Noname manuscript No.
(will be inserted by the editor)

Diabetes Prediction & Analysis through Machine Learning

Badiuzzaman Pranto · Sk. Maliha Mehnaz · Esha Bintee Mahid · Imran Mahmud Sadman

the date of receipt and acceptance should be inserted later

Abstract Machine Learning has a significant impact in every aspect of science and technology which includes medical researches and life science. Diabetes Mellitus, more commonly known as diabetes is a chronic disease which involves abnormally high levels of glucose sugar in the blood. This paper has focused upon detection of diabetes using different machine learning approaches to build up a model based on PIMA Indian Data set. The model has been tested on an unseen portion on PIMA and also on the data set collected from Kurmitola General Hospital, Bangladesh. This will help us to understand the stability of the model as well as how well the model is performing on patients from our country. The process of building this model will also show us how we can make effective machine learning models and also the impact of different models on the detection of diabetes.

Keywords Machine learning · Diabetes prediction · Decision tree · K-NN · Gaussian Naive Bayes · Kurmitola General Hospital · Pima Indian Dataset

Badiuzzaman Pranto
Department of ECE
E-mail: prantoamt@gmail.com

Sk. Maliha Mehnaz
Department of ECE
E-mail: mehnazlasmi213@gmail.com

Esha Bintee Mahid
Department of ECE
E-mail: esha.mahid@northsouth.edu

Imran Mahmud Sadman
Department of ECE
E-mail: mahmud.sadman@gmail.com

1 Introduction

Diabetes is the root cause of many associated health diseases such as heart attacks, liver and kidney failure, nerves damage, diabetic retinopathy, diabetic neuropathy and Polycystic Ovary Syndrome(PCOS). PCOS has recently become a common occurring in women. Due to PCOS, many female are already suffering from diabetes at a young age. According to the International Diabetes Federation(IDF) atlas 2019 [6], 1 in 11 adults(20-79 years) have diabetes which can be approximated as 463 million people.According to the WHO report, in 2016, 1.6 million deaths occurred directly due to diabetes and in 2012, high blood glucose resulted in 2.2 million deaths. [10] Around 80lac people in Bangladesh already have diabetes. [5] A survey by the Bangabandhu Sheikh Mujib Medical University(BSMMU) on 2000 adults in Dhaka slums found out in 2016 that 19% adult among which 15.6 percent male and 22.5% female had suffered from diabetes. [7]Diabetes is mainly of three types:
a) Type 1 diabetes- It can occur at any age but most likely to happen to children and adolescents.In this case, the body produces very little or no insulin at all. As a result,daily insulin injections are needed to keep glucose levels under control. Frequent urination, sudden weight loss, abnormal thirst, constant hunger, blurred vision and tiredness are common symptoms. This can be treated by the help of insulin therapy.
b) Type 2 diabetes- It mostly occurs in adults and comprises of around 90 percent diabetes cases. The body does not fully respond to insulin resulting to higher glucose levels. Maintaining a healthy diet and increased physical activity can help. Obesity,unhealthy diet, high blood pressure, physical inactivity are the major risk factors. Insulin injections are required when oral medication is not sufficient enough to control the blood sugar levels.
c) Gestational Diabetes(GDM)- This diabetes consists of high blood pressure during pregnancy and can cause health complications to both mother and child. It usually disappears during the pregnancy stage but the affected ones along with their child have a risk of developing Type 2 diabetes in their later life. According to a survey on 2017, there were an estimate of 204 million women having diabetes. About 21.3million live births had some form of hyperglycaemia in pregnancy, among which about 85.1percent happened due to gestational diabetes. Gestational diabetes affected around 1 in 7 births.

2 Related Work

A variety of machine learning algorithms have been used to predict diabetes. Various classifiers have been used to detect diabetes such as Random Forest, IBK, J48, and fuzzy approaches. The dataset available at the UCI repository based on India has been mostly used for this detection. In one research paper [3] by Md. Aminul Islam and Nusrat Jahan on "Prediction of Onset Diabetes using Machine Learning Techniques" investigated different types of machine learning classification algorithms and make a comparative analysis.

