International
Journal
Of Advanced Research in Engineering& Management (IJAREM)
ISSN: 2456-2033 || PP. 01-12

Diabetes Prediction using Machine Learning

Sujal Pose
Department of Computer Engineering
School of Engineering & Applied Sciences

Dr. Uttara Gogate
Department of Computer Engineering
SSJCOE, Dombivli

Abstract: Diabetes is one of the most serious illnesses and many people today suffer from it. Diabetes is caused by high blood sugar in the human body. At the same time, it is due to age, obesity, lifestyle, high blood pressure, etc. Hospitals typically perform various tests and treatments to collect the information needed to diagnose diabetes. Untreated diabetes can cause some major problems in people such as heart-related illnesses, blood pressure, kidney problems, eye damage, and in many forms it can affect the organs of the body. Diabetes can be controlled with early prediction and treatment. To achieve this goal, the project will apply a variety of machine learning techniques to develop a complete system for predicting diabetes in the human body or in patients.

Machine learning plays an important role in the medical industry. The healthcare industry has a large dataset. With the help of machine learning, you can study these datasets, gain knowledge from hidden patterns, data, and find information to predict outcomes accordingly. Machine learning is promising in the future, providing better predictive results by building models and applying algorithms directly to data collected from patients. This task uses machine learning techniques in the dataset to predict diabetes. Techniques include K-nearest neighbor method (KNN), support vector machine (SVM), logistic regression (LR), random forest (RF), decision tree classifier, and many more. By comparison, each model has different accuracy and can prove help worthy to predict accurately.

Keywords: diabetes, prediction, machine learning, healthcare, dataset

I. INTRODUCTION

The healthcare department has a large database. Such databases may contain structured, semi-structured, or unstructured data. Given the current scenario, diabetes (DM) is a very serious illness in developing countries like India. Diabetes (DM) is classified as a non-communicable disease (NCB) and many people suffer from it. Since the end of 2019, a newly identified disease called COVID-19 has spread rapidly to China and other countries. Severe Acute Respiratory Syndrome A new beta coronavirus, known as coronavirus 2 (SARS-CoV-2), has been identified as a COVID-19 pathogen, causing severe pneumonia and acute and even fatal lung failure. [1]

Diabetes is one of the leading causes of morbidity worldwide and is expected to increase significantly in the coming decades. [2] Several studies have shown that diabetic patients are more susceptible to infections such as Staphylococcus aureus and Mycobacterium tuberculosis [3]-[4]. Probably due to dysregulation of the immune system. [5] Plasma glucose levels and diabetes have been reported to be independent predictors of mortality and morbidity in SARS patients. [6] A retrospective study in Wuhan, China, found that of 41 COVID-19 patients, 32% had underlying illness and 20% had diabetes. [7] Therefore, these diabetics are at increased risk of COVID-19 and may have a poor prognosis.

174 COVIDs admitted to Wuhan Union Hospital from February 10, 2020 to February 29, 2020 to understand whether diabetes is a risk factor affecting the progression and prognosis of COVID-19- 19 patients were included in this study according to selection criteria. Their basic information, laboratory tests, computed tomography (CT) scans of the chest, and treatments were collected and analyzed. Diabetes has been found to be associated with a poor prognosis as a common underlying disorder in COVID-19 patients.

More than 425 million people worldwide have diabetes, and this number is predicted to increase to 629 million by 2045 [8]. Type 1 (T1D) known as insulin-dependent diabetes mellitus (IDDM). The reason behind this type of DM is that the human body cannot produce enough insulin, so it is necessary to inject insulin into the patient. Type 2 (T2D) is also known as non-insulin dependent diabetes mellitus (NIDDM). This type of diabetes is seen when somatic cells cannot use insulin properly. Type 3 gestational diabetes, elevated blood sugar levels in

