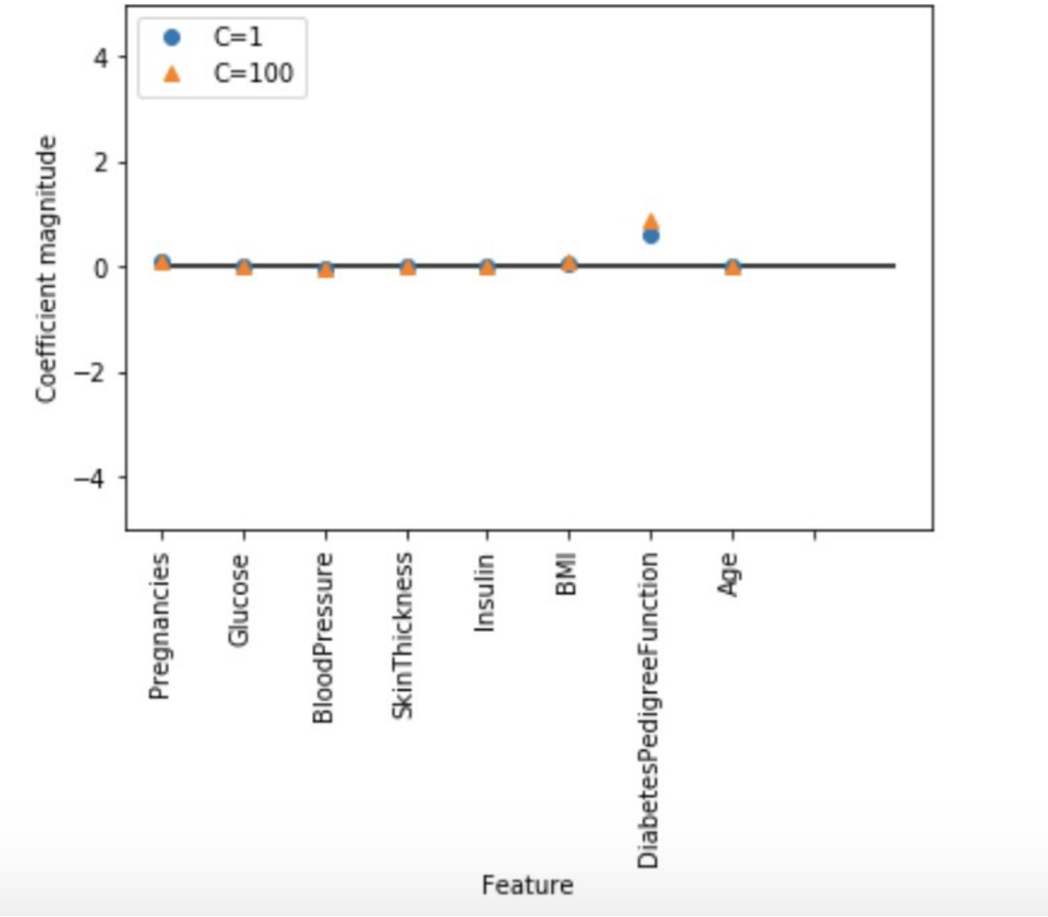
To obtain best performing model, several classification models like Logistic Regression, Decision Tree classifier, k-Nearest Neighbors, Random Forest Classifier and Neural Network were implemented.

Logistic Regression- It is a classic and common classification algorithm.

The trade-off parameter c in logistic regression is the inverse regularization parameter which is 1/λ and also determines the regularization strength. As shown in the picture below, the higher the value of c (c=100) results in less regularization and the default value of c (c=1) was comparatively better and selected as increasing the model complexity had not necessarily generalized it better. Henceforth, the feature DisbetesPedigreeFunction which was observed to be high was related and contributed significantly to the sample of the dataset irrespective of which model had to be chosen.