Based on their results, Logistic Regression along with SMO and MLP could be used to predict the onset diabetes having an accuracy of 78.0%. Bagging performed well but it was not enough. In another research paper by Aakansha Rathore, Simran Chauhan and Sashki Gujral on "Detecting and Predicting Diabetes using Supervised Learning:An approach towards better healthcare for Women" [2] they used the Support Vector Machines(SVM) and Decision Tree to detect as well as predict the risk of diabetes. Detection was done successfully with the SVM classifier with an accuracy of 82%. Some frequently used techniques for the Diabetes Prediction are given below:

A. Support Vector Machine(SVM): SVM is a machine learning algorithm that is used for supervised learning. The data set is trained such that it gives the predetermined output. The main goal is to construct a hyper-plane that divided the entire data set into several categories. Thus the data mining and classification are effectively performed. [4] In some cases, the data sets are not linearly separable for which Kernel function is used to improve the model and tackle over-fitting. It is comparatively slow in execution. [2]

B. Principal Component Analysis (PCA): PCA is a supervised machine learning algorithm that classifies the data set such that there remains maximum co-relation. The goal is to create an orthogonal plane where the data can be classified, another plane belongs perpendicular to this plane. It helps in feature extraction and also uses the eigenvalues and eigenvectors [8] to find the principal component. The main disadvantage for the PCA classifier is it gives poor results in diabetes detection when used alone. [4]

C. Naive Bayes Classifier: It is known for its space efficiency. It assumes that the features in the data set are independent of one another. The conditional probability remains that it determines the likelihood of an event to take place given that some of the events have already occurred. The values are taken as average resulting in reduced noise, also giving rise to probability values resulting in higher accuracy. [4]

D. Decision Trees(J48): It is one of the best machine learning algorithms resulting in higher accuracy rates. A huge amount of data can be easily handled and also run efficiently. Until the criteria is met, the decision is run from root to leaf. The method of growing an ensemble of trees and letting them vote for the most popular class results in better accuracy. A disadvantage is that even with a small change in input, the decision tree gets unstable [4].

E. Artificial Neural Network(ANN): ANN can learn from input and predict the output. Combining ANN with fuzzy logic, uncertainties can be handled. But training for ANN requires more effort. Besides, it is difficult to guarantee whether all the inputs have been trained. [4]

F. Instance Based Learner(IBk): [1] Also known as KNN and lazy learning, it is a classification based machine learning method depending on the closest training example in the feature space which the value of K. A distance measure is required to determine the closeness of the instances. The instances are classified by finding the nearest neighbours and then the most popular class among the neighbours is chosen. [4]Although performance is very accu-

rate, it is comparatively slow. With large values of K, the performance also gets better.

3 Methodology

3.1 Data Pre-processing

Data set have been collected from Kaggle which is on PIMA Indians Diabetes Database. This data set was originally collected from the National Institute of Diabetes, Digestive and Kidney Diseases based on India. Alongside, we collected more real data from the general ward of Kurmitola General Hospital(KGH) based on Bangladesh in order to predict diabetes.
An important factor in both the datasets was that it was based on female patients whose number of times of pregnancy is one of the important attribute for our analysis. We have used different Machine Learning algorithms on this particular data to predict whether a patient is suffering from Diabetes or not.

Data set	Number of instances	Number of features	Positive	Negative
PIMA Indian	768	8	268	518
Kurmitola Hospital	182	5	50	131

Table 1: Data set summery

Firstly, we analyzed the data for errors and found some issues in general. A total 8 attributes were present containing 768 instances in the data set of PIMA. These 8 particular attributes are considered as important factors and the reason behind the occurrence of diabetes.

