DIABETES PREDICTION USING MACHINE LEARNING

[1]Atadjanova Nozima Sultan-Muratovna, [2]Abdukhakimov Fayzulla Kudratulla ugli

[1]Senior Lecturer at Tashkent University of Information Technologies “Computer engineering” faculty, Department of Computer Systems

[2]4th year BSc student at Tashkent University of Information Technologies “Computer engineering” faculty, Department of Computer Systems

https://doi.org/10.5281/zenodo.7612356

Abstract. *Diabetes is a chronic disease which occurs when the level of glucose rises above a certain amount. In other words, this is the case when the pancreas stops producing the necessary amount of insulin which controls the level of blood glucose. According to International Diabetes Federation, almost 382 million people are living with diabetes across the whole world. By 2035, the number of people with diabetes is forecast to increase up to 592 million. Diabetes is considered to be the major cause of blindness, stroke, kidney failure and many other fatal illnesses. When we consume food, our blood turns the food into sugar, or glucose. At that point, our pancreas normally releases a hormone called insulin. Insulin allows the glucose from a person’s food to access the cells in their body to supply energy. However, when a person has diabetes, this system does not work. It is generally known that Type 1 and Type 2 are the most common forms of diabetes among the elderly as well as adults, but there are also other forms such as gestational diabetes which occurs during pregnancy, and others.*

Machine learning is becoming a leading scientific filed in data science dealing with the ways in which machines themselves learn from experience and gradually develop. The purpose of this project is to develop a system which enables to diagnose diabetes in patients at early stage so that doctors can prevent serious consequences effectively. There are various algorithms such as K nearest neighbor, Logistic Regression, Decision Tree, Super Vector Machine and others to predict illness with high accuracy. In this project, we have decided to use Super Vector Machine algorithm in order to predict diabetes based on several symptoms in patients, and we have used a dataset which involves the symptoms of diabetes.

Keywords: *diabetes, machine learning, super vector machine, accuracy.*

I. INTRODUCTION

In today’s modern world, diabetes is the fast-growing disease among the elderly, and even youngsters are facing to this. In order to find a solution to treat the disease, we need to analyze it deeply and understand how it develops, what makes the process critical, and what we need to do so that we can prevent it. Sugar, or glucose, comes from food we consume on a daily basis. More specifically, these foods are called carbohydrate foods. Carbohydrate foods provide our body with its main energy source, even those people with diabetes needs carbohydrate to produce enough energy. Carbohydrate foods include bread, cereal, pasta, rice, fruit, dairy products and vegetables (particularly starchy vegetables). When we eat these foods, our blood breaks the down to glucose. Afterwards, the ‘sugar’ moves around the body with the help of blood circulation, some of which is taken to our brain to maintain cognitive development and several critical functions to live. The remainder of the glucose is taken to the cells of our body so as to produce energy and also to our liver, where it is stored as energy that is used later by the body. In order for the body to use glucose for energy, insulin is required. Insulin is a hormone that is produced by the cells in the pancreas.

We can imagine the process of how insulin works like a key to the door. Insulin attaches itself to doors on the cell, opening the door to allow glucose to move from the blood stream, through the door, and into the cell. If the pancreas is not able to produce enough insulin (insulin deficiency) or if the body cannot use the insulin it produces (insulin resistance), glucose builds up in the bloodstream (hyperglycemia) and diabetes develops.

Types of Diabetes

Type 1 diabetes is a serious condition in which insulin is not produced in a body. This means that the level of sugar(glucose) is considerably high in bloodstream. This happens because the body attacks the cells in pancreas that produce the insulin, meaning that the body stops producing at all. When a person has Type 1 diabetes, his or her body still breaks down the carbohydrate from food and drink and turns it into glucose. But when the glucose enters his/her bloodstream, there is no insulin to allow it into the body's cells. More and more glucose then builds up in the bloodstream, leading to high blood sugar levels. There are no exact causes of developing Type 1 diabetes, and little research has been done so far.

Type 2 diabetes develops when the pancreas makes less insulin than the body needs, and the body cells stop responding to insulin. They do not take in sugar as they should. Sugar builds up in your blood. When cells do not respond to insulin, this is called insulin resistance. It is usually caused by lifestyle factors, including obesity and a lack of exercise, or genetics that prevent cells from working normally. This type of diabetes is considered to be the most common among older adults, though it is becoming more common in children.

