PROJECT REPORT

Diabetes Prediction

Written by	Shubham Luharuka
Document Version	0.1
Last revised date	17-01-2021

Document control

Version	Date	Author	Comment (By viewer)
0.1	17-10-2021	Shubham Luharuka	

ABSTRACT

Diabetes, a disease caused when there is deficiency of insulin in our body. Insulin is produced by pancreas which control the blood glucose level in the body and prevent body from hyperglycemia. Our *daily activity, food and routine effect the production of insulin in our body. Junk foods and sugar reduce the production of insulin. In this paper I used PIMA dataset to predict whether a person is healthy or diabetic. Different multiclass classification algorithms like logistic regression, Support vector machine (SVM), decision tree, boosting algorithm like Adaboost and ensemble learning methods are used. I also implemented one neural network to predict whether the person is healthy or diabetic.*

KEYWORDS

PIMA Dataset, Ensemble Learning Methods, Gaussian Distribution, Boosting Algorithm, Data Mining, Neural Network, Diabetes Prediction, Artificial Neural Network.

Table of Contents

1. INTRODUCTION

In today’s world, youngsters are liking junk food. They are preferring soft drinks which contains a lot of sugar or alcohol. They are ignoring the fact it can put a bad effect on their body. Diabetes is one of the most common diseases. It occurs when there is a lack of insulin in our body or blood glucose level increases. Insulin is a hormone produced by the pancreas to neutralize the blood glucose level. The risk of hyperglycemia increases with the level of blood glucose. There are 3 types of diabetes. a) Type1 b) Type2 c) Gestational diabetes. If the body is not able to produce insulin it is type1. Generally, these types of diabetes are found in children and young adults.

If the body is not able to produce insulin as well as not able to use the insulin then it is of type2. Middle-aged and older people are targeted by this type. Lastly, Gestational diabetes develops in pregnant women. Generally, it goes away after the baby birth but later has the chance to develop type2. Different other health problem comes with diabetes like heart diseases, kidney problems, eye problem and many more. The chance of getting stroke also increases with high glucose levels in our body. Junk food contains a high per cent of trans and saturated fats which can increase the level of triglycerides. As you can see in Figure 1 the number of diabetic persons increases over the past 5 years in many countries, especially in developed countries.

In this project, I used multiclass algorithms like logistics regression, Support Vector Machine (SVM), Decision Tree, and Adaboost to predict whether the person is healthy or diabetic. I used different combinations of machine learning algorithms applied with Ensemble learning. In last I tried to make one neural network to classify the healthy and diabetic person.

Pima Indians Diabetes Database [4] from the National Institute of Diabetes and Kidney diseases is used for this project and has 8 feature input vectors and the last column is the Boolean value (1 or 0) which tells whether the person is diabetic or healthy. The 8 input features are Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, DiabetesPedigreeFunction, Age, and the result vector is named Outcome. In this dataset, all patients are females at least 21 years old. There is a total of 2000 data in this dataset. This dataset contains 1316 data of healthy people and 684 data of the diabetic patient.