SL	Attribute Name
1	Number of pregnancy
2	Glucose concentration
3	Blood Pressure
4	Skin thickness
5	Serum Insulin
6	BMI
7	Diabetes Pedigree Fuction
8	Age

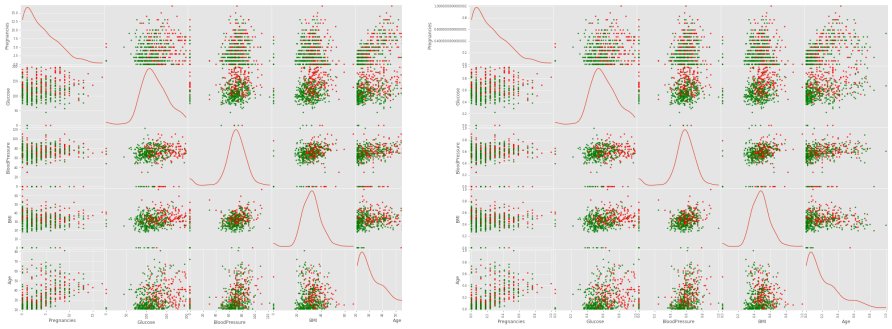
Table 2: Pima data set Attributes

Data in the real world are mostly inconsistent or unclean. And it is important to pre-process those data either by filling missing values or by removing them. Kurmitola General Hospital's data set had the attribute, Glucose in

mmol/L so we converted that value into mg/dL to match with the data set from PIMA. The data set of PIMA had around 374 null values for the attribute, Insulin. Moreover, 227 null values were found for the attribute Skin Thickness. Skin Thickness, Insulin and Diabetes Pedigree Function, all three attributes had a huge number of zero entries in the PIMA data set. On top of that, the data we have collected from Kurmitola General Hospital have fewer values available for those three attributes and most of them were null. One way to handle this error is to use the mean value of those features in the place of those null entries. But we aim to use this data set and find out the stability of our model based on this test set from Kurmitola General Hospital. Henceforth we removed those three features from both the data set and finally worked on the rest 5 features which were:
a) number of times the patient got pregnant/Pregnancy
b) blood sugar or Glucose,
c) Blood pressure(BP),
d) Body mass-index(BMI) and
e) age of the patient.
Our target attribute was denoted as the outcome. The PIMA Indian diabetes dataset was divided on a ratio of 70:30 where 70% of the dataset was used for training purpose while rest 30% was used for test purpose. In addition, some of the values were too high or had a massive difference in the same featured attribute which could result in wrong prediction or may have dominating behavior in the prediction output. For this reason, normalization of the features was done using the formula 1

$$x_{normalized} = \frac{x - x_{minimum}}{(x_{maximum} - x_{minimum})} \tag{1}$$

Figure 1a 1b on the other hand, explains the relationship between attributes after and before normalization. Relationships are pretty noticeable in the data set. The red plots indicates the instances with outcome 1 or who has diabetes and green plots indicates who doesn't have diabetes.



(a) Scatter Matrix Withouth Normalization (b) Scatter Matrix After Normalization

Fig. 1: Scatter Matrix of Data set

3.2 Machine Learning Model Implementation

Scikit-learn [9], a python based machine learning library has been used to classify the data in our study. Various classification, regression and clustering algorithms including support vector machines, random forests, gradient boosting, k-means and DBSCAN are featured through Scikit-learn. It is designed to operate with the Python numerical and also with scientific libraries such as NumPy, Pandas, SciPy etc. We followed the process illustrated at FIG: 2 while implementing and evaluating the model. We used Decision tree, K-Nearest Neighbor and Naive Bayes to train the model and then compared the models to find best model. These algorithms are explained shortly in 3.3.1, 3.3.2 and 3.3.3. After completing the pre-processing stage, the dataset from PIMA was first divided into 70% training set and 30% test set. Then test set consisting of 538 instances was used to train the model. A 3-fold cross validation was performed on this training set to find the Hyper Parameters. In each cross-validation test, the model was trained with 358 instances and the tested with 179 instances. Once the hyper parameters were tuned, we fed data to our model for the mentioned algorithms with best hyper parameters. Tuning hyper parameter for each algorithm is explained at 3.3.1 & 3.3.2. After training stage, we first applied rest 30% data of Pima data set to the model for testing and recorded the result. Then we applied the Kurmitola General Hospital data set and recorded the result.

