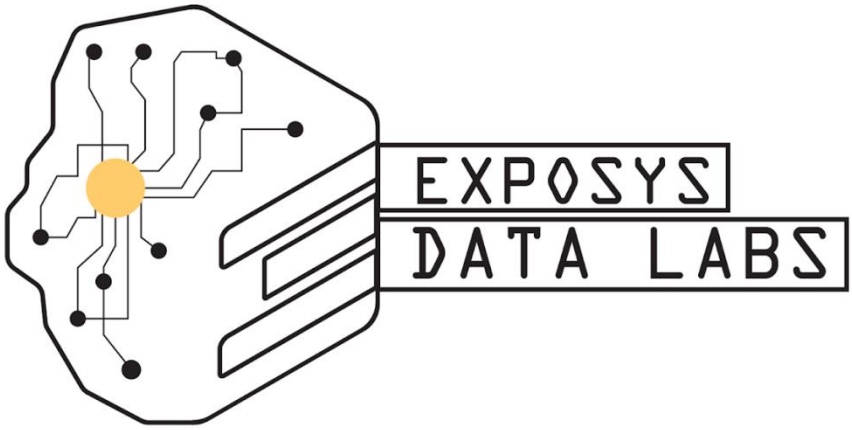

DIABETES DISEASE PREDICTION

REPORT

SUBMITTED BY:

PRATHMESH SALUNKHE

pssalunkhe@mitaoe.ac.in

ABSTRACT

This report addresses the prediction of diabetes diseases using various data science and deep learning algorithms. In India, around 72.96 million cases of diabetes are found in adults. India is second most affected country by diabetes in the world. Diabetes is disease in which the glucose level becomes high. It becomes very hard for diabetes patients to do day to day tasks. Diabetes doesn't have a cure, but if detected in early stages it can be reversed. Doctors have said that if diabetes is detected in early stage and patient follows the advice of your doctors and nutritionist, diabetes can be reversed. So it is very important to detect diabetes in early stages. We have used data science and deep learning to predict disease using available datasets. Using this model we can predict whether a person has diabetes by supplying some fields to the model. This can be were useful to detect the possibility of disease in early stages.

TABLE OF CONTENT

1. INTRODUCTION

India is the second most affected country by diabetes in the world. Diabetes detection at early stage can be helpful for a person to reverse diabetes or reduce the impact on the life of patient. A diabetic person has various restrictions on his life style. It is one of the world's fastest growing public health issues as well as one of the leading causes of death. Diabetes can be transferred genetically also. According to the recent facts and figures from International Diabetes Federation (IDF), about 463 million people are living with diabetes worldwide, with that figure expected to rise to 578 million by 2030 and 700 million by 2045 which accounts for an increase in prevalence by 51% in future(IDF Diabetes, 2019). This statistics is further exacerbated in Asian countries particularly India, which ranks second after China in terms of the number of people living with diabetes. One out of every six diabetic adults worldwide belongs to India. The prevalence of diabetes among people aged 20–79 years is estimated to be 88 million with that number expected to rise to 115 million by 2030 and 153 million by 2045 (International Diabetes Federation, 2019). These startling statistics point towards the emergence of a diabetes epidemic. Diabetes comes along with additional complications such as diabetic nephropathy, retinopathy, neuropathy, and angiopathy.

1.1 MOTIVATION

Due to the rise in diabetes cases in our country, our country can soon see a diabetes epidemic. It is difficult to identify diabetes in early stages. If identified, it can be very helpful. The signs and symptoms are easily neglected by people. Also people don't tend to go for testing for diseases out of fear, especially in India. Without proper identification of diabetes disease, disease control is not possible. Few sections of our society can't even afford to consult a doctor for such diseases. With the growth of artificial intelligence technology, disease prediction has been more accurate. Various systems are built to predict diseases using data science and machine learning models. Same can be done to predict diabetes efficiently. This model will be easily accessible to all and also be cost efficient. More and more people can check the possibility of diabetes in there body, helping them to fight with it early.

1.2 PROBLEM STATEMENT

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

1.3 OBJECTIVE

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

2. EXISTING METHOD

Diabetes is predicted using the glucose level in the blood. The main method is fasting plasma glucose test. In fasting plasma glucose test, the blood sugar level is testing after overnight fasting. Another test used to detect diabetes is A1C test. In A1C test the blood sugar level over past 2 – 3 months is measured. The third test is Glucose tolerance test in which the blood sugar level is tested before and after drinking a liquid containing Glucose.