Gestational diabetes appears in pregnant women who suddenly develop high blood sugar. In two thirds of the cases, it will reappear during subsequent pregnancies. There is a great chance that type 1 or type 2 diabetes will occur after a pregnancy affected by gestational diabetes.

A list of some symptoms of diabetes is provided below:

- Increased thirst
- Loss of weight
- Sleepiness
- Confusion and difficulty concentrating
- Hunger
- Dry skin
- Having slow-healing sores
- Blurred vision
- Having itchy genitals or thrush that keeps coming back

Causes of diabetes

The cause of Type 1 diabetes is still unknown. Scientists are aware of how Type 1 diabetes causes the body to attack the cells that make insulin. As a result, the organism cannot produce insulin to break glucose down in the bloodstream. The cause of Type 2 diabetes is linked with a problem in the amount of insulin being made or used. The body cannot supervise its blood sugar levels, so they keep rising gradually. Some of reasons are:

- being physically inactive
- having excess weight or obesity
- having a family history of type 2 diabetes
- having pre-diabetes – when blood sugar levels are above the normal range but not high enough to be diagnosed as diabetes.

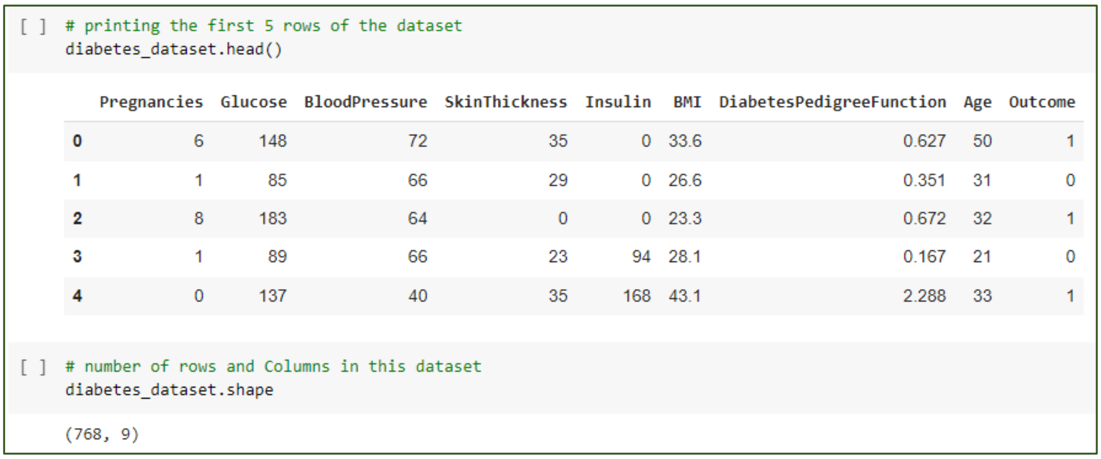
II. METHODOLOGY

In this section, we shall introduce our own methodology so as to increase the accuracy of our proposed algorithm used to predict diabetes based on several symptoms. The dataset we used in our project is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage. The dataset consists of several medical predictor (independent) variables and one target (dependent) variable, **Outcome**. Independent variables include the number of pregnancies the patient has had, their BMI, insulin level, age, and so on.

The dataset consists of 769 cases (Table 1). The aim is to predict whether a patient is diabetic or not based on the measures.

Table 1.

The overview of the dataset.



```
[ ] # printing the first 5 rows of the dataset
    diabetes_dataset.head()
```

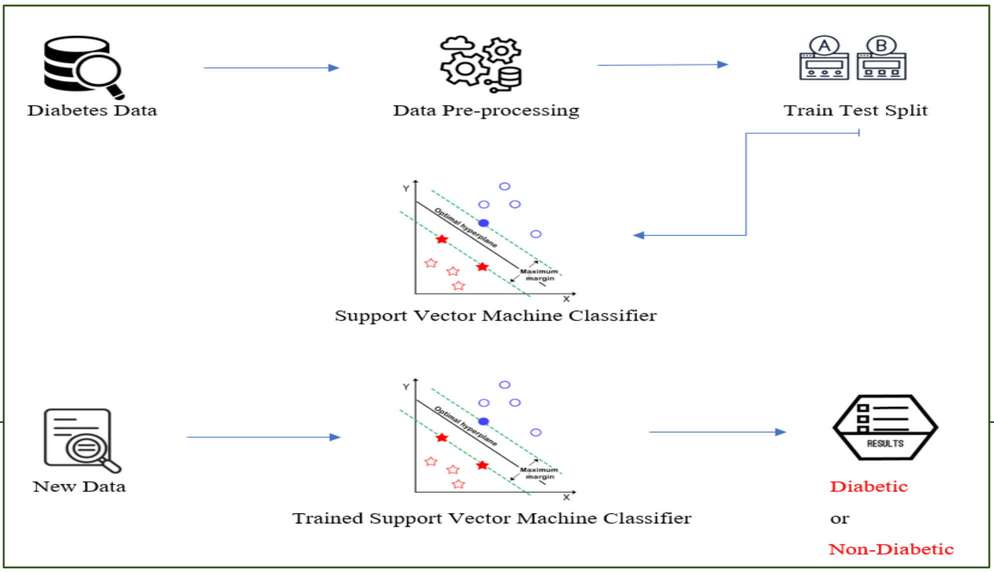
	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

```
[ ] # number of rows and Columns in this dataset
    diabetes_dataset.shape

    (768, 9)
```