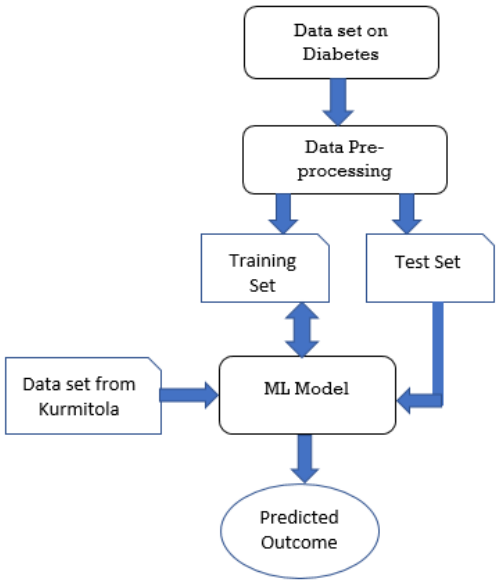


Fig. 2: Workflow

3.3 Algorithms

3.3.1 Decision Tree

Decision tree learning is a method for approximating discrete-valued target functions, in which the learned function is represented as sets of if-else/then rules to improve human readability. Decision Trees classify instances by sorting them down the tree from root node to some leaf node. Each node specifies a test of some attribute of the instance, and each branch descending from that node corresponds to one of the possible values for this attribute. The trees selects root node based on a statistical calculation called Information Gain. Information gain of a node can be measured through gini or entropy. We have used entropy to calculate the information gain. Entropy basically tells us how impure a collection of data is. The term impure here defines non-homogeneity. In other word we can say, "Entropy is the measurement of homogeneity. It returns us the information about a arbitrary dataset that, how impure/non-homogeneous the data set is." [11]

$$Entropy(S) = -(P_{\oplus} log_2 P_{\oplus} + P_{\ominus} log_2 P_{\ominus}) \tag{2}$$

$$Gain(S, A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v) \tag{3}$$

Maximum depth is considered as the hyper-parameter or the described as the length of the longest path from the root of a tree to its leaf. The tree will over fit on the training set while increasing the depth of the tree. If we set the maximum depth too high then it will increase the chance of over fitting the training data without capturing useful patterns as well as it will cause the testing error. This scenario is visualized in FIG: 3. So to fix those issues, we observed the cross-validation accuracy. And we found out the best accuracy of cross-validation is at maximum depth of 2. Further we have used Grid Search Algorithm to ensure the maximum depth for best cross-validation accuracy. Here we have found out the training accuracy of around 76% and the test accuracy to be 73% after a cross validation score of 72% for the model using Decision Tree Algorithm.

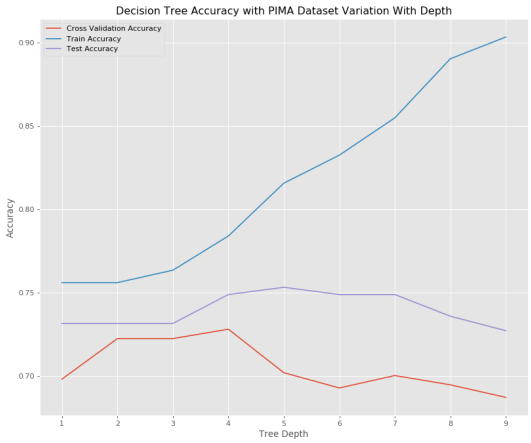


Fig. 3: Decision Tree Hyper Parameter Tuning

3.3.2 K-Nearest Neighbor

Next we have used the K-NN algorithm. K-NN is also known as Instance Based Learning or Lazy Learning algorithm. K-NN can be used for both classification and regression. But K-NN is suggested best for classification problem. The way K-NN works is, it calculates distance of one instance from "N" neighbors and classifies the new instances according to the class of most nearest neighbors. So here, the number of neighbors from which we want to calculate distance is the hyper parameter. We need to choose a tuned value for K or number of neighbors so that our model does not over fit or under fit. For analysing this problem, we plotted test, train and 3 fold cross-validation accuracy on FIG:4. This test set is Pima test set.