Training accuracy is 0.7812

Testing accuracy is 0.7708



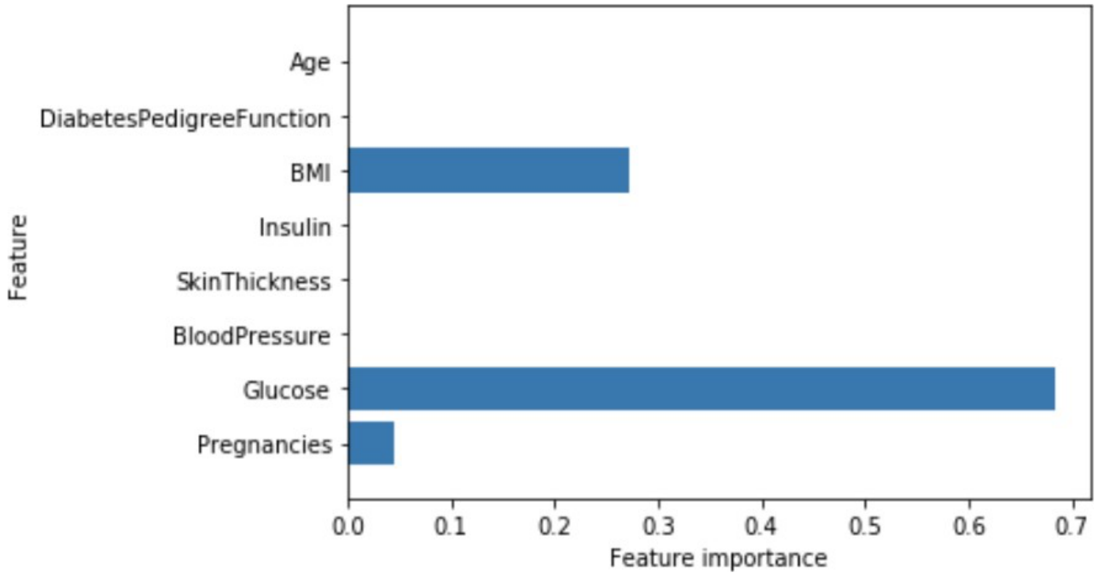
Decision Tree Classifier-

Decision trees are very easy to understand and resistant to noisy data. It's a recommended option to use decision tree for classification. As decision trees are simple, feature importance towards making decision can be obtained.

Training accuracy is 0.77.

Testing accuracy is 0.73.

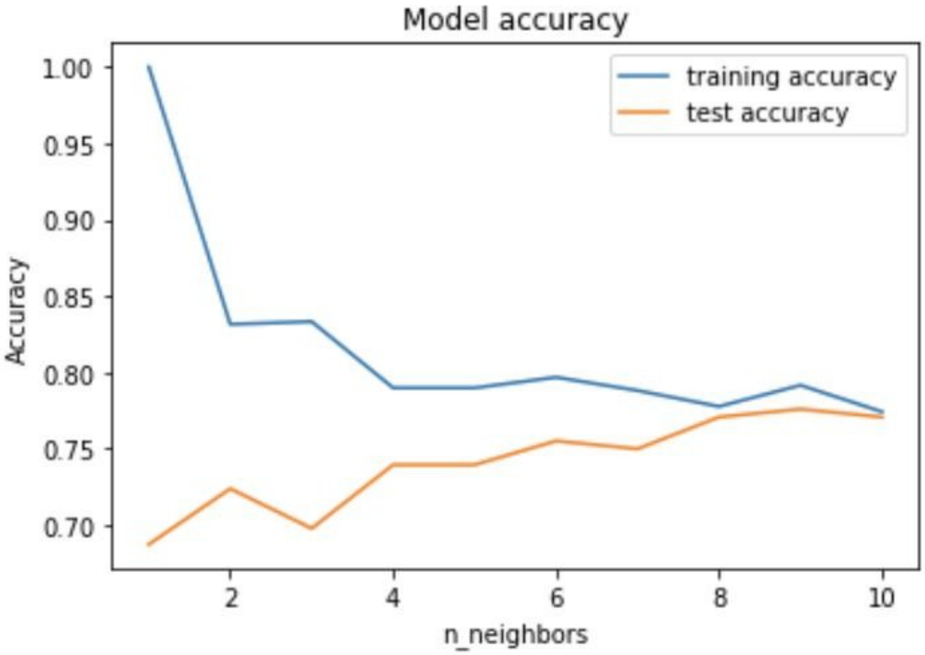
Feature importance by decision tree model:



Glucose and BMI were crucial factors contributing to cause diabetes. Therefore, individuals facing potential risk of suffering from Diabetes should pay attention to these two factors.

K - Nearest Neighbor-

The K-NN algorithm is the simplest machine learning algorithm. It predicts considering neighbors characteristics.