Result	A1C	Fasting Blood sugar test	Glucose Tolerance test	Random blood sugar test
Diabetes	6.5% or above	126 mg/dL or above	200 mg/dL or above	200 mg/dL or above
Pre-diabetes	5.7% – 6.4%	100 – 125 mg/dL	140 – 199 mg	N/A
Normal	Below 5.7%	99 mg/dl or below	140 mg/dL or below	N/A

Table 1: Various tests for diabetes detection

3. PROPOSED METHOD WITH ARCHITECTURE

Deep learning has shown immense growth in past decade. We have used deep learning models for the prediction purpose. Using these models we can predict the possibility of diabetes in a person. The model requires some input fields related to the persons such as weight, age, blood pressure etc. These fields can be easily determined for any patient. Using these input fields the model will give an output which is a probability of diabetes in a patient. We used Pima Indians diabetes database for training out model. The dataset has 8 input columns and an output column. The dataset is visualized and cleaning is performed using Pandas library. After this the dataset is passed to the model. Model consists of input layer with 8 input nodes and an output layer with one output node. In between these two layers there are hidden layers. The specifications of hidden layer are needed to be tuned to get maximum accuracy. We tested different architectures with different hidden layers and got different accuracies. After training model, we use this model for prediction of diabetes in a person by giving required fields to the model.