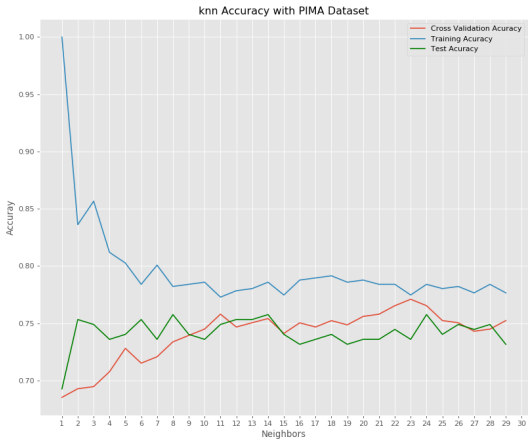


Fig. 4: KNN Hyper Parameter Tuning

The FIG:4 illustrates that error is Zero in training accuracy while k=1 but the model over fits on training set. On the other hand, while the value for "K" increases, accuracy reduces. We find out that both training and test accuracy does not meet at any specific point so that we can easily choose the neighbor for our K-NN. Best way to choose a value for K is by observing cross-validation accuracy. The red line on FIG: 4 indicates cross validation accuracy and it is clearly visible that accuracy is best at k=23. In order to confirm the value of K, we have undergone the GridSearch algorithm for hyperparameter tuning. In both cases, it is found out that the value of K being 23 gives the best fit. In case of K-NN, the training accuracy was 77% while the test accuracy was 74% and cross validation accuracy was 77% after prediction of diabetes using this model.

3.3.3 Naive Bayes

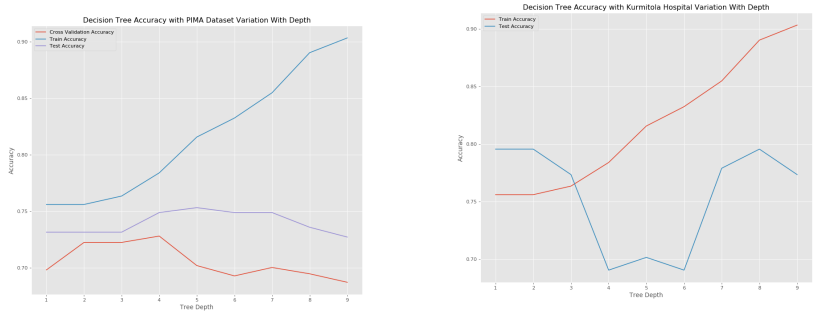
Lastly we performed the Naive Bayes Algorithm for the prediction of diabetes. There are three types of Naive Bayes algorithm: Gaussian Bayes, Multinomial Bayes and Bernoulli Bayes. After observing our data set, we found that Gaussian Naive Bayes is best fit for our data set. Because, All the attributes of our data set is continuous and normally distributed. So, Gaussian Naive Bayes in this case should fit best on our data set. Equation 4 is the formula of Gaussian Naive Bayes for mathematically computing probability of an event given another event.

$$p(x = v|c_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} e^{-(\frac{(v-\mu)^2}{2\sigma_k^2})} \tag{4}$$

After evaluating the algorithm, the training accuracy was at a rate of 76% while the test accuracy was at a rate of 73%

4 Results

We have found the accuracy on both our data sets by using Decision Tree Algorithm, K-NN algorithm and Naive Bayes Algorithm. Table: 12 shows us the list of resultant final accuracy we got after all the processes. We already discussed about the graph at FIG: 5a in Section 3.3.1, which is again mentioned here for showing the comparison of PIMA and KGH test set. FIG: 5 illustrated us the final accuracy graph of Pima test set and KGH test set with respect to tree depth. Here again it is proved that our choice of hyper parameter, max depth=2 was quite better.