We have proposed our own model that involves several steps to implement to obtain the necessary result in the end. As you can see the diagram (Figure 1), our model consists of three major stages beginning with collecting data about diabetes and ending with preparing Support Vector Machine Classifier for prediction. After implementing all steps, we shall be able to test our model through inserting new data into our project. Our model will automatically release the result based on inserted data attributes in a relatively short time as we have used Python programming language to create our project.

Figure 1.

Proposed Model Diagram.

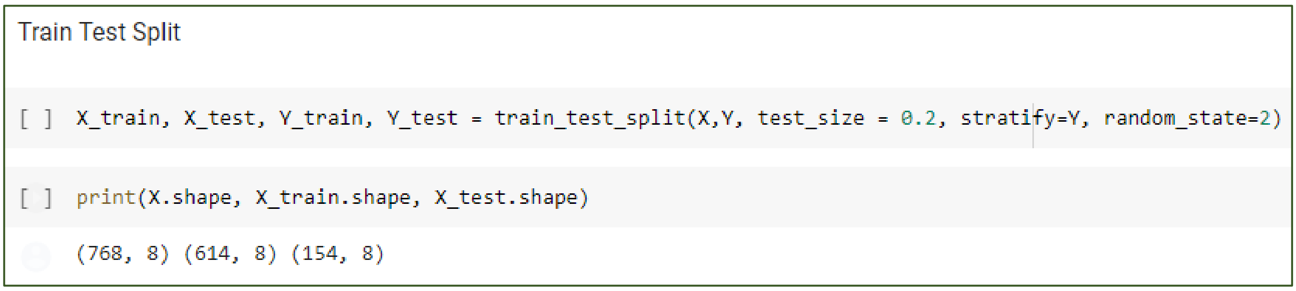


In Step 1, we use Diabetes Data for Pre-processing, meaning that we analyze the provided data based on features such as amount of glucose, skin thickness, age BMI and others using Python libraries. Through this way, we aim to prepare the data for another data processing procedure. Real-world data is not well-organized, and is frequently created, processed and stored by a variety of humans, business processes and applications. As a result, a dataset may be missing individual fields, contain manual input errors, or have duplicate data or different names to describe the same thing. Humans can often identify and rectify these problems in the data they use in the line of business, however, data used to train machine learning or deep learning algorithms needs to be automatically preprocessed. Therefore, we propose starting the process from pre-processing stage.

In Step 2, we split the processed data into two separate groups: test and train groups. We have divided these groups in 20:80 ratio so that we can train more data to obtain a satisfactory result with good accuracy. This ensures that both sets are representative of the entire dataset, and gives us a good way to measure the accuracy of our model (Figure 2). A train test is the way of structuring our machine learning project so that we can test our hypothesis quickly and inexpensively. Basically, it is a way to divide the training data so that we can try our algorithm accurately. A training set is normally used in order to train our existing data, and a testing set is responsible for checking the provided data. In order to gain high accuracy in machine learning, it is recommended to train more data than testing, therefore, we have agreed to split the data in 20:80 ratio (80% for the training set, and 20% for the testing set).

Figure 2.

Train Test Split.