pregnant women who do not detect diabetes early, causes this type of diabetes. Type 1 diabetes (T1D) is characterized by autoimmune-mediated destruction of insulin-producing β-cells, whereas type 2 diabetes (T2D) is associated with insulin resistance and impaired β-cell insulin secretion in long-term β-cells. It arises from the combination. Depletion and eventually destruction [9].Diabetes is the world's leading non-infectious chronic pandemic disease with complications. Over time, hyperglycemia can damage small and large blood vessels, increasing the risk of microvascular and large blood vessel complications [10]. A study of more than 1.3 million participants found that 98% of adults with type 2 diabetes had at least one co-existing chronic illness and nearly 90% had at least two chronic illnesses. Was shown [11,12]. The most common conditions in T2D patients are hypertension (82.1%), overweight / obesity (78.2%), hyperlipidemia (77.2%), chronic kidney disease (24.1%), and cardiovascular disease (21.6%) was included [11]. In a longitudinal study involving 915 individuals with T1D and 3,590 children in the reference cohort, the incidence of 6 chronic diseases was significantly higher in children with T1D [13].T1D is associated with increased risk of hospitalization (HR; 95% CI) due to comorbidity (3.7; 2.5 to 5.5), thyroid disease (14.2; 6.7 to 31.0), non-communicable enteritis and colitis (5.9; 3.0 to 11.5). It was related. , Cardiovascular disorders (3.1; 2.3 to 4.2), psychiatric disorders (2.0; 1.4 to 3.1), epilepsy (2.0; 1.1 to 3.7), and (obstructive) lung diseases (1.5; 1.2 to 2.0). Observational studies also show that cerebrovascular mortality is elevated at all ages in T1D patients [14].

Poor management of diabetes increases the risk of skin, bones, eyes, ears, gastrointestinal, urinary tract, respiratory infections, etc., and significantly increases hospitalization and mortality [15], [16] , [17], [18].]

DM has long-term complications. Also, diabetics are at increased risk of various health problems. A technique called predictive analytics incorporates a variety of machine learning algorithms, data mining techniques, and statistical techniques that use current and historical data to find knowledge and predict future events. By applying predictive analytics to medical data, you can make important decisions and make predictions. Predictive analytics can be performed using machine learning and regression techniques. Predictive analytics aims to diagnose disease with the highest possible accuracy, enhance patient care, optimize resources, and improve clinical outcomes. [19]Machine learning is considered to be one of the most important artificial intelligence features, supporting the development of computer systems with the ability to gain knowledge from past experience without the need for programming in all cases. increase. Machine learning is considered an urgent need for today's situation to eliminate human effort by supporting automation with minimal flaws. The existing method for detecting diabetes is to use clinical tests such as fasting blood glucose and oral glucose tolerance tests. However, this method is time consuming. This white paper focuses on building predictive models using machine learning algorithms and techniques for predicting diabetes.

The paper is organized as follows Section II-gives literature review of the work done on diabetes prediction earlier and taxonomy of machine learning algorithms. Section III-presents motivation behind working on this topic. Section IV gives diabetes prediction proposed model is discussed. Section V gives results of experiment followed by Conclusion and References.

II. RELATED WORK

Analysis of related work yielded results on different medical data sets, and analyzes and predictions were performed using different methods and techniques. Different predictive models have been developed and implemented by different researchers using variations of data mining techniques, machine learning algorithms, or combinations of these techniques.

Veena Vijayan V. And Anjali C discussed diabetes caused by increased sugar content in plasma. Various computerized information systems that utilize classifiers to predict and diagnose diabetes using decision trees, SVMs, naive bays and ANN algorithms have been outlined [20] Song et al. [21] Different algorithms are described using different parameters such as glucose, blood pressure (BP), skin thickness (ST), insulin, body mass index (BMI), diabetic pedigree function (DPF), age, etc. All parameters were not included. Only small sample data is used. ANN, EM, GMM, logistic regression, and SVM have been applied to diabetes datasets. ANN (Artificial Neural Network) provided better accuracy and performance than other algorithms. P. Suresh Kumar and V. Umatejaswi have announced algorithms such as Decision Tree, SVM, and Naive Bayes for identifying diabetes using data mining techniques. [22]

Xue-Hui Meng et al. [23] Predict diabetes using real-world datasets by distributing questioners and gathering information using a variety of data mining techniques. SPSS and weka tools were used for data analysis and forecasting, respectively. Tao et al. [24] KNN, Naive Bayes, Random Forest, Decision Trees, Swimming, and Logistic Regression were applied early on to predict diabetes (DM). Focus on filtering. Mani Butwall and

