

TACKLING DIABETES USING MACHINE LEARNING

Shaunak Mulay
NBN Sinhagad School of Engineering, Pune

Amrut Kulkarni
Vishwakarma Institute of Information Technology, Pune

Abstract—Diabetes is a chronic disease that has affected a huge mass of people worldwide. Diabetes mellitus or simply Diabetes is caused due to increase in the level of blood glucose. Weight, inactivity, family history, Race, age, gestational diabetes, high blood pressure, abnormal cholesterol, etc. are some of the causes of diabetes among most of the people. Diabetes is one of the most fatal and common disease. Machine Learning is transforming all spheres of our life, including the healthcare sector. Application of Machine Learning has a potential to vastly enhance the diabetes care methodology and also improve its efficiency. In case of diabetes, Machine Learning plays a crucial role in the diagnosis process. Using Machine Learning algorithms for diagnosis of diabetes can give quick and accurate results. Though the accuracy as of now is not that good we can try to improve on it. In this study, we have taken PIMA Indian dataset that is from National Institute of Diabetes and Digestive and Kidney Diseases. We have implemented algorithms like Logistic Regression, Random Forest and KNN on this dataset.

I. INTRODUCTION

Diabetes is the state of our body in which our body cells are unable to absorb sugar or glucose and utilize them in the form of energy. Extra sugar builds up in your system as a result of this. Insulin, a hormone produced by the pancreas, aids glucose absorption into cells for use as energy. In the United States, 34.2 million people of all ages – around one in ten – have diabetes. 7.3 million persons aged 18 and up (about 1 in 5) have no idea they have diabetes (just under 3 percent of all U.S. adults). The number of people who are diagnosed with diabetes increases with age. Diabetes affects more than 26% of persons aged 65 and over (about 1 in 4). Diabetes affected 8.5 percent of persons aged 18 and above in 2014. Diabetes was the direct cause of 1.5 million fatalities in 2019, with 48 percent of all diabetes-related deaths occurring before the age of 70.

The following are the most prevalent kinds of diabetes:

- Type 1 diabetes is an autoimmune illness, which means your body is attacking itself. The insulin-producing cells in your pancreas are damaged in this situation. Type 1 diabetes affects up to 10% of patients with diabetes.

- Type 2 diabetes occurs when your body either does not produce enough insulin or when your cells do not respond to insulin properly. Diabetes mellitus is the most frequent form of the disease. Type 2 diabetes affects up to 95% of diabetics.
- Prediabetes: This condition occurs before Type 2 diabetes develops. Your blood glucose levels are higher than normal, but not high enough for Type 2 diabetes to be diagnosed. normally disappears following the birth of a child. If you have gestational diabetes, though, you're more likely to develop Type 2 diabetes later in life.

Artificial intelligence techniques combined with cutting-edge technologies such as medical devices, mobile computing, and sensor technologies have the potential to improve the creation and delivery of chronic disease management services. Diabetes mellitus, which is defined by a malfunction of glucose homeostasis and can be identified using Machine Learning, is one of the most dangerous and prevalent chronic diseases.

II. LITERATURE SURVEY

The analysis of various healthcare datasets provides insights into various aspects of the industry. Various prediction models were developed and implemented using various techniques and methods.

Health facilities are an ideal platform to implement AI and ML trends to improve diagnostic systems.

Performance of these approaches is similar to that of healthcare professionals. They can help doctors diagnose and improve their patients' conditions by studying vast amounts of data.

Humar Kahramanli and Novruz Allahverdi used Artificial neural network (ANN) in combination with fuzzy logic to predict diabetes. [9] B.M. Patil, R.C. Joshi and Durga Toshniwal proposed Hybrid Prediction Model which includes Simple K-means clustering algorithm, followed by application of classification algorithm to the result obtained from clustering algorithm. In order to build classifiers C4.5 decision tree algorithm is used. El Jerjawi et al [2] established a neural network model for diabetes prediction. They used some attributes such as PG Concentration (Plasma glucose at 2 hours in an oral glucose tolerance test), Diastolic BP

(Diastolic Blood Pressure (mm Hg)). Most of them need some professional medical test, so it is not accessible for every person.

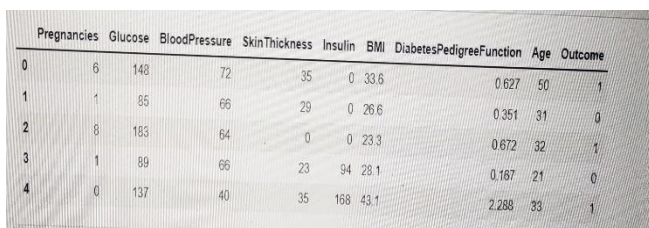
In [2], Aljumah, A.A., Ahamad, M.G., Siddiqui, M.K, proposed an approach data mining analysis on the Saudi Arabia NCD data using the Oracle Data Miner tool. The five age groups were re-classified into two age groups: Young and Old.