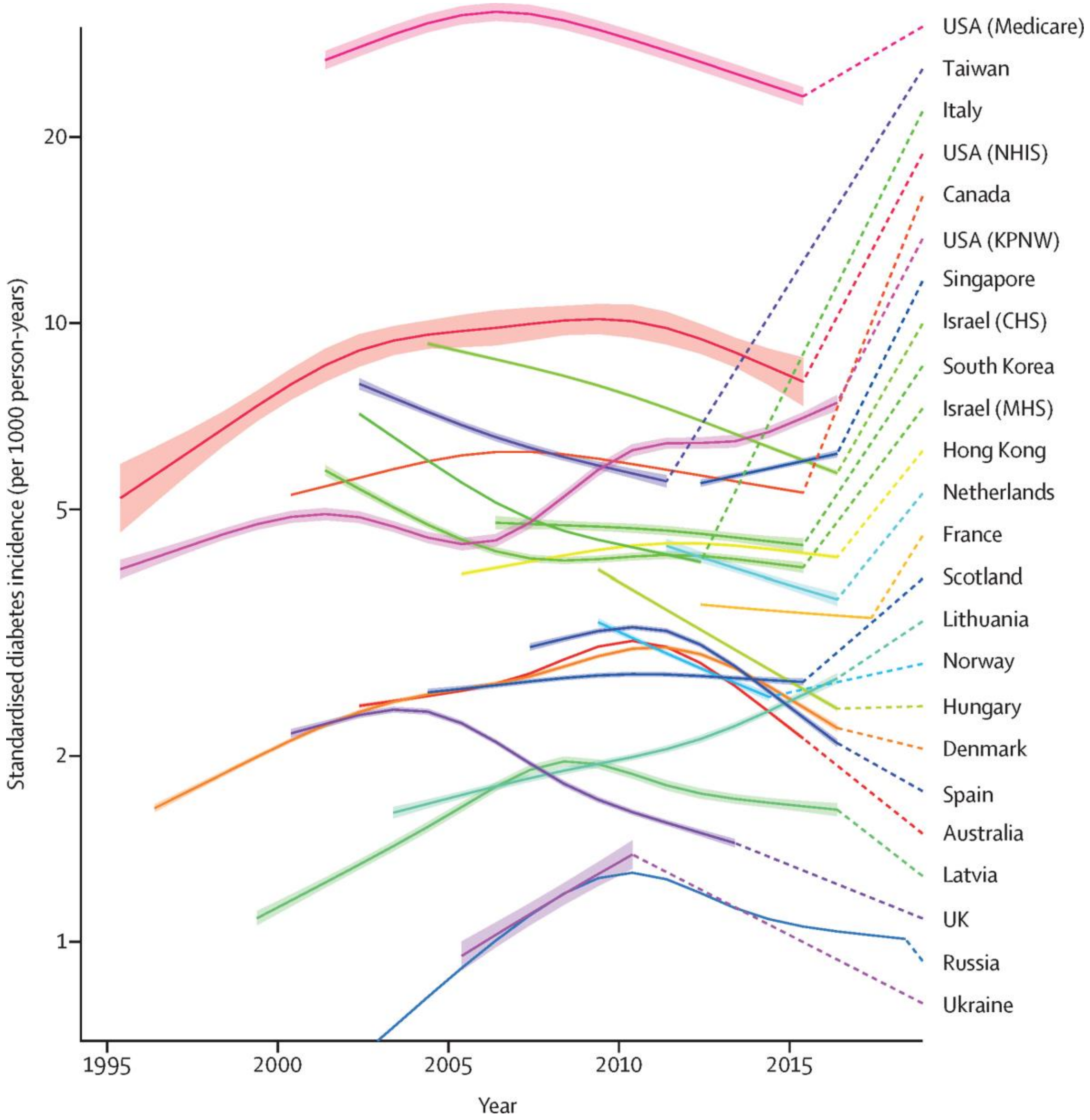


Figure 1: - Trends in the incidence of diagnosed diabetes: a multicountry analysis of aggregate data from 22 million diagnoses in high-income and middle-income settings (From: -The LANCET)

2. EXISTING METHODS

Vandana C Bavkar et.al., [1] got 72.22 % accuracy with SVM classifier, 78.57 % with Naïve Bays and 70.35 % with KNN classifier. Their algorithm achieved maximum accuracy of 89.97% with Decision Tree classifier. They used one other dataset which is created in vivo experiment on 182 patients having diabetes and not. Different other features like regression coefficient, peak distance, power spectral density, pulse transit time extracted from PPG signal and feed into neural network. These all features are different in time and frequency domain.

Huma Naz et.al.,[2] achieved the accuracy of 98.07% with decision tree classifier. They implemented other algorithms also like Naïve Byes which give an accuracy of 76.33. They also proposed an Artificial Neural Network which is able to predict the diabetic and healthy person with 90.34% accuracy.

Souad Larabi-Marie-Sainte et.al.,[3] achieved a best accuracy of 74.48% with REPTree. In this paper she used weka software tools to evaluate diabetes prediction. They used others algorithm like Kstar, OneR, PART, SMO, BayesNet which give accuracy of 68.23, 70.83, 74.35, 72.14, 73.83 percent respectively. Null values in the dataset are filled with the average values.

3. PROPOSED METHOD

In this paper I proposed a method to classify the healthy person and diabetic person from Pima Indians Diabetes Database from National Institute of Diabetes and Kidney diseases. It consists of 2000 data of female patients. Before implementing any algorithm, I did some data preprocessing. It consists of many duplicates' values. So first I removed all the duplicates values. I am doing this because after doing train test split, there is chance that duplicate values will come in train as well as test set also. It will not able to make model perfect. In this dataset some columns like Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Age can be equal to zero. Therefor I change it will null to make our process smooth. Box Plotting of dataset features show there is some outliers present in the dataset. I also change the values of outliers to null. Finally, all null values are replaced with the mean of that feature. Some features are converted to their square root values so that I get the gaussian distribution of the dataset.

