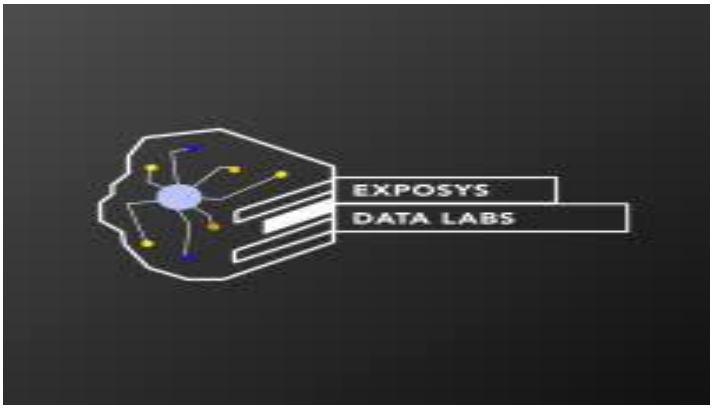


Exposys Data Labs

DIABETES DISEASE PREDICTION REPORT

SUBMITTED BY:

Vaishnavi Pulate

vdpulate@mitaoe.ac.in

ABSTRACT

Diabetes is considered as one of the deadliest and chronic diseases which causes an increase in blood sugar. Many complications occur if diabetes remains untreated and unidentified. India is second most affected country by diabetes in the world. Diabetes is disease in which the glucose level becomes high. The motive of study is to build a suitable Model using Deep Learning in order to predict the disease at early stage. Therefore using Py-Torch framework neural network is build which has different number of neurons helping to predict the disease at early stage. The prediction of disease is done based on different parameters such as Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function**,** Age. The accuracy Given by the model is around 65% to predict the disease.

.

TABLE OF CONTENT

1. INTRODUCTION

Diabetes is an illness which affects the ability of the body in producing the hormone insulin, which in turn makes the metabolism of carbohydrate abnormal and raise the levels of glucose in the blood. In Diabetes a person generally suffers from high blood sugar. Intensify thirst, Intensify hunger and Frequent urination are some of the symptoms caused due to high blood sugar. Many complications occur if diabetes remains untreated. Some of the severe complications include diabetic ketoacidosis and nonketotic hyperosmolar coma. .This could be a clear evidence that, according to WHO, the number of the diabetic patient had been sharply increased from 108 million in 1980 to 422 million in 2014.The World Health Organization predicts that by 2030, diabetes will become the seventh leading cause of death in the world. The global prevalence of diabetes among adults over 18 years of age increased from 4.7% in 1980 to 8.5% in 2014. Early prediction of disease like diabetes can be controlled and save the human life. To accomplish this, this work explores prediction of diabetes by taking various attributes related to diabetes disease. For this purpose we use the Pima Indian Diabetes Dataset, we apply various Deep Learning classification and ensemble Techniques to predict diabetes. Machine Learning Is a method that is used to train computers or machines explicitly. Various Deep Learning Techniques provide efficient result to collect Knowledge by building various classification and ensemble models from collected dataset. Such collected data can be useful to predict diabetes. Various techniques of Machine Learning can capable to do prediction, however its tough to choose best technique. Thus for this purpose we apply popular classification and ensemble methods on dataset for prediction.

1.1 PROBLEM STATEMENT

To prepare dataset and build a diabetes disease prediction model using Data Science.

1.2 OBJECTIVE

1. Prepare dataset using several methods
2. Build a model for disease prediction
3. Training and testing model

2. EXISTING METHOD

There are many method based on Machine Learning as well as some medical methods to predict the diabetes. Few of them are enlisted below:

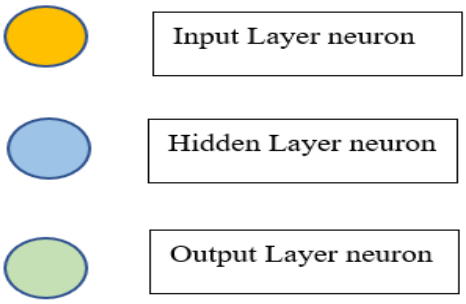
K.Vijiya Kumar proposed random Forest algorithm for the Prediction of diabetes develop a system which can perform early prediction of diabetes for a patient with a higher accuracy by using Random Forest algorithm in ma- chine learning technique. The proposed model gives the best results for diabetic prediction and the result showed that the prediction system is capable of predicting the diabetes disease effectively, efficiently and most importantly, instantly. Nonso Nnamoko presented predicting diabetes onset: an ensemble supervised learning approach they used five widely used classifiers are employed for the ensembles and a meta-classifier is used to aggregate their outputs. The results are presented and compared with similar studies that used the same dataset within the literature. It is shown that by using the proposed method, diabetes onset prediction can be done with higher accuracy. Tejas N. Joshi presented Diabetes Prediction Using Machine Learning Techniques aims to predict diabetes via three different supervised machine learning methods including: SVM, Logistic regression, ANN. This project proposes an effective technique for earlier detection of the diabetes disease.

3. PROPOSED METHOD WITH ARCHITECTURE

In order to solve given problem statement we have used deep learning models to predict the disease at early stage. Dataset has been downloaded from Kaggle and then there is some preprocessing of the dataset is done. As Py-Torch framework is used for building model the pandas dataframe is converted to torch to do further computations. Dataset is being splitted into test-train in-order to check the testing accuracy of the model. The model is build using Py-Torch framework. The model consist of different different number of layer such as Linear layer and sigmoid layer. Linear layer has different number input and output neurons as parameters whereas sigmoid layer is applying activation on the neuron. The number of neurons in the first linear(input layer) layer is decided on various number of inputs and the number of neurons in the last layer(output layer) is one as the problem statement is based on Binary classification. Between the input layer and output layer there are some hidden layers. In this neural network every neuron in the previous layer is connected to every other neuron in the next layer. Then this model is trained and tested and accuracies are predicted.

3.1Model Architectures

Notations:



Model architectures:

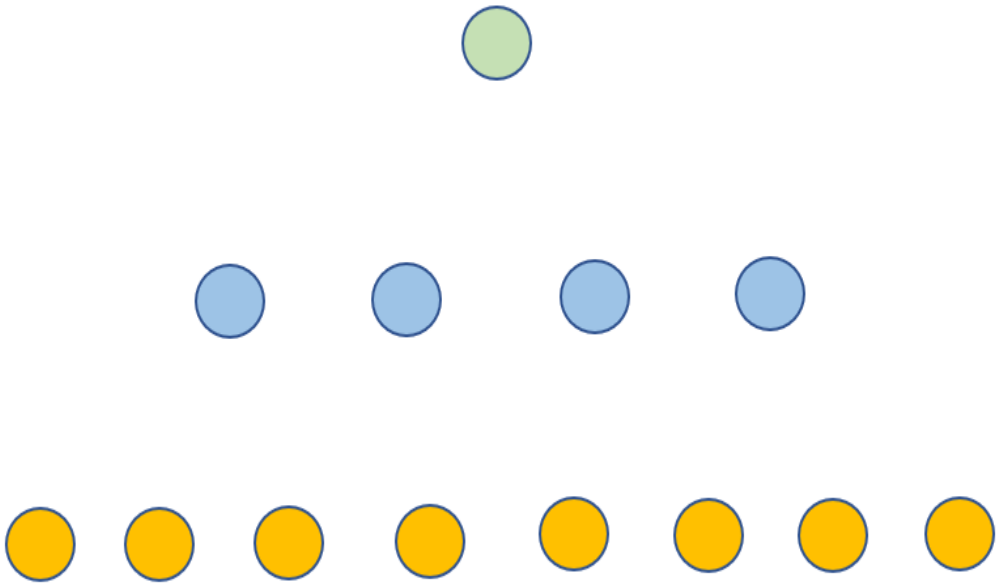


Fig 1: Model 1

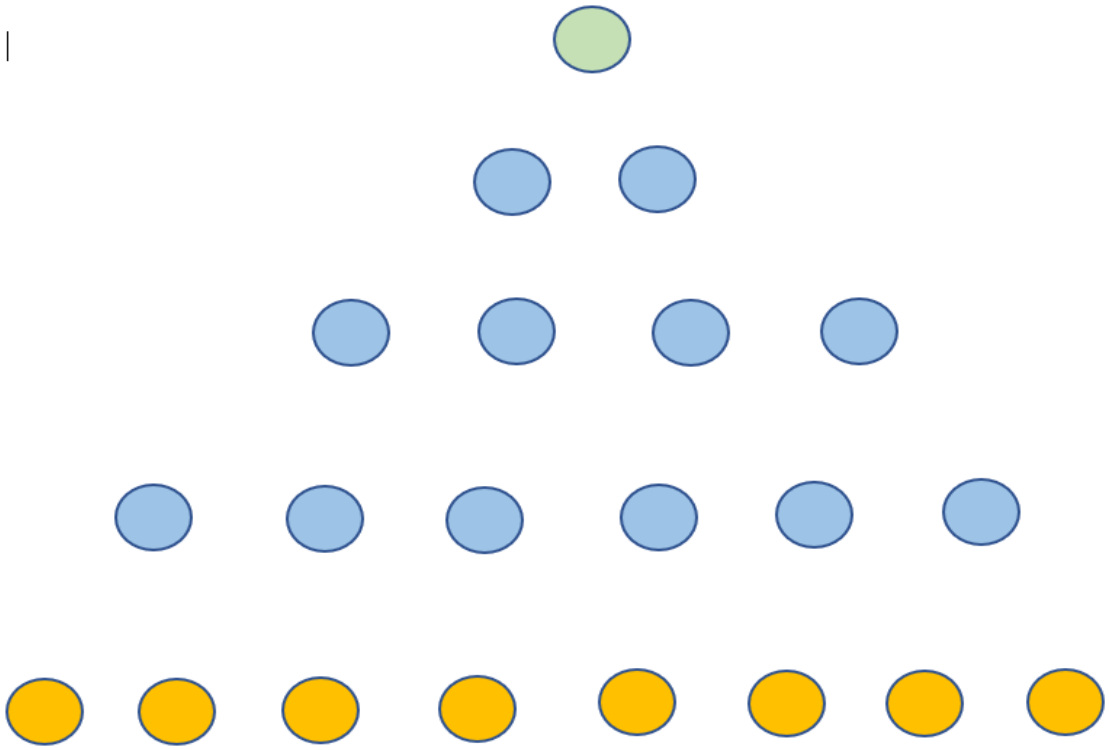


Fig 2: Model 2

4. METHODOLOGY

I. **Dataset Collection and Pre-processing:** Dataset used is Pima Indians diabetes dataset. Dataset was downloaded form Kaggle datasets. There are 768 rows and 9 columns. Out of 9 columns, 8 columns are used for prediction and 9^{th} column is outcome. Therefore "Outcome" is separated and stored in different variable. Pandas data-frame is then converted to tensor for future computations. And finally train-test splitting is done.

II. **Building Model:** Programming platform used is Google Colab. For Building model Py-torch framework has been used. Model is build using nn.sequentials from py-torch library. The model consist of different number of layer such as Linear layer and sigmoid layer. Linear layer has different number input and output neurons as parameters whereas sigmoid layer is applying activation on the neuron. The number of neurons in the first linear(input layer) layer is decided on various number of inputs and the number of neurons in the last layer(output layer) is one as the problem statement is based on Binary classification. Between the input layer and output layer there are some hidden layers. In this neural network every neuron in the previous layer is connected to every other neuron in the next layer. Then this model is trained and tested and accuracies are predicted.

III. **Training and Testing:** For function for training data passes through the model for given number of epochs and loss is computed and updated after every epoch. Model can be trained to best accuracy by changing some hyperparameters like learning rate, number of epochs, optimizers, Activation functions, etc. Stochastic Gradient descent algorithm has been used for forward pass. Accuracy function is used to find the accuracy of the model.

5. IMPLEMENTATION

Implementation of the code has been done on Google Colab. Framework used for building model is Py-Torch.

Model 1:

```
[19] class FirstNetwork_1(nn.Module):

       def __init__(self):
         super().__init__()
         self.net = nn.Sequential(

             nn.Linear(8, 4),
             nn.Sigmoid(),
             nn.Linear(4, 1),
             nn.Sigmoid()
             #nn.Softmax()
         )

       def forward(self, X):
         return self.net(X)
```

Model 2:

```
[20] class FirstNetwork_2(nn.Module):

       def __init__(self):
         super().__init__()

         self.net = nn.Sequential(

             nn.Linear(8, 6),
             nn.Sigmoid(),
             nn.Linear(6, 4),
             nn.Sigmoid(),
             nn.Linear(4, 2),
             nn.Sigmoid(),
             nn.Linear(2, 1),
             nn.Sigmoid()
             #nn.Softmax()
         )

       def forward(self, X):
         return self.net(X)
```