Train Test Split

```
[ ]  X_train, X_test, Y_train, Y_test = train_test_split(X,Y, test_size = 0.2, stratify=Y, random_state=2)

[ ]  print(X.shape, X_train.shape, X_test.shape)

     (768, 8) (614, 8) (154, 8)
```

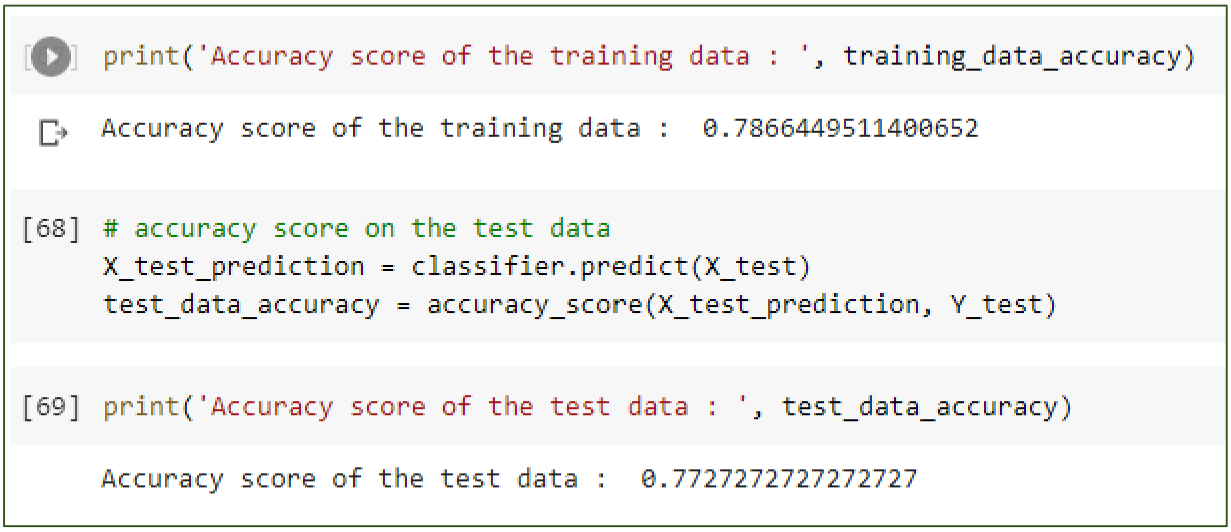
In Step 3, we implement the process of evaluating Support Vector Machine algorithm so as to finalize our model creation. We identify the maximum width between support vectors, as this will increase the efficiency of our project at the end of the evaluation of the process. As a result, we will be able to form Classifier which will enable us to predict diabetes.

III. RESULTS AND DISCUSSION

After evaluating our model, we have calculated our model accuracy so that we can predict whether a patient has diabetes or not based our SVM algorithm. According the results, our training group showed about 79 per cent accuracy, which means it is fairly acceptable to diagnose a patient. As for test group, it showed almost 77 per cent accuracy, which is also acceptable. In Figure 3, we have showed our implementation of calculating accuracy of the model:

Figure 3.

Model Accuracy.

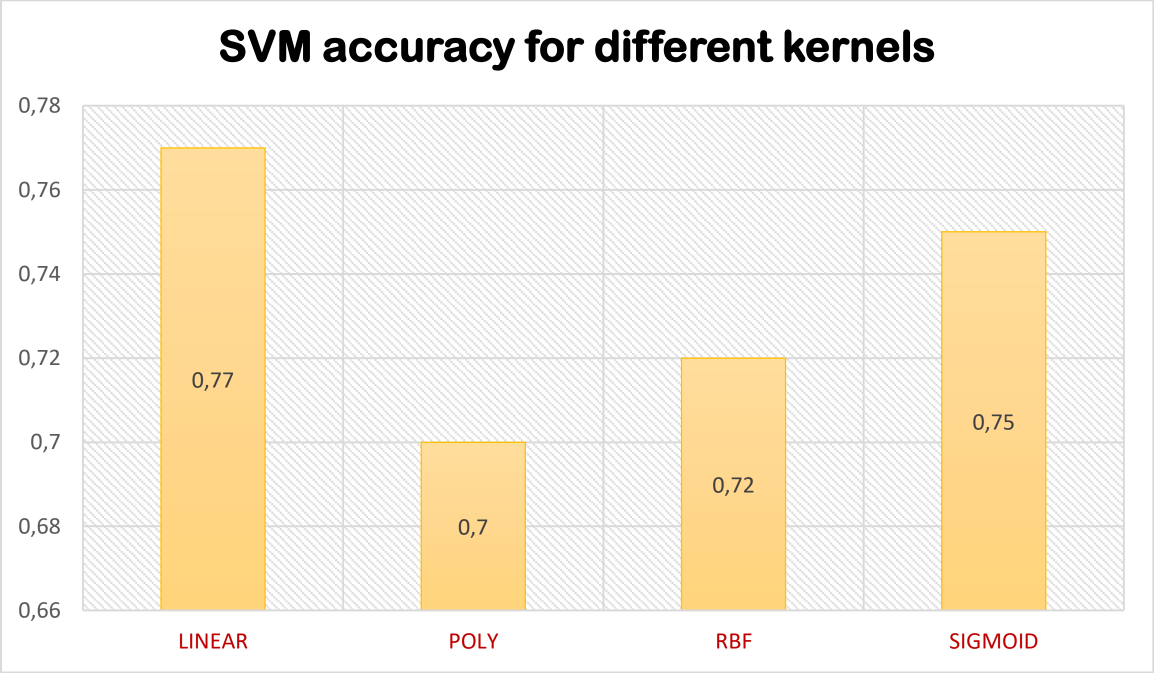


```
     print('Accuracy score of the training data : ', training_data_accuracy)

     Accuracy score of the training data :  0.7866449511400652

[68] # accuracy score on the test data
     X_test_prediction = classifier.predict(X_test)
     test_data_accuracy = accuracy_score(X_test_prediction, Y_test)

[69] print('Accuracy score of the test data : ', test_data_accuracy)

     Accuracy score of the test data :  0.7727272727272727
```