Training accuracy is 0.7743

Testing accuracy is 0.7708

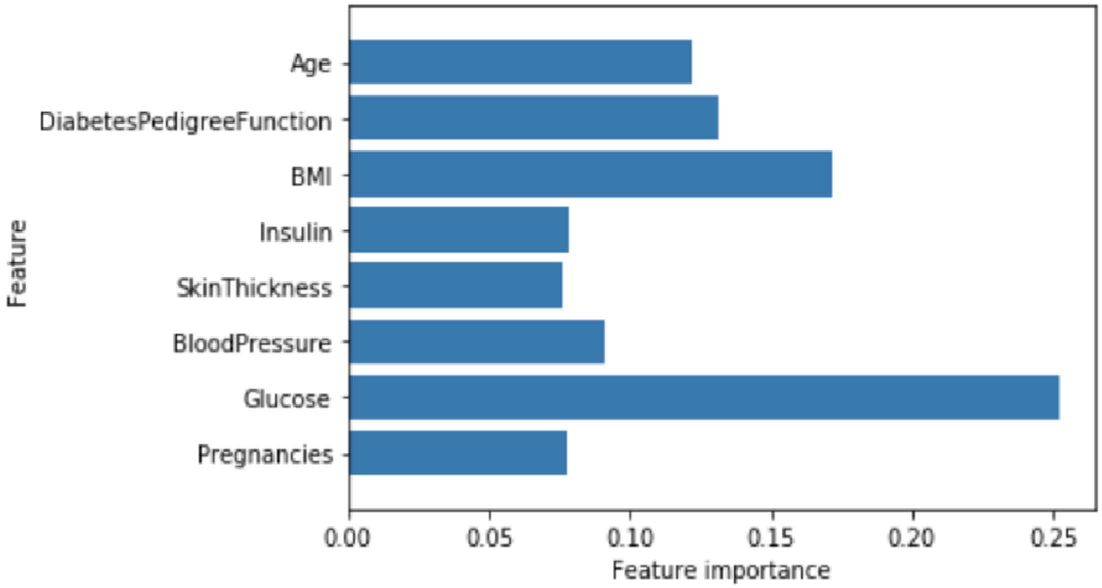
Random Forest-

Random forest is similar to decision tree but instead of making decision from single decision tree it makes decision by considering the decisions of hundreds of trees called as Random Forest.

Training accuracy is 0.80

Testing accuracy is 0.755

Feature importance:



According to random forest the features Glucose as well as BMI influences the decision most. The extent of effect of other features in decision making can be seen as demonstrated above.

Neural Network Model-

Neural network is a very complex but very efficient in building predictive function to accurately perform classification task. When it comes to business platform, it is very stable model unlike decision trees.

Initial neural network model gave the accuracy of 0.72 which was comparatively less than other baseline models.

After application of parameter tuning approach, optimal hyper parameter's values obtained are the following,