Shraddha Kumar (2015) proposed a model that uses a random forest classifier to predict diabetic behavior. [25] Nawaz Mohamudally 1 and Dost Muhammad (2011) predicted diabetes using the C4.5 decision tree algorithm, neural networks, K-means clustering algorithms, and visualizations. [26] Weifeng Xu et al. [27] Various machine learning algorithms have been applied to predict diabetes. These algorithms provided RF with greater accuracy than other data mining techniques. Swarupa et al. [28] .Naive Bayes (NB) provided excellent accuracy with an accuracy value of 77.01%. Sajida et al. [9] Adaboost provided better performance and accuracy. Loannis et al. [29] Ten-fold cross-validation was used as an evaluation method for three different algorithms: logistic regression, naive Bayes, and Svm. From these three different algorithms, svm provided higher accuracy and performance than other methods.

The outcome of the literature survey is different kind of methods and algorithms are used in the path to predict the diabetes with better accuracy, most of the time the same dataset is being used in the past to work. We would be learning from those survey and similarly implements the model but on two different datasets, one with medical parameters and other with symptoms and try to develop a interactive system that can be used by the user to know if he or she has diabetes or not.

III. MOTIVATION

Since 10 years, the proportion of people suffering from diabetes has increased dramatically. The current human lifestyle is the main reason behind the growth of diabetes. Current medical diagnostic methods can result in three different types of errors. 1. In reality, the patient is already diabetic, but the test results are a false-negative type that indicates that the person is not diabetic. 2. False positive type. In this type, the actual patient is not a diabetic, but according to test reports he / she is a diabetic. 3. The third type is an unclassifiable type in which the system cannot diagnose a particular case. This occurs due to inadequate extraction of knowledge from historical data and can be predicted for certain patients with unclassified types. However, in reality, patients should be expected to be in either the diabetic or non-diabetic category. Such misdiagnosis can lead to unnecessary treatment or no treatment at all when needed. To avoid or mitigate the severity of these effects, you need to create your system using machine learning algorithms and data mining techniques that provide accurate results and reduce manpower.

IV. PROPOSED SYSTEM

Based on the problems explained in the introductory part, we propose an interactive system that can predict the presence or absence of diabetes more accurately. This model employs various classifiers such as SVM, Random Forest, Logical Regression, Decision Tree, and KNN. The main focus is on the well-known Benchmark Diabetes Dataset from the PIMA Indian Diabetes Dataset in the UCI Machine Learning Repository, which consists of other datasets with 8 and 17 attributes for maximum accuracy with machine learning techniques. The framework consists of the following important phases, as shown in the figure below. The data from the dataset is first preprocessed to avoid null or redundant data that can affect the prediction, then split later, and then the features that make the prediction even more annoying are selected. Machine learning algorithms have been applied to the training and test sets, resulting in the desired accuracy. You can later implement those models and use them to develop the system.

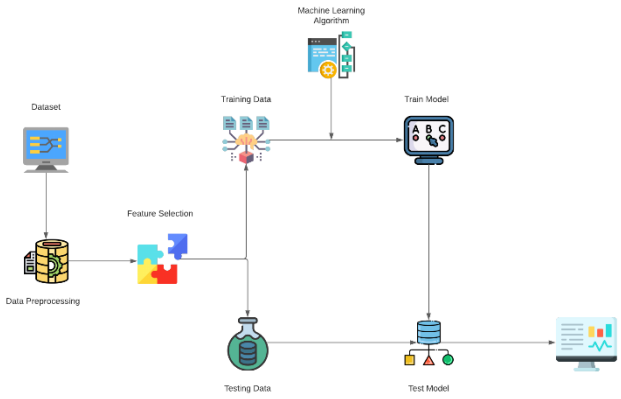


Fig 1. Process followed for evaluation of algorithms Diabetes Datasets

A. Datasets

This research paper uses publicly available datasets [30] [31] downloaded from the UCI Machine Learning Repository. The dataset selected is part of a larger dataset held by the National Institute of Diabetes, Gastroenterology and Kidney Disease. In predictive analytics, many researchers used this dataset for their studies. This dataset contains 768 patient records of Pima Indian women with nine attributes.