This SVM classifier aims at forming a hyper plane that can separate the classes as much as possible by adjusting the distance between the data points and the hyper plane. In order to find the most appropriate accuracy for prediction, we have tried four different kernel values so as to increase the accuracy of the model. These kernels are linear, poly, rbf and sigmoid.

Figure 4.

SVM accuracy for different kernels.



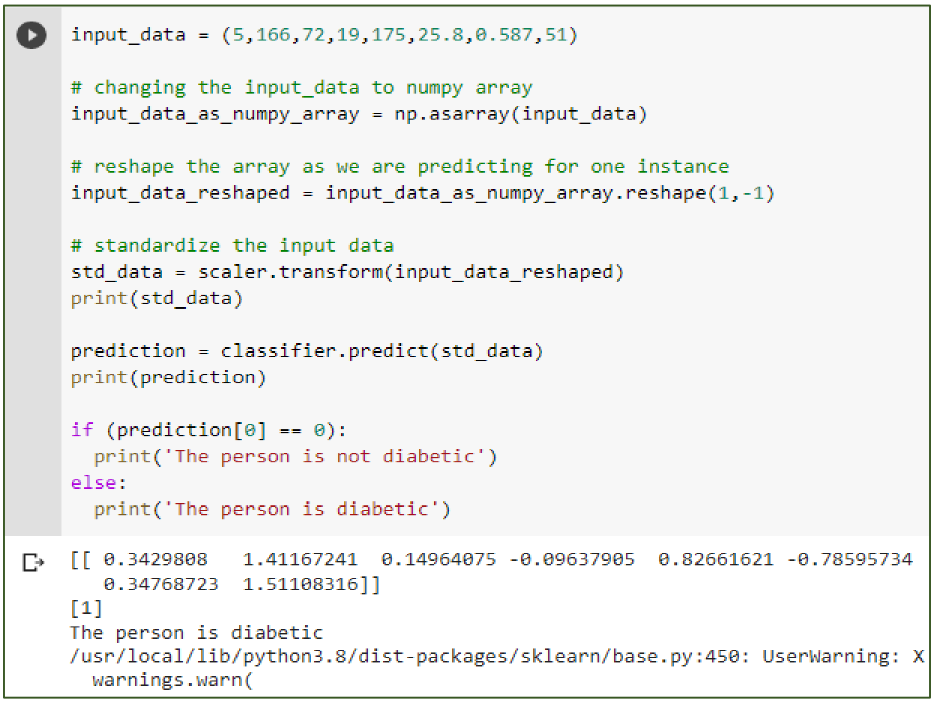
As can be seen from the histogram, the linear kernel performed the best for the dataset we used and achieved an overall score of 77 per cent. The other kernels such as poly, rbf and sigmoid showed slightly low results – 0.7, 0.72 and 0.75 per cents consecutively. Therefore, we have concluded that using the linear kernel is the most appropriate option to predict the disease using SVM algorithm. Importantly, this has helped us diagnose diabetes with minor inaccuracies, which is probably better than other algorithms.