- In [1], Aishwarya, R., Gayathri, P., Jaisankar, N et al. proposed an approach Matlab 10 is used for implementation of SVM classification. Pre-processing and classification used bioinformatics tool available in Matlab. Bioinformatics tool was used to check prevalence of diabetes in patients' data is very important before preprocessing, using raw data.

III. DATA DESCRIPTION

The National Institute of Diabetes and Digestive and Kidney Diseases provided this data. The goal is to determine whether a patient has diabetes based on diagnostic parameters. All of the patients at this clinic are Pima Indian women who are at least 21 years old.

Each training instance has eight features as well as a class variable that serves as the training instance's label (see Table 1). Number of pregnancies, plasma glucose concentration, diabetes pedigree function, triceps skin fold thickness (mm), diastolic blood pressure (mmHg), 2-hour serum insulin (m U/mL), body mass index (kg/m²), and years of age are among the characteristics. The binary value is assigned to the class variable.



	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

Fig.1

(The above fig 1. shows the first 5 rows from whole dataset)

IV. PROPOSED WORK

To study the PIMA dataset we first performed Exploratory Data Analysis (EDA) on the dataset. We fetched a lot of insights about the correlation of different features among themselves. To achieve so, we used a variety of tools and strategies. We used techniques from Matplotlib, Python, Seaborn, and others.

This dataset didn't contain any null values so handling missing values work was avoided. We simply had to focus on getting insights and implementing different algorithms.

The figure 2. Shows the count-plot of the outcome of all the population. We can see that nearly 500 of them were tested negative and positive outcomes were around 275.

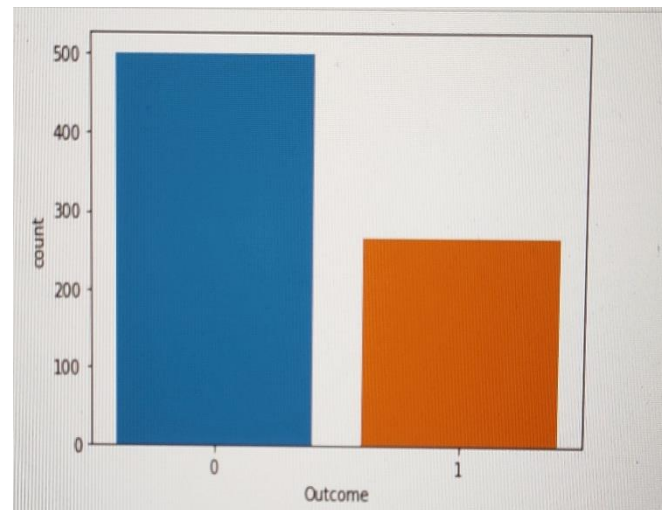


Fig.2

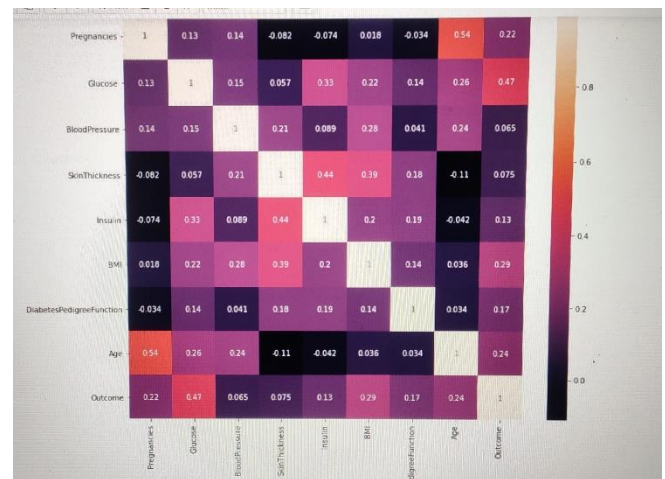


Fig.3

In this study we have implemented five algorithms on our dataset. SVM, Logistic Regression, Random Forest, Decision Tree and KNN are the five algorithms.

- **SVM:**

Support Vector Machine (SVM) is a supervised machine learning algorithm that can be used for both classification or regression challenges. Each data item is plotted as a point in n-dimensional space in the SVM algorithms, where n is the number of features you have and the value of each feature is the value of a given coordinate. Then by finding the hyperplane we perform classification that will easily differentiate between two different classes.