Another dataset was downloaded from the early stage diabetes risk prediction. It was collected using a direct questionnaire from patients at the Sylhet Diabetes Hospital in Sylhet, Bangladesh and approved by a physician. You can use this dataset for symptom-based predictions. This dataset contains 520 patient records with 17 attributes.

Table 1 describes the attributes used in the dataset, and Table 2 describes the basic data statistics. In particular, all patients in this dataset are women over the age of 21 who belong to the Pima Indian heritage. The purpose is to predict whether a person has diabetes based on the patient's diagnostic measurements.

The dataset consists of several medical prediction variables and one target variable, the result. Predictive variables include the patient's pregnancy count, body mass index, insulin level, age, and so on.

TABLE 1. PIMA INDIAN DIABETES DATASET ATTRIBUTES

Column	Information
Pregnancies	Number of times pregnant
Glucose	Plasma glucose concentration a 2 hours in an oral glucose tolerance test
BloodPressure	Diastolic blood pressure (mm Hg)
SkinThickness	Triceps skin fold thickness (mm)
Insulin	2-Hour serum insulin (mu U/ml)
BMI	Body mass index (weight in kg/(height in m)^2)
DiabetesPedigree Function	Diabetes pedigree function (Numeric)
Age	No. of years (Numeric)
Outcome	Class variable (0 or 1) 268 of 768 are 1, the others are 0

TABLE II. PIMA INDIAN DIABETES DATASET STATISTICS

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
count	768.00	768.00	768.00	768.00	768.00	768.00	768.00	768.00	768.00
mean	3.85	120.89	69.11	20.54	79.80	31.99	0.47	33.24	0.35
std	3.37	31.97	19.36	15.95	115.24	7.88	0.33	11.76	0.48
min	0.00	0.00	0.00	0.00	0.00	0.00	0.08	21.00	0.00
25%	1.00	99.00	62.00	0.00	0.00	27.30	0.24	24.00	0.00
50%	3.00	117.00	72.00	23.00	30.50	32.00	0.37	29.00	0.00
75%	6.00	140.25	80.00	32.00	127.25	36.60	0.63	41.00	1.00
max	17.00	199.00	122.00	99.00	846.00	67.10	2.42	81.00	1.00

Table 3 describes the attributes used in the dataset, and Table 4 describes the basic data statistics. In particular, all patients in this dataset are between the ages of 20 and 65. The purpose is to predict whether a person has diabetes based on the patient's symptoms.

This dataset contains data on signs and symptoms of new diabetes or potential diabetes. It was collected using a direct questionnaire from patients at the Sylhet Diabetes Hospital in Sylhet, Bangladesh and approved by a physician.

TABLE III. EARLY STAGE DIABETES RISK PREDICTION DATASET ATTRIBUTES

Column	Information
Age	Age in years ranging from (20years to 65 years)
Gender	Male / Female

Polyuria	**Polyuria** is a condition where the body urinates more than usual (Yes / No)
Polydipsia	**Polydipsia** is a medical name for the feeling of extreme thirstiness. (Yes / No)
Sudden weight loss	Losing weight frequently (Yes / No)
Weakness	Feeling weak physically (Yes / No)
Polyphagia	Medical term for excessive or extreme hunger. (Yes / No)
Genital Thrush	Affects the vagina, though may affect the penis too, and can be irritating and painful. (Yes / No)
Visual blurring	Disturbance in a person's eyesight (Yes / No)
Itching	irritating sensation that makes you want to scratch your skin (Yes / No)
Irritability	Irritability involves feelings of anger or frustration (Yes / No)
Delayed healing	Wound takes time to heal (Yes / No)
Partial Paresis	Weakening of a muscle or group of muscles Yes / No)
Muscle stiffness	Muscles feel tight (Yes / No)
Alopecia	**C**ondition that causes hair to fall out in small patches, which can be unnoticeable. (Yes / No)
Obesity	Complex disease involving an excessive amount of body fat (Yes / No)
Class	Diabetes outcome (Positive / Negative)