Making a predictive system is considered to be the main part of our project, as all implementations have been done to obtain a necessary result. We need to understand the exact purpose of building a predictive system. Predictive analytics encompasses a variety of statistical techniques (including machine learning, predictive modelling and data mining) and uses statistics (both historical and current) to estimate, or ‘predict’, future outcomes. These outcomes might be behaviors a customer is likely to exhibit or possible changes in the market, for example. Predictive analytics help us to understand possible future occurrences by analyzing the past. In medicine, we can effectively use machine learning techniques to treat patients more efficiently than traditional methods. Machine learning models that are trained on a larger scale can show high performance in improving patients’ health in hospitals, and this will contribute to the improvement in medicine of a country. Furthermore, hospitals can use predictive systems to provide the best care by pre-determining increase of hospital bed availability or staff shortage.

As is shown below (Figure 5), we have written several lines of codes, and we have tested the project in the end.

Figure 5.

Predictive system



```
input_data = (5,166,72,19,175,25.8,0.587,51)

# changing the input_data to numpy array
input_data_as_numpy_array = np.asarray(input_data)

# reshape the array as we are predicting for one instance
input_data_reshaped = input_data_as_numpy_array.reshape(1,-1)

# standardize the input data
std_data = scaler.transform(input_data_reshaped)
print(std_data)

prediction = classifier.predict(std_data)
print(prediction)

if (prediction[0] == 0):
  print('The person is not diabetic')
else:
  print('The person is diabetic')
```

```
[[ 0.3429808   1.41167241  0.14964075 -0.09637905  0.82661621 -0.78595734
   0.34768723  1.51108316]]
[1]
The person is diabetic
/usr/local/lib/python3.8/dist-packages/sklearn/base.py:450: UserWarning: X
  warnings.warn(
```

It is clear from Figure 5 that we have inserted input_data variable into our predictive system and filled it with random numbers. Then, we changed, reshaped and standardized the data inserted. Using the classifier we created, we have tried to reduce the lines of codes to make it comfortable for other readers to understand. The result is that a person is diabetic according to the data we provided.

IV. CONCLUSION AND FUTURE WORK

One of the important real-world medical problems is the detection of diabetes at its early stage. In this project, analytical efforts have been made in designing a system which enhances the probability of predicting diabetes more accurately. During this work, SVM algorithm has been implemented using Python programming language. All the experiments have been done using PIMA Diabetes Dataset which includes 768 samples. The accuracy of our predictive system is 77%, and this is enough to diagnose a patient.

In future, this project can be utilized in local hospitals in Uzbekistan so that doctors can diagnose their patients using machine learning–based system, and this will ensure patients to live longer and happier than ever before. Currently, the hospitals in Uzbekistan is not using machine learning to effectively prevent diabetes, as they usually use conventional methods to diagnose and treat illnesses. We would like the government to help us introduce the new technique we created to local hospitals, as this would contribute to the significant decrease in the number of people with diabetes in Uzbekistan. Early action can reduce the probability of getting diabetes among adults as well as youngsters.

REFERENCES

1. KM Jyoti Rani, "Diabetes Prediction Using Machine Learning", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN: 2456-3307, Volume 6 Issue 4, pp. 294-305, July-August 2020.
2. K. Papatheodorou, M. Banach, M. Edmonds, N. Papanas, D. Papazoglou, Complications of diabetes,]. Diabetes Res. 2015 (2015) 1-5.
3. S. Park, D. Choi, M. Kim, W. Cha, C. Kim, I.C. Moon, *Identifying prescription patterns with a topic model of diseases and medications*, J. Biomed. Informat. 75 (2017) 35–47.
4. Chaki J, Ganesh ST, Cidham SK, Theertan SA. *Machine learning and artificial intelligence based diabetes mellitus detection and self-management: a systematic review*. J King Saud Univ Comput Inf Sci. 2020.
5. AD Association. Classification and diagnosis of diabetes: standards of medical care in diabetes-2020. Diabetes Care. 2019.
6. Muhammad LJ, Algehyne EA, Usman SS. Predictive supervised machine learning models for diabetes mellitus. SN Comput Sci. 2020;1(5):1–10.
7. Bernardini M, Romeo L, Misericordia P, Frontoni E. Discovering the type 2 diabetes in electronic health records using the sparse balanced support vector machine. IEEE J Biomed Health Inform. 2020;24(1):235–46.