In last whole dataset's feature is standardize by removing the mean and scaling to unit variance. Data preprocessing completed here. Now I feed this data into the different algorithm. I tested our algorithm with different train-test split ratio, from 70:30 to 95:05. After pre-processing the dataset feed in the Artificial Neural Network (ANN) and tried to get the accuracy more than other classifiers.

4. ARCHITECTURE

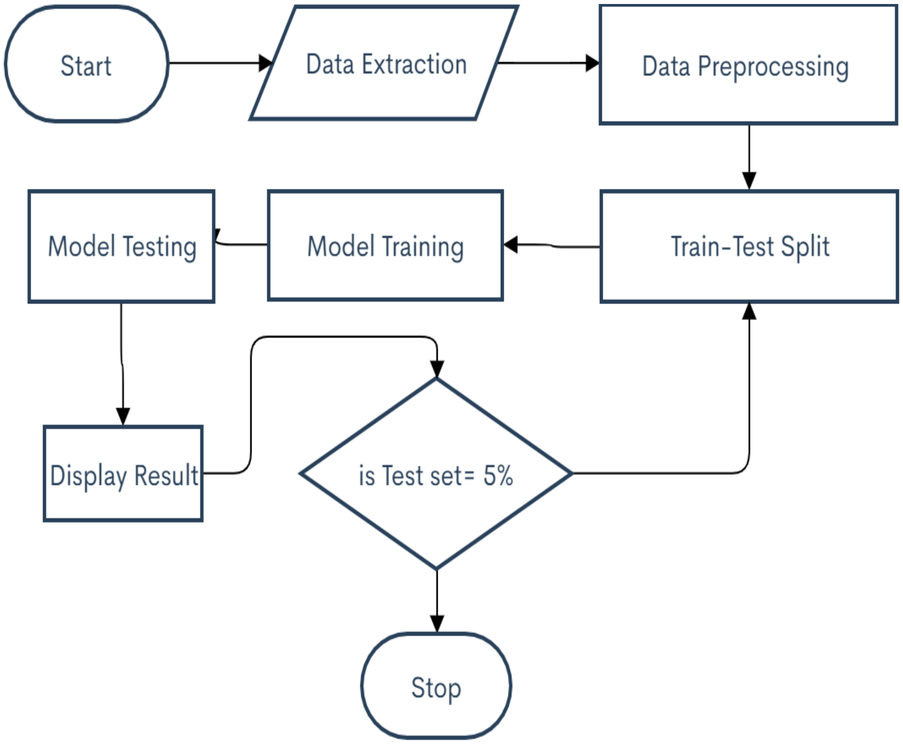
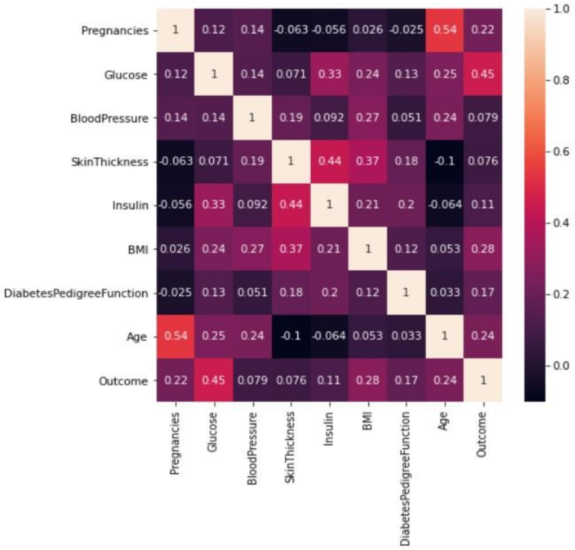


Figure 2: Flow Chart of the proposed model

5. METHODOLOGY

5.1 Data Collection: Dataset used in this project is the Pima Indians Diabetes Database [4] extracted from Kaggle website. It has 2000 data of female patients. Some are repeated some times. So, after removing duplicated data from dataset, I have 744 unique data.

Out of which 491 is the person who is healthy or non-diabetic and 253 patients are diabetic. Figure (3) is the covariance matrix of the dataset. From this matrix we can see that on Outcome, Blood Glucose level put a maximum effect. Blood Pressure and Skin Thickness put less impact on Outcome vector.