3.1 Model Architectures

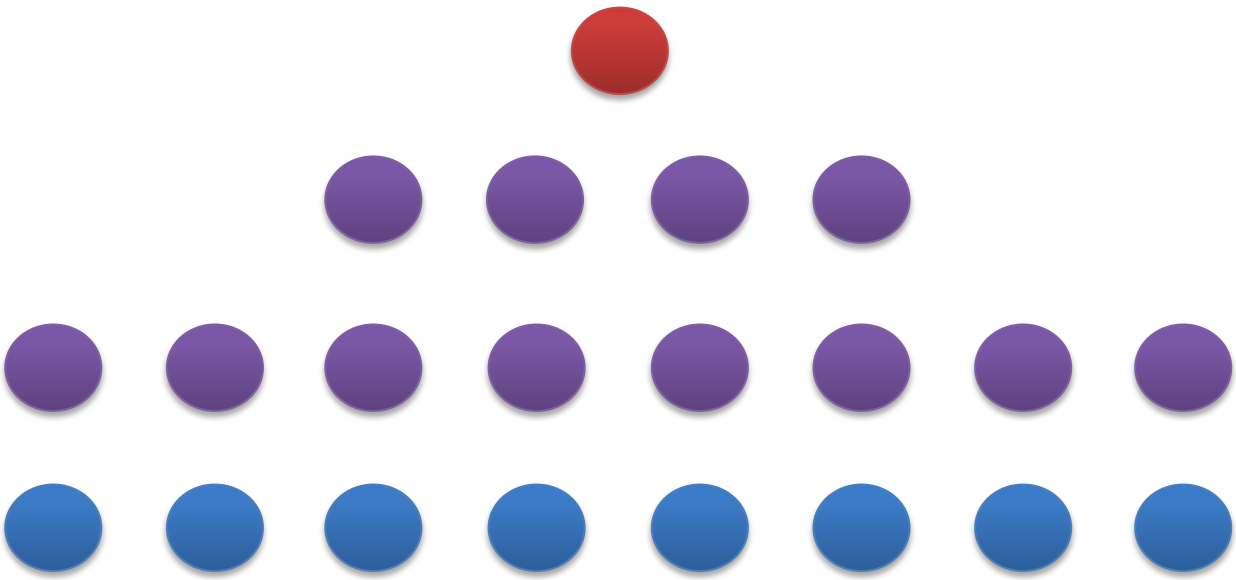

Fig 1: Architecture of model 1

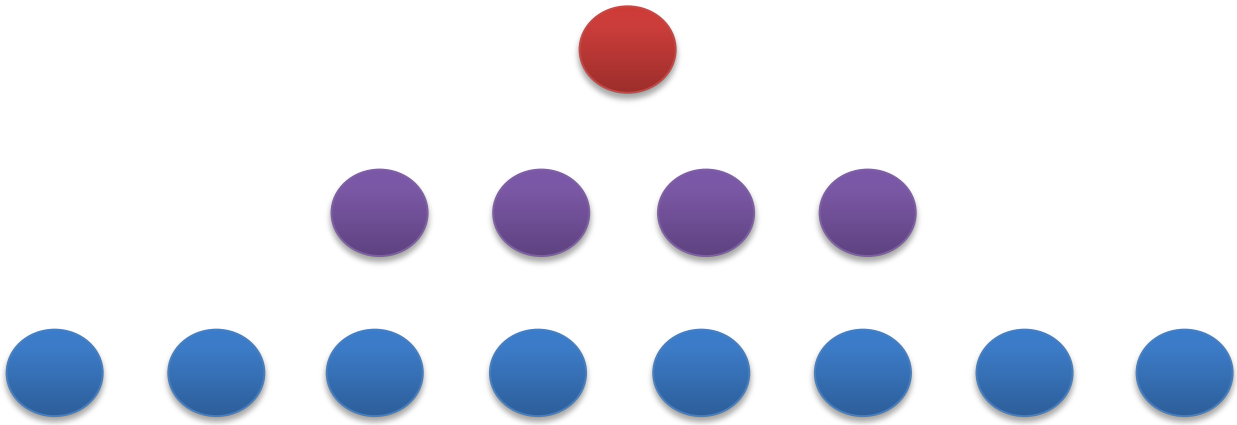

Fig 2: Architecture of model 2

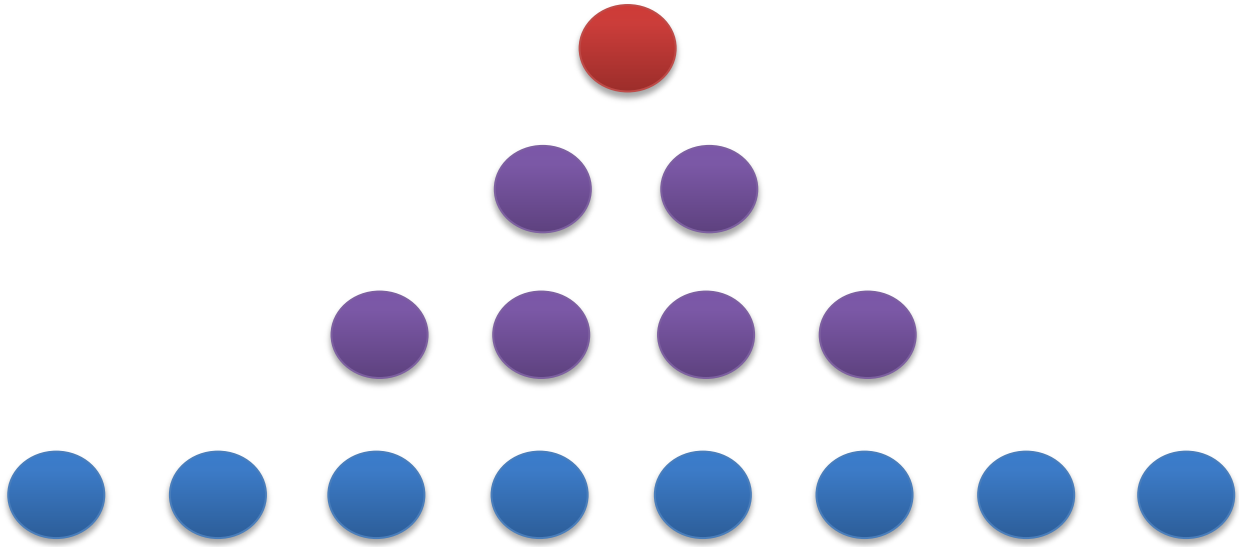

Fig 3: Architecture of model 3

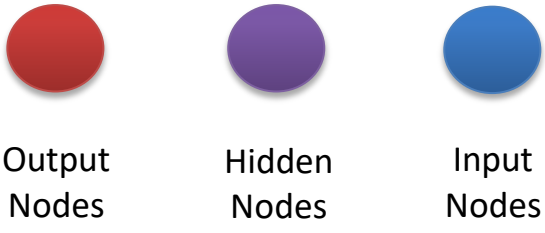

Fig 4: Architectural notations

4. METHODOLOGY

The complete system is divided into different smaller tasks. At first we need to gather as much dataset possible for the project. The dataset we used is from Pima Indians diabetes dataset which was available on Kaggle. This dataset had many empty entries. These empty entries were replaced by median values of that column. Dataset visualization is performed using available libraries, and the correlation between the various columns of dataset is derived using libraries. After dataset cleaning and visualization, the dataset is used to train our model. The dataset is split into three parts i.e. train, test and validation for different purposes. The model is built using deep learning principles. We tried different architectures for our model. The models are then trained on the train dataset. We are using GPU for faster fitting of model as it provides high computational power as compared to CPU. After training the model we test the accuracy of model on the test dataset. The model hasn't been trained on test dataset so it provides an unbiased accuracy. Hyper parameter tuning is done using the test accuracy results on the model. After the hyper parameters are tuned then the final validation accuracy is calculated. Then this model can be hosted and used for predicting the diabetes in a patient.