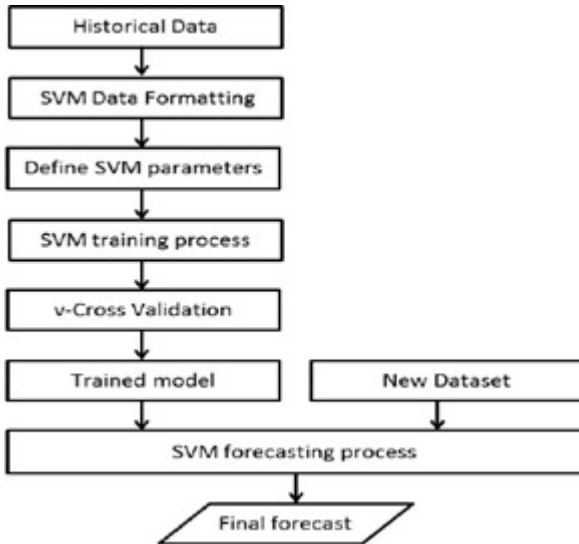


Fig.4

On fitting the train and test data we get the prediction results that are shown in the adjoining diagram. We got an accuracy of 83.4% on training data and 74.8% on test data. Here we can see a decrease of around 10% in the accuracy.

• **Logistic Regression:**

Logistic regression is a "supervised machine learning" approach that can be used to model the likelihood of a specific class or occurrence. It is used when the data is linearly separable and the outcome is binary or dichotomous in nature. This means that logistic regression models are models that have a certain fixed number of parameters that depend on the number they use a set of input features and produce a categorical prediction, such as whether a plant belongs to a specific species or not.

	Precision	Recall	F1-score	Support
0	0.78	0.88	0.83	147
1	0.73	0.56	0.64	84

Fig.5

Algorithm:

1. Randomly initialize parameters for the hypothesis function
2. Apply Logistic function to linear hypothesis function
3. Calculate the Partial Derivative
4. Update parameters
5. Repeat 2-4 for n number of iterations (Until cost function is minimized otherwise)

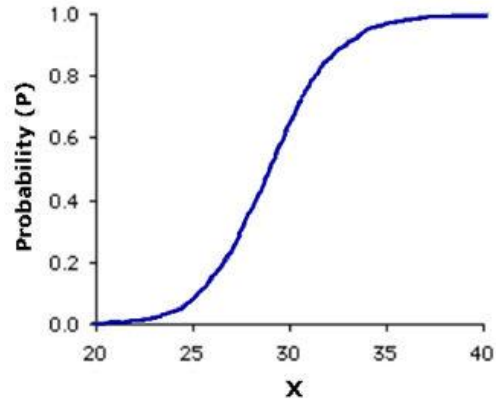


Fig.6

• **Random Forest:**

Random forest, like its name implies, consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model's prediction. Random forest adds additional randomness to the model, while growing the trees. Instead of searching for the most important feature while splitting a node, it searches for the best feature among a random subset of features. This results in a wide diversity that generally results in a better model.

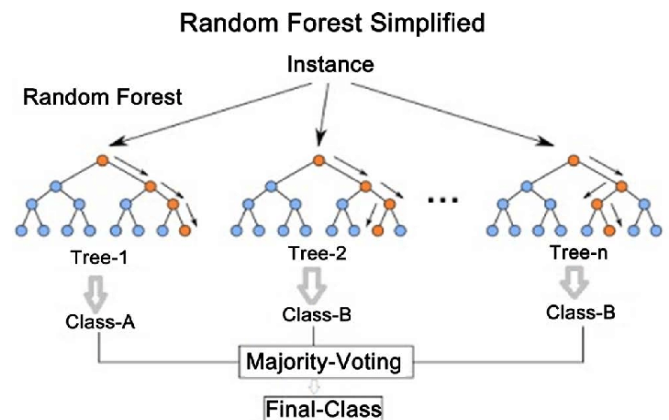


Fig.7

Algorithm:

- 1: In Random Forest n number of random records are taken from the data set having k number of records.
- 2: Individual decision trees are constructed for each sample.
- 3: Each decision tree will generate an output.
- 4: Final output is considered based on Majority Voting or Averaging for Classification and regression respectively.

	Precision	Recall	F1-score	Support
--	-----------	--------	----------	---------

0	0.77	0.89	0.82	147
1	0.73	0.52	0.61	84

Fig 8

Decision Tree:

Decision tree is the most powerful and popular tool for classification and prediction. A Decision tree is a flowchart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label. On each step or node of a decision tree, used for classification, we try to form a condition on the features to separate all the labels or classes contained in the dataset to the fullest purity.