(a) Decision Tree accuracy with Pima

(b) Decision Tree accuracy with Kurmitola Hospital

Fig. 5: Decision Tree Analysis

.	Yes	No
Yes	141	16
No	46	28

Table 3: Confusion Matrix for PIMA test set using Decision Tree

.	Yes	No
Yes	131	0
No	37	13

Table 4: Confusion Matrix for KGH test set using Decision Tree

Table: 3 and 4 are the confusion matrix for decision tree of both test sets. Though Pima test set predicts 16 false positive but KGH test set doesn't predict any single false positive.

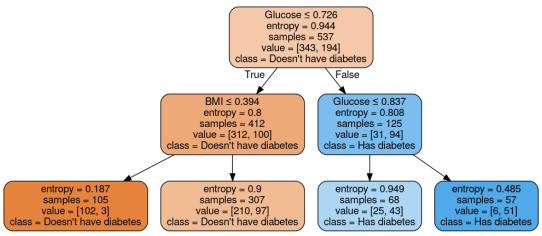
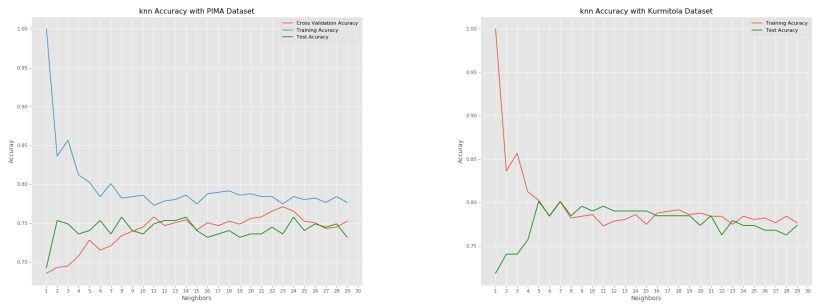


Fig. 6: Visualization of Decision Tree

Our decision tree of max depth = 2 is visualized at FIG: 6. Now, on the other hand, FIG:7 shows us the final accuracy analysis of K-Nearest Neighbor Algorithm with respect to the number of neighbors. Table: 5, 6 are the confusion matrix for Pima and KGH data set respectively. Again the confusion matrix at Table: 6 tells us that there was no false positive prediction.



(a) KNN accuracy with Pima (b) KNN accuracy with Kurmitola Hospital

Fig. 7: KNN Analysis

.	Yes	No
Yes	135	22
No	39	35

Table 5: Confusion Matrix for PIMA test set using KNN

.	Yes	No
Yes	131	0
No	40	10

Table 6: Confusion Matrix for KGH test set using KNN

	precision	recall	f1-score	support
0	0.77	1.00	0.87	131
1	1.00	0.220	0.33	50
Micro avg	0.78	0.78	0.78	181
Macro avg	0.88	0.60	0.60	181
Weighted avg	0.83	0.78	0.72	181

Table 7: Classification report for KNN on KGH set

We didn't face any problem of hyper parameter in Gaussian Naive Bayes algorithm. The algorithm is very good at measuring probability based on given events. Though we didn't have to worry about hyper parameter, but the confusion matrix of KGH test set at Table: 9 again says, our model didn't predict any false positives. Which is a good sign.