TABLE IV. EARLY STAGE DIABETES RISK PREDICTION DATASET STATISTICS

	Age	Polyuria	Polydipsia	sudden weight loss	weakness	Polyphagia	Genital thrush	visual blurring	Itching	Irritability	delayed healing	partial paresis	muscle stiffness	Alopecia	Obesity	class
count	520.00	520.00	520.00	520.00	520.00	520.00	520.00	520.00	520.00	520.00	520.00	520.00	520.00	520.00	520.00	520.00
mean	48.03	0.50	0.45	0.42	0.59	0.46	0.22	0.45	0.49	0.24	0.46	0.43	0.38	0.34	0.17	0.62
std	12.15	0.50	0.50	0.49	0.49	0.50	0.42	0.50	0.50	0.43	0.50	0.50	0.48	0.48	0.38	0.49
min	16.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
25%	39.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
50%	47.50	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
75%	57.00	1.00	1.00	1.00	1.00	1.00	0.00	1.00	1.00	0.00	1.00	1.00	1.00	1.00	0.00	1.00
max	90.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

B. Training Data and test Data

Machine learning training datasets are used to train models to perform rich actions. Detailed features are obtained from the training set to train the model. Therefore, these structures are combined with the prototype. Sentiment analysis retrieves a single word or a sequence of consecutive words from a tweet. Therefore, if the training set is properly labeled, the model can get something from the feature. Therefore, to test your model, use that type of data to see if your model is responding correctly.

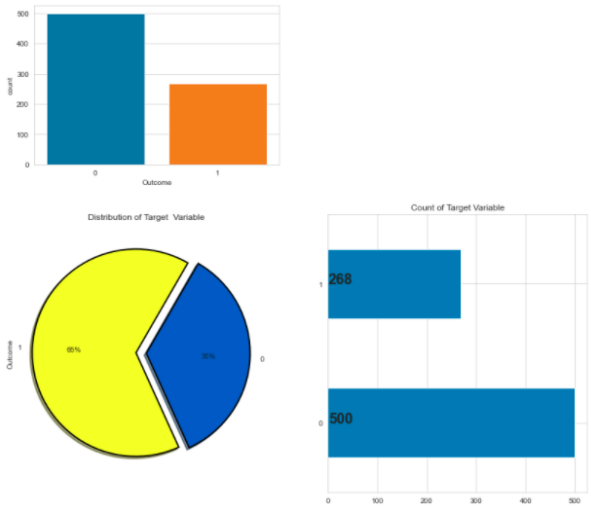


Fig 2. Distribution & count of target variable on PIMA Diabetes Dataset

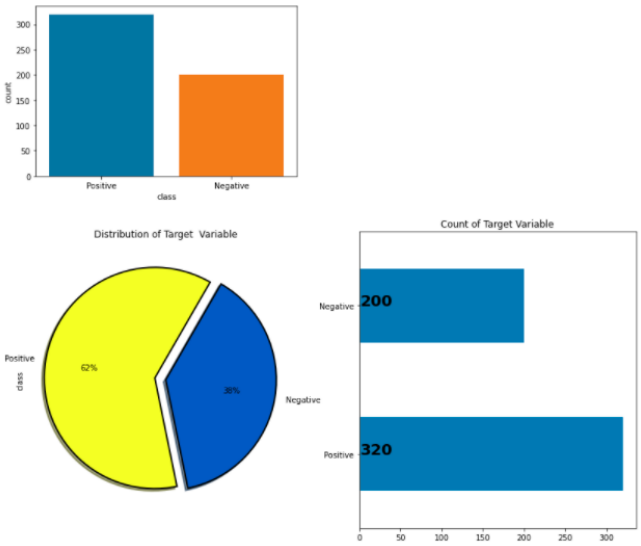
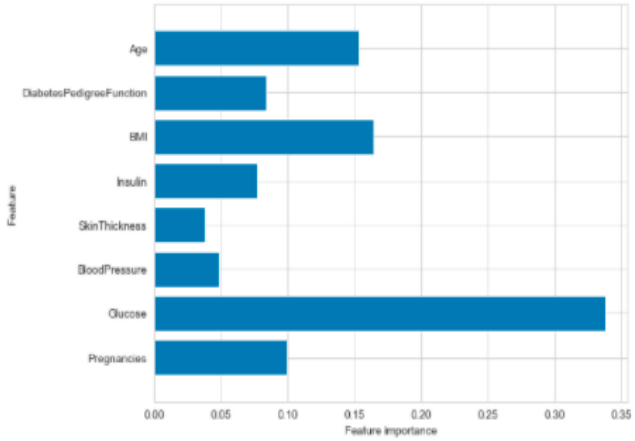