	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
Pregnancies	1	0.12	0.14	-0.063	-0.056	0.026	-0.025	0.54	0.22
Glucose	0.12	1	0.14	0.071	0.33	0.24	0.13	0.25	0.45
BloodPressure	0.14	0.14	1	0.19	0.092	0.27	0.051	0.24	0.079
SkinThickness	-0.063	0.071	0.19	1	0.44	0.37	0.18	-0.1	0.076
Insulin	-0.056	0.33	0.092	0.44	1	0.21	0.2	-0.064	0.11
BMI	0.026	0.24	0.27	0.37	0.21	1	0.12	0.053	0.28
DiabetesPedigreeFunction	-0.025	0.13	0.051	0.18	0.2	0.12	1	0.033	0.17
Age	0.54	0.25	0.24	-0.1	-0.064	0.053	0.033	1	0.24
Outcome	0.22	0.45	0.079	0.076	0.11	0.28	0.17	0.24	1

Figure 3: - Covariance matrix of the dataset

5.2 Data Preprocessing

Some features values can’t be zero. In dataset all feature columns contains the zero values except DiabetesPedigreeFunction column. Pregnancies column can contain zero value. So, I converted all the data from of the feature columns except DiabetesPedigreeFunction and Pregnancies with null to make our process smooth. Column Insulin contains a lot of null values.

Figure 4. is the box plot diagram of features. We can see values Pregnancies, SkinThickness, Insulin, BMI, DiebetesPedigreeFunction and Age feature is skewed. There are some outliers in this section. I placed the null values in the outlier's features.

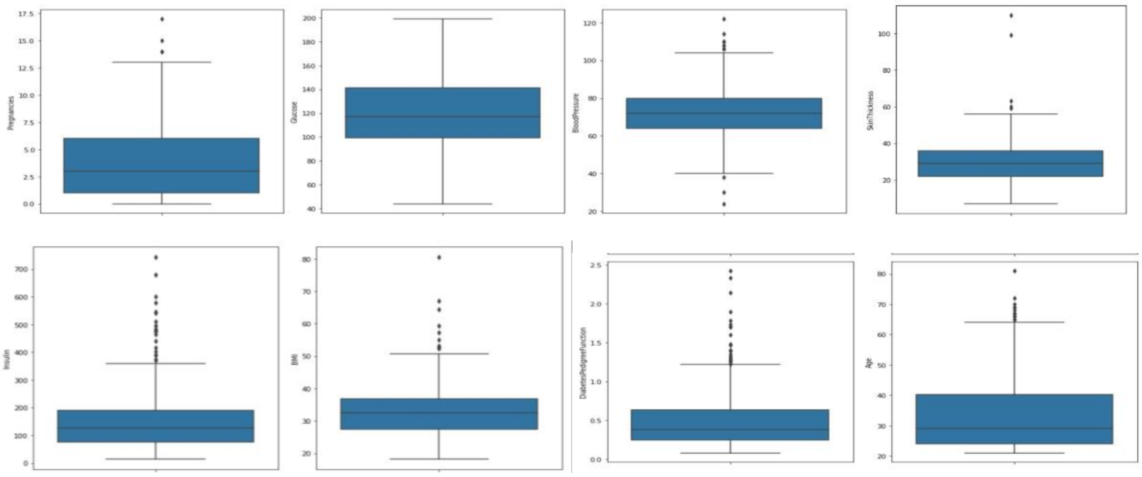


Figure 4: Box Plot of Columns in PIMA dataset.

In Figure 5, you can see the gaussian distribution of data. Some features like Glucose, BloodPressure, SkinThickness, BMI are normally distributed, but other features like Pregnancies, DiabetesPedigreeFunction, Age, Insulin are not. There are several methods to convert the features into normally distributed function like log, square root, exponential etc. I used square root function to change the feature value of columns which is not normally distributed.