Activation function- softmax

Optimizer- adam

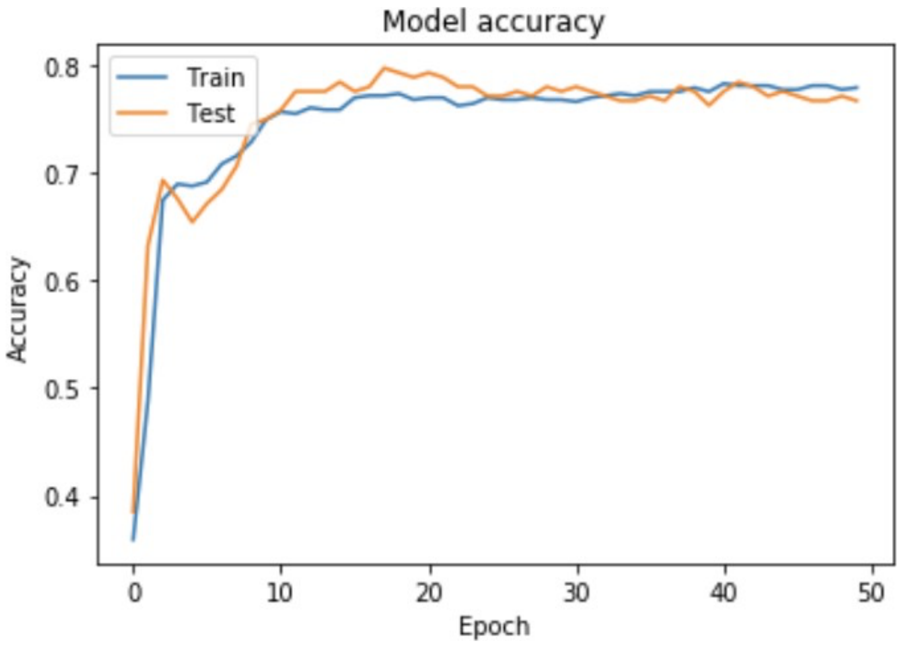
Epoch -50

Batch size- 32

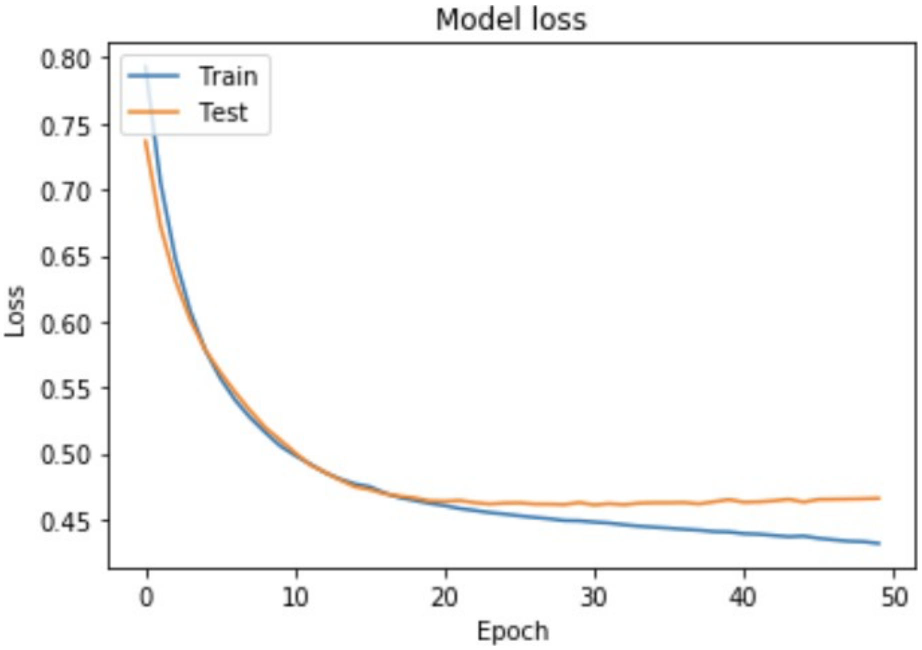
Also using the optimal values of hyper parameters, neural network model's accuracy 0.7520 was obtained which gave out a substantial increase.

Performance measuring graphs:

Model Accuracy:



Model loss:



c. **Best Model and Performance**

Neural network was selected as superior model because of its stability and chance of getting a better model after applying some techniques like parameter tuning. Neural network model was observed to be 0.7520 with regards to testing accuracy.

6. **Conclusion**

- Various classical classification algorithms were used to address the prediction of diabetes such as Logistic Regression, Decision Trees, K-Nearest Neighbors, Random Forests and Neural Networks.

- The feature Glucose was observed to be one of the most influential features towards diabetes as it had a high importance in many of the classification algorithms used in this project.
- It was found that the most stable model considering business platform with better accuracy was Neural Network.
- Considering the size of the dataset one can assume Random Forests to be an appropriate classification method since it′s simpler and effective. However, with appropriate tuning and training the Neural Network; Deep Neural Networks haa a significant impact on the property of statistical strength which Random Forest lacks.

7. Preferences

- charleshsliao, V., 2020. *Logistic Regression In Python To Tune Parameter C*. [online] Charles' Hodgepodge. Available at: <https://charleshsliao.wordpress.com/2017/05/20/logistic-regression-in-python-to-tune-parameter-c/> [Accessed 24 April 2020].
- Colab.research.google.com. 2020. Google Colaboratory. [online] Available at: <https://colab.research.google.com/github/GoogleCloudPlatform/tensorflow-without-a-phd/blob/master/tensorflow-mnist-tutorial/keras_04_mnist_convolutional.ipynb#scrollTo=TTwH_P-ZJ_xx> [Accessed 24 April 2020].

- Medium. 2020. Machine Learning for Diabetes. [online] Available at: <https://towardsdatascience.com/machine-learning-for-diabetes-562dd7df4d42> [Accessed 24 April 2020].
- Medium. 2020. Building Our First Neural Network In Keras. [online] Available at: <https://towardsdatascience.com/building-our-first-neural-network-in-keras-bdc8abbc17f5> [Accessed 24 April 2020].