Fig 3. Distribution & count of target variable on Early Stage Diabetes Risk Prediction Dataset

C. Feature extraction

Feature extraction is used to transform input information as a result of features. The squared measure of an attribute is a property of the input design that facilitates the distinction between classes of the input design. If the input data is too large for the algorithm to process, it is suspected to be redundant because the image represented as a pixel will recur and change to a condensed set of attributes. You can use the extracted features instead of the complete initial data to perform the selected task. From the figure below, we can see that it is most important that attribute glucose is the leading cause of diabetes in particular.



D. Machine Learning Algorithms Used

- **Logistic Regression**

Logistic regression is a supervised learning classification algorithm used to predict the probabilities of target variables. The nature of the target or dependent variable is dichotomized. That is, only two classes are possible.

Simply put, the dependent variable is binary in nature and the data is encoded as either 1 (representing success / yes) or 0 (representing failure / no).

Mathematically, the logistic regression model predicts P (Y = 1) as a function of X. It is one of the simplest ML algorithms that can be used for various classification problems such as spam detection, diabetes prediction, and cancer detection.

- **Support Vector Machine**

Support Vector Machine (SVM) is a powerful yet flexible supervised machine learning algorithm used for both classification and regression. However, they are commonly used in classification problems. SVM was first introduced in the 1960s, but was subsequently improved in 1990. SVM has its own implementation method compared to other machine learning algorithms. These days it is very popular because it can handle multiple continuous and categorical variables.

- **Decision Tree**

In general, decision tree analysis is a predictive modeling tool that can be applied in many areas. Decision trees can be built with an algorithmic approach that allows you to divide your dataset in different ways based on different conditions. Decision stress is the most powerful algorithm in the category of supervised algorithms.

These can be used for both classification and regression tasks. The two main entities in the tree are the decision nodes where the data is split and separated to produce results.

- **Random Forest**

Random forest is a supervised learning algorithm used for both classification and regression. However, it is mainly used for classification issues. As we know, forests are made up of trees, and the more trees there are, the stronger the forest. Similarly, the Random Forest algorithm creates a decision tree in the data sample, gets predictions from each, and finally votes to select the best solution. This is a better ensemble technique than a single decision tree because it reduces overfitting by averaging the results.

- **K-nearest neighbors**

The K-nearest neighbor (KNN) algorithm is a type of supervised ML algorithm that can be used for both classification and regression prediction problems. However, it is mainly used for industry classification prediction problems. The following two properties define KNN properly-

Lazy Learning Algorithm-KNN is a lazy learning algorithm because there is no special training phase and all data is used for training during classification.

Nonparametric Learning Algorithm-KNN is also a nonparametric learning algorithm because it makes no assumptions about the underlying data.

V. RESULT AND DISCUSSION

Several machine learning algorithms were used in this experimental study. These algorithms are KNN, SVM, LR, DT, and RF. All of these algorithms were applied to the PIMA Indian dataset and the early diabetes risk prediction dataset. The data is divided into two parts, training data and test data, which consist of 70% and 30% of the data, respectively. All of these algorithms were applied to the same dataset using Jupyter Notebook and the results were obtained. Predicting accuracy is the main evaluation parameter used in this task. The accuracy can be ignored using Equation 1. Accuracy is the overall success rate of the algorithm.

All predicted true positives and true negatives divided by all positives and negatives. The true positives (TP), true negatives (TN), false negatives (FN), and false positives (FP) predicted by all algorithms are shown in Figures 3 and 4. In this case, TP means real diabetes and predicted diabetes. FN, real diabetes, but not predicted to be diabetic. I predicted FP and diabetes, but I'm not actually diabetic. TN, not really diabetic, prediction is not diabetic.