5. IMPLEMENTATION

We used Google Colab for implementation as it provides GPU service. We used python programming. PyTorch framework is used for model building and Pandas library is used for dataset cleaning.

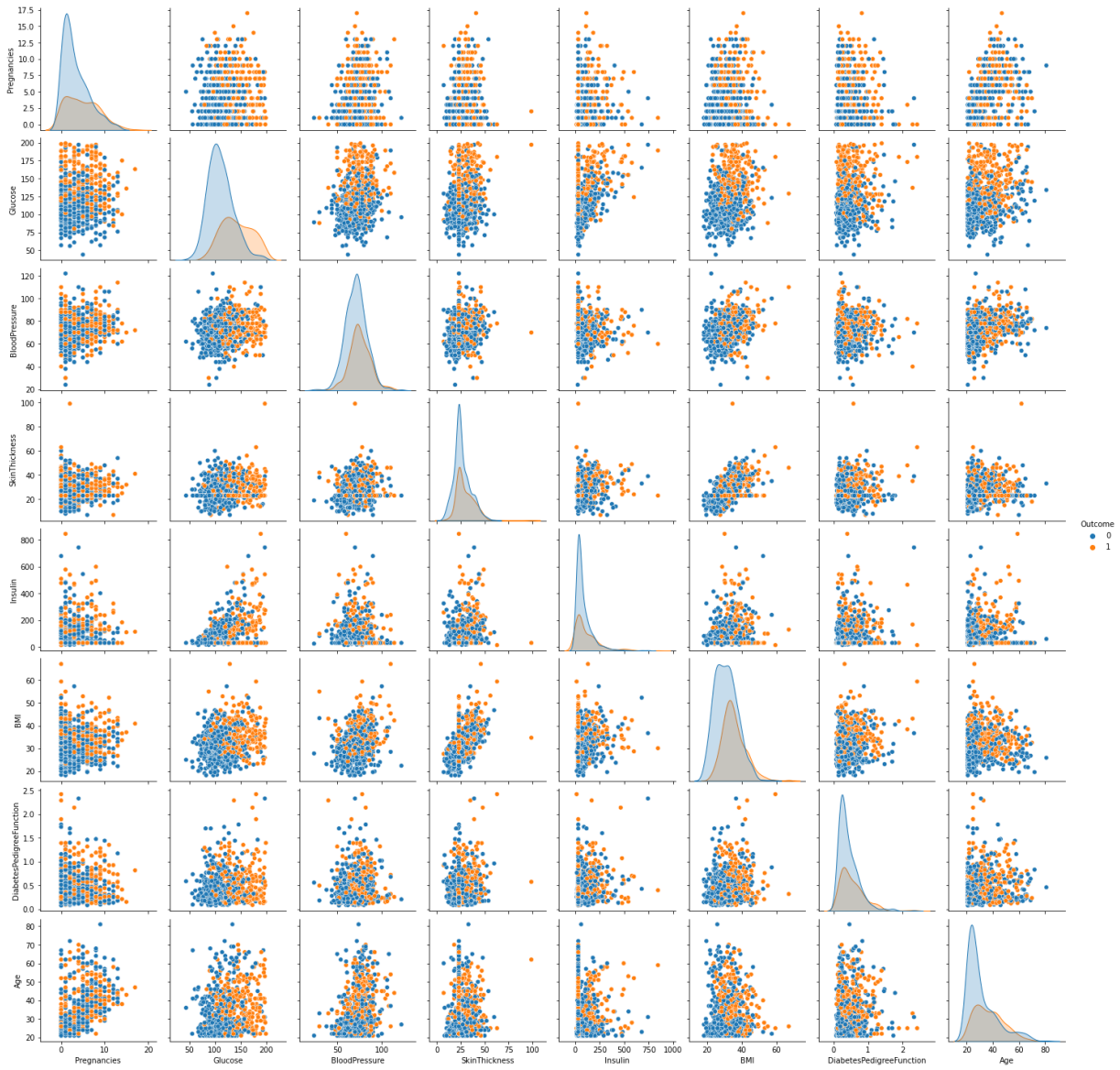

Fig 5: Pair plots of dataset columns

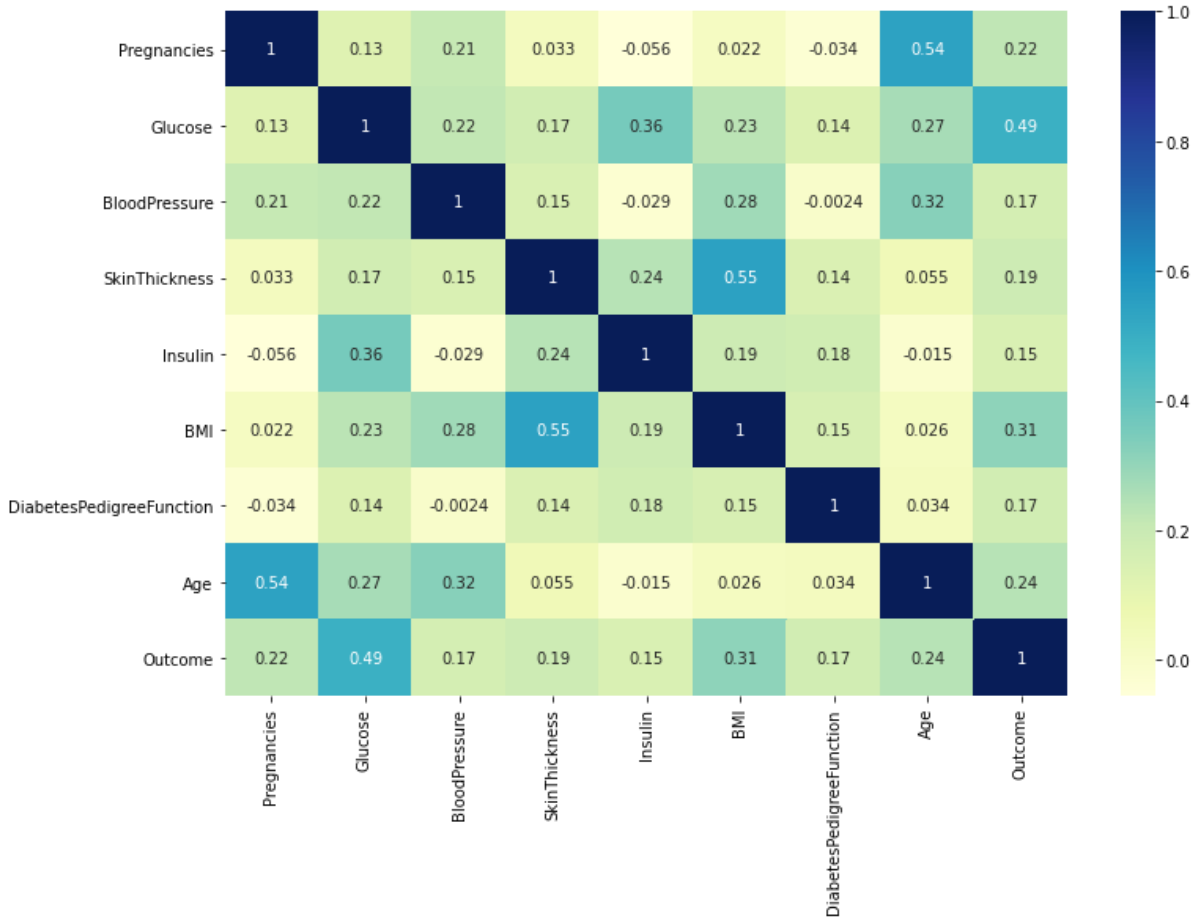

Fig 6: Correlation between dataset columns

5.1 SOFTWARE REQUIREMENTS

- Programming language: Python 3
- Platform: Google Colab
- Framework: PyTorch 1.9.0v
- Plotting: Mathplotlib 3.2.2v, Seaborn 0.11.1v
- Dataset analysis/cleaning: Pandas 1.1.5v, Sklearn 0.22.2v, Numpy 1.19.5v

5.2 RESULTS

Implementation of model using PyTorch we got a maximum train accuracy of accuracy of 77% and test accuracy of 70%.

6. CONCLUSION

In this project we have used data science and deep learning algorithms to predict possibility of diabetes in a person. Early detection of diabetes is helpful for reduction of effect of the diabetes disease on a person's life. Conventional methods for diabetes detection are accurate but people ignore the signs and symptoms of disease. We have built a model which can be used by providing minimum input fields related to patient. This model can help people to detect diabetes in its early stages.