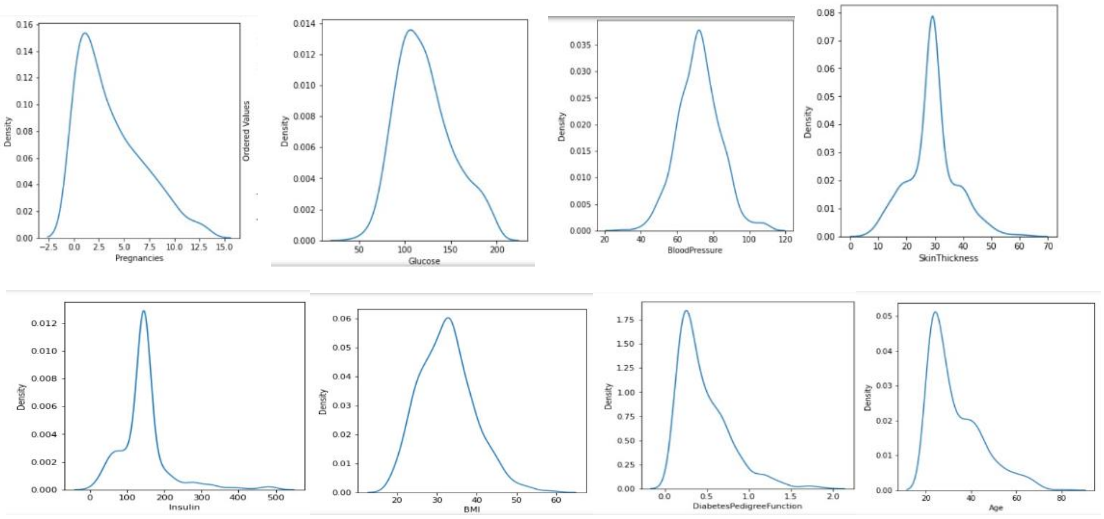


Figure 5: Gaussian Distribution of all feature in PIMA dataset.

After that I filled all null values with mean values of that features. Now my data is ready to train the classifiers and ANN.

5.3 Training of model:

For this project I took Logistic Regression Classifiers, Linear Support Vector Machine (LSVM), Decision Tree (DT), Adaboost Classifier, K-Nearest Neighbour (KNN), Artificial Neural Network (ANN), and different combinations of above classifiers with ensemble learning methods. Adaboost is a boosting algorithm and there is a challenge to overcome from its overfitting nature. I checked different models for different train-test ratio. Starting with test of size 30% of total size, I achieved maximum accuracy of 80 % with ANN 78 % from Logistic regression.

6. IMPLEMENTATION AND RESULT

6.1 Tools used for Implementation purpose

These are special libraries and frameworks which I used in python3 programming language.

- Pandas (1.3.1)
- NumPy (1.19.5)
- Sk-Learn (0.24.2)
- TensorFlow (2.4.0)
- Matplotlib (3.4.2)
- Seaborn (0.11.1)

6.2 Results and Evaluation

This whole project is developed on the normal PC with 8 GB RAM, Intel Core i5 Processor with 4 GB Nvidia GPU GEFORCE GTX 1650 Ti. Pre-processed data is used to trained the model. With Logistic Regression maximum training accuracy was 79.81 % and testing accuracy was 78.21 %, which was achieved when test size is 30 %.

With LSVM maximum training accuracy was 78.5% and testing score was 76.41 % when test size is 25% of dataset. Decision Tree gives the maximum training accuracy of 98.65% but testing score was only 72.56 %. KNN was able to achieved the training accuracy of 85.41% with the testing accuracy of 75.65% when test size is 5%.

Adaboost is the boosting algorithm. Basically, boosting algorithm are a set of the low accurate classifier to create a highly accurate classifier. It can keep track of the model who failed the true prediction. Through Adaboost I am able to get 99.83 % of accuracy with 72.88 % of testing accuracy.

Stacking methods of ensemble learning methods are used with different combinations of above classifiers. Table (1) is showing the combination used, maximum training accuracy, testing accuracy and the train-test split ratio of 70:30.

Table 1 : Different combination in ensemble learning classifier and attained accuracy.

Combinations	Best. Training Accuracy	Testing Accuracy
Logistic Regression and LSVM	0.7715	0.7500
Logistic Regression and Adaboost	0.8538	0.7823
LSVM and Adaboost	0.8558	0.75
Linear Regression, LSVM and Adaboost	0.8134	0.759

Finally, my proposed model, an ANN was able to attained an accuracy of 84.01% with testing score 80.23%. This is achieved when the train-test split ratio was 70:30. In my proposed model There is one Input layer takes total 8 features as input. There is two hidden layers of nodes. One is of size 264 and other is of size 132. Rectified Linear Unit (ReLU) is the activation function used by these layers. In last a dense layer of 2 nodes gives the probability distribution of weather a person is diabetic or healthy. Table (2) is showing summary of our proposed ANN.