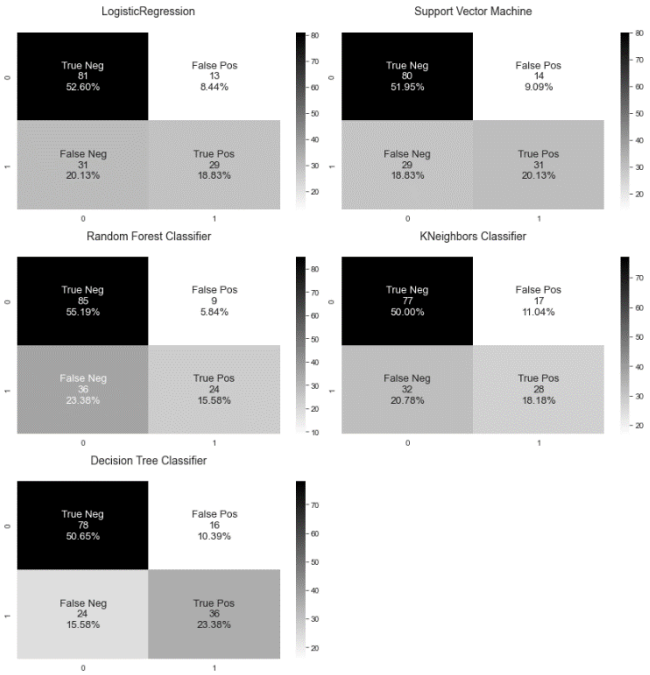


Fig 4. Confusion matrix for the labelled algorithms implemented on PIMA Diabetes Dataset

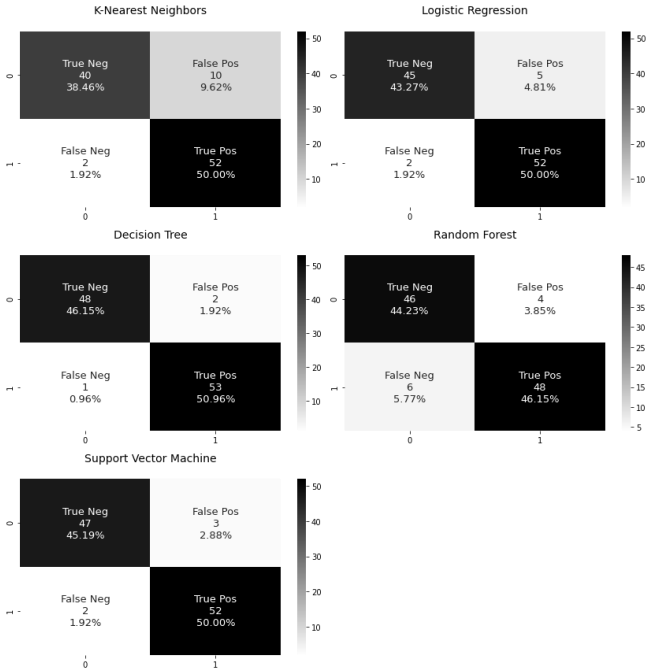


Fig 5. Confusion matrix for the labelled algorithms implemented on Early Stage Diabetes Risk Prediction Dataset

To summarize the results of the algorithms on both the dataset we have plotted a relational bar graph which demonstrates the algorithms with their respective accuracies.

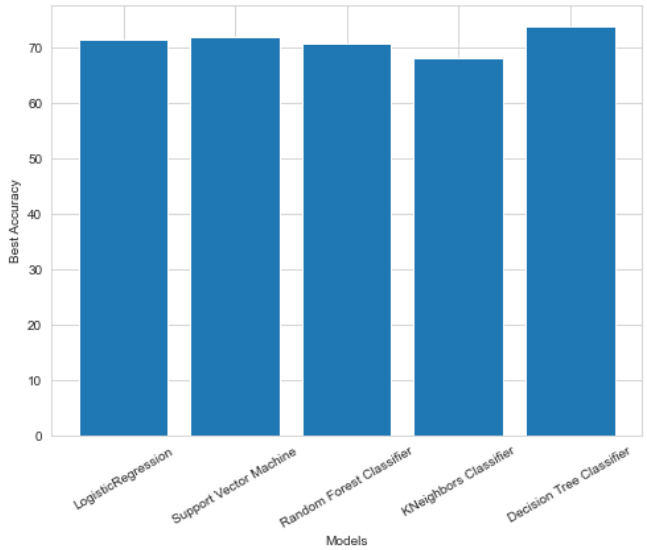


Fig 6. Accuracy Comparison for the algorithms on PIMA Diabetes Dataset