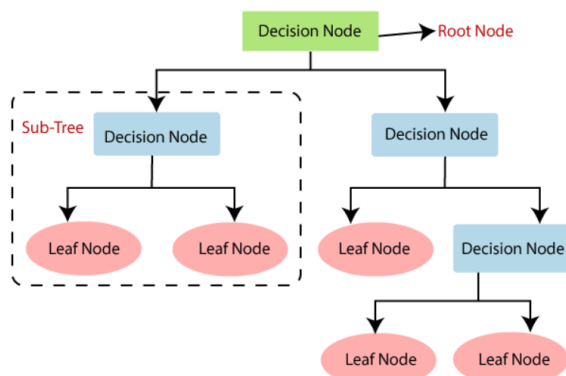


Fig.9

A general algorithm for a decision tree can be described as follows:

1. Pick the best attribute/feature. The best attribute is one which best splits or separates the data.
2. Ask the relevant question.
3. Follow the answer path.
4. Go to step 1 until you arrive to the answer.

	Precision	Recall	F1-score	Support
0	0.81	0.82	0.81	147
1	0.67	0.65	0.66	84

Fig.10

KNN:

The k-nearest neighbors (KNN) algorithm is a simple, easy-to-implement supervised machine learning algorithm that can be used to solve both classification and regression problems. The KNN algorithm assumes that similar things exist in close proximity. In other words, similar things are near to each other.

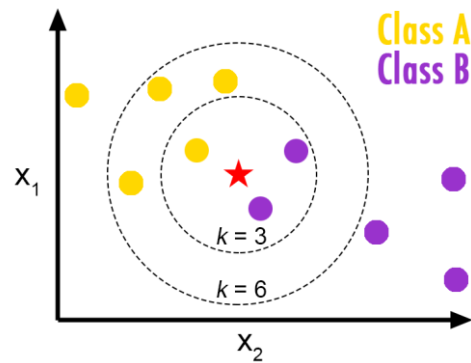


Fig 11

The following algorithm can be used to describe how K-NN works:

- 1: Choose the neighbor's number K.
- 2: Determine the Euclidean distance between K neighbors.
- 3: Using the obtained Euclidean distance, select the K closest neighbors.
- 4: Among these k neighbors, count the number of the data points in each category.
- 5: Assign the new data points to that category for which the number of the neighbor is maximum.
- 6: Our model is ready.

	Precision	Recall	F1-score	Support
0	0.76	0.90	0.83	147
1	0.75	0.51	0.61	84

Fig 12

V. CONCLUSION

In this study we implemented five different algorithms on the Pima dataset and we compared the train and test results of all the algorithms with each other. We observed that the accuracy should be improved in order to implement this model in real life. A good diabetes prediction model will aid doctors in making accurate diagnosis and ensuring that patients receive prompt treatment. For diabetes risk, we use descriptive statistics. Figure 13. lists the train and test results for different algorithms. For the first three models, logistic regression, SVM, and Decision trees are straightforward and intuitive, and they have fewer drawbacks.

	Models	Score
1	SVM	0.748918
2	Logistic Regression	0.766234
3	Random Forest	0.770563
4	Decision Tree	0.757576
5	KNN	0.761905

Fig.13

VI. REFERENCES

- [1] Design of a hybrid system for the diabetes and heart diseases - Humar Kahramanli , Novruz Allahverdi.
- [2] Application of data mining: Diabetes health care in young and old patients Abdullah A. Aljumah, Mohammed Gulam Ahamad, Mohammad Khubeb Siddiqui*College of Computer Engineering and Sciences, Salman bin Abdulaziz University, Saudi Arabia
- [3] A novel pattern extraction techniques used for classification of type-2 diabetic patients with back-propagation B. M. Patil, R. C. Joshi & Durga Toshniwal
- [4] A Method for Classification Using Machine Learning Technique for Diabetes Aishwarya. R 1 , Gayathri. P 2 and N. Jaisankar 3 M.Tech Student 1 , Assistant Professor (Senior) 2 and Professor 3 School of Computing Science and Engineering, VIT University, Vellore – 632014, Tamil Nadu, India.
- [5] Machine Learning in Predicting Diabetes in the Early Stage Juncheng Ma University of California Irvine Irvine, CA, 92697, United States
- [6] Prediction of Type 2 Diabetes Based on Machine Learning Algorithm Henock M. Deberneh and Intaek Kim *
- [7] el_Jerjawi, Nesreen Samer & Abu-Naser, Samy S. (2018). Diabetes Prediction Using Artificial Neural Network. International Journal of Advanced Science and Technology
- [8] Data mining a diabetic data warehouse Joseph L. Breault,a,b,*; Colin R. Goodall,c,d; Peter J. Fose, b, pg, Artificial Intelligence in Medicine.
- [9] Dr Saravana kumar N M, Eswari T, Sampath P and Lavanya S,” Predictive Methodology for Diabetic Data Analysis in Big Data”,
- [10] K. Rajesh and V. Sangeetha, “Application of Data Mining Methods and Techniques for Diabetes Diagnosis”
- [11] <https://www.who.int/health-topics/diabetes>
- [12] <https://www.analyticsvidhya.com/>
- [13] www.kaggle.com