Table 2: Type of layer, their respective parameter and output shape

Type of Layer	**Output Shape**	**Parameter**
Input Layer	8	-
Dense (activation=ReLU)	264	2376
Dense (activation=ReLU)	132	34980
Dense (activation=SoftMax)	2	266

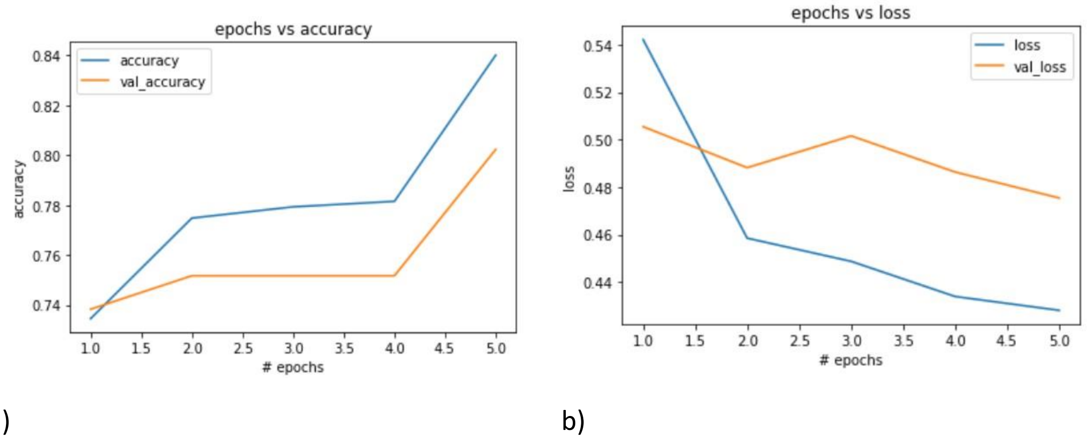


Figure 6: Result of proposed model attained by a) graph plotted between no of epochs and accuracy (left) b) graph plotted between no of epochs and loss (right).

As result, I got 80.23 % test accuracy. Figure (7) is the plot of confusion matrix of my result. Length of dataset used for testing is 224, out of which my proposed model is able to predict 121 True Negative and 57 True Positive value.

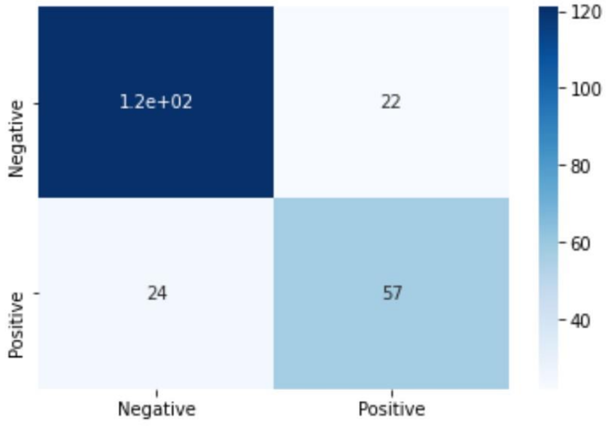


Figure 7: Confusion matrix of the predicted value.

7. CONCLUSIONS

As described earlier, the diabetic person is increasing every year. So, there is need of system that can predict the diabetes in early stage of life. If we maintain a proper daily routine, eat healthy food and exercise chance of being diabetic reduces. In my system I implemented many classifications algorithm along with one artificial neural network (ANN). Highest accuracy I achieved from ANN is 82.14 %. This accuracy is achieved when 80 % of randomly shuffled data is used as training and remaining for testing. In this neural network I achieved an accuracy of 92% if only 95 % of data is used for training. Only 744 data from dataset is used. In future we can do the data augmentation by GAN or a new dataset which has more features as well as available for men used to increase the accuracy of the system.

8. REFERENCES:

[1] Vandana C Bavkar, Arundhati A Shinde, “Machine learning algorithms for Diabetes prediction and neural network method for blood glucose measurement”, Indian Journal of Science and Technology 14(10): 869-880, 2021.

[2] Huma Naz, Sachin Ahuja, “Deep learning approach for diabetes prediction using PIMA Indian dataset”, Journal of Diabetes & Metabolic Disorders ,2020.

[3] Souad Larabi-Marie-Sainte , Linah Aburahmah, Rana Almohaini, Tanzila Saba, “Current Techniques for Diabetes Prediction: Review and Case Study”, 29 october 2019.

[4] Dataset source:- https://www.kaggle.com/uciml/pima-indians-diabetes-database