Fit function – Used for computations

```
[22] def fit_v1(epochs = 10000, learning_rate = 0.01):
       loss_arr = []
       opt = optim.SGD(fn.parameters(), lr=learning_rate,momentum = 0.01)
       loss_new = nn.MSELoss()

       for epoch in range(epochs):
         y_hat = fn(X_train)

         loss = loss_new(y_hat, Y_train)
         loss_arr.append(loss.item())


         loss.backward()
         opt.step()
         opt.zero_grad()


       plt.plot(loss_arr)
       plt.show()
       print('Loss before training', loss_arr[0])
       print('Loss after training', loss_arr[-1])
```

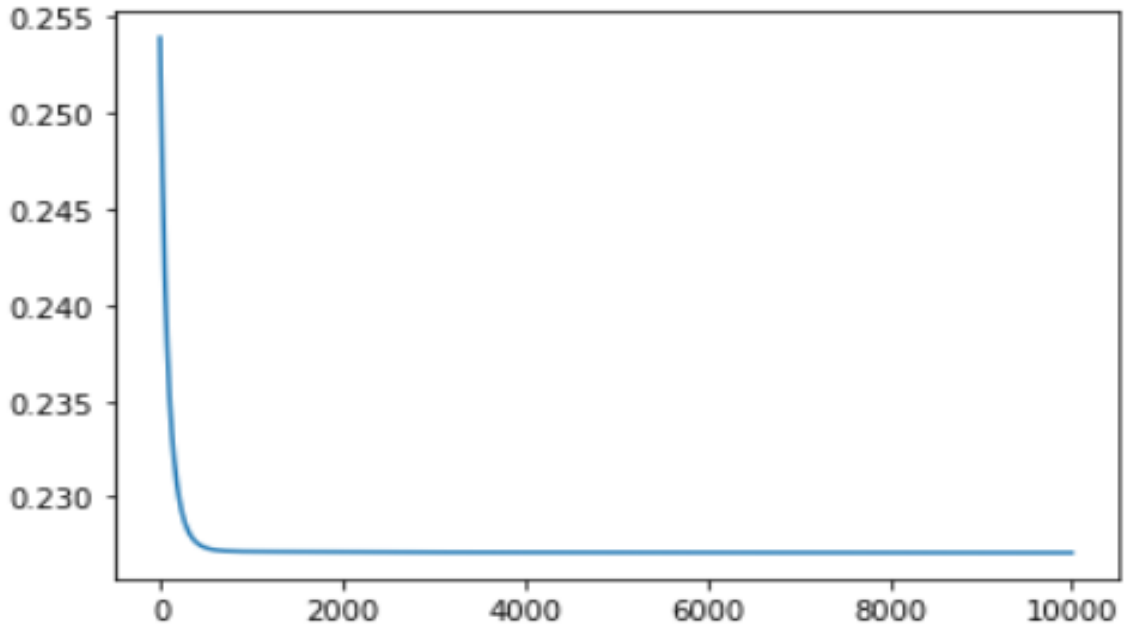
Evaluation Function:

```
[24] def evaluation(input,label,model):
         correct = 0
         total = label.shape[0]

         output = model(input)
         for x,y in zip(output,label):
             if(x<0.5 and y==0):
                 correct+=1
             elif(x>=0.5 and y==1):
                 correct+=1

         return 100 * correct / total
```

6. RESULTS

The best model out of 2 is 2nd model which gives accuracy of **Traning Accuracy** : 65.14657980456026 and **Test Accuracy:** 64.93506493506493

Loss Plot :




```
Loss before training 0.2539098262786865
Loss after training 0.22707775235176086
Time taken:  15.93601942062378
```

7. CONCLUSION

In the given problem statement we have used Deep Learning for building a suitable model in order to predict diabetes at early stage. Both the models were build using same techniques differing in the number of hidden layers. For better accuracy we need to change some hidden layer parameters also and hyperparameters such as learning rate, number of epochs, optimizers, Gradient Descent algorithms, etc. From the results Model 2 was giving good accuracy in-order to predict the disease. Prediction was done based on certain characteristics of individual's health.