.	Yes	No
Yes	131	26
No	36	38

Table 8: Confusion Matrix for PIMA test set using Gaussian Naive Bayes

.	Yes	No
Yes	131	0
No	39	11

Table 9: Confusion Matrix for KGH test set using Gaussian Naive Bayes

	precision	recall	f1-score	support
0	0.77	1.00	0.87	131
1	1.00	0.220	0.33	50
Micro avg	0.78	0.78	0.78	181
Macro avg	0.89	0.61	0.62	181
Weighted avg	0.83	0.78	0.73	181

Table 10: Gaussian Naive Bayes classification report on KGH Dataset

	precision	recall	f1-score	support
0	0.78	0.83	o.81	157
1	0.59	0.51	0.55	74
Micro avg	0.73	0.73	0.73	231
Macro avg	0.69	0.67	0.68	231
Weighted avg	0.72	0.73	0.73	231

Table 11: Classification report for Gaussian Naive Bayes on Pima Testset

Classifier	Training Accuracy	Test Accuracy	KGH Accuracy
KNN	0.7746741154562383	0.7359307359307359	0.7790055248618785
Decision Tree	0.7560521415270018	0.7316017316017316	0.7955801104972375
Gaussian Naive Bayes	0.7690875232774674	0.7316017316017316	0.7845303867403315

Table 12: Accuracy rates for training, Pima test set and KGH set

5 Conclusion

Prevention is better than cure. And unfortunately, there's no curable treatment for Diabetes till now! However, taking preventive measures and having proper awareness can help reduce the risk of diabetes. In a developing country like Bangladesh, most of the people are unaware of the fact that they are having Diabetes. The goal of this paper is to analyze the data and predict whether a patient has diabetes or not so that they can take early preventive care measures. Because Diabetes is a slow killer and if not kept under strict control then it can lead to hyperglycemia that creates many severe complications than an increased level of blood glucose. Here by using up three classifiers throughout of implementing the machine learning method the best accuracy is found by using the decision tree algorithm which is of 80%, the minimum accuracy is found by using K-NN algorithm which is of 78%. The other method that we used provided us the accuracy of 79%. So, we can finalise by saying that decision tree provided us the best accuracy for our whole project

References

1. Krati Saxena, Dr. Zubair Khan and Shefali Singh,Diagnosis of Diabetes Mellitus using K Nearest Neighbor Algorithm, *International Journal of Computer Science Trends and Technology*,Vol:2 (2014).

2. Aakansha Rathore, Simran Chauhan and Sakshi Gujral, Detecting and Predicting Diabetes Using Supervised Learning: An Approach towards Better Healthcare for Women, *International Journal of Advanced Research in Computer Science* ,Vol:8 (2017).
3. Md. Aminul Islam and Nusrat Jahan, Prediction of Onset Diabetes using Machine Learning Techniques, *International Journal of Computer Applications* ,Vol:180 (2017).
4. Sakshi Gujral,Early Diabetes Detection using Machine Learning: A Review, *IJIRST –International Journal for Innovative Research in Science & Technology*, Volume 3, Issue 10, March 2017, ISSN (online): 2349-6010.
5. The Daily Star, 21 December, 2019, *A worrying picture of diabetes in Bangladesh*, https://www.diabetesatlas.org/en/sections/worldwide-toll-of-diabetes.html
6. IDF Diabetes Atlas, 9th edition, 2019, *International Diabetes Federation*, https://www.diabetesatlas.org/en/sections/worldwide-toll-of-diabetes.html
7. IDF Diabetes Atlas, 9th edition, 2019, *What is Diabetes*, https://www.idf.org/aboutdiabetes/what-is-diabetes.html
8. Towards Science, *PCA: Eigenvectors and Eigenvalues*, https://towardsdatascience.com/pca-eigenvectors-and-eigenvalues-1f968bc6777a
9. Wikipedia, *Scikit-learn*, https://en.wikipedia.org/wiki/Scikit-learn
10. *World Health Organization*, https://www.who.int/news-room/fact-sheets/detail/diabetes
11. Tom M. Mitchell, *"Machine Learning"*, McGraw-Hill Science/Engineering/Math, March 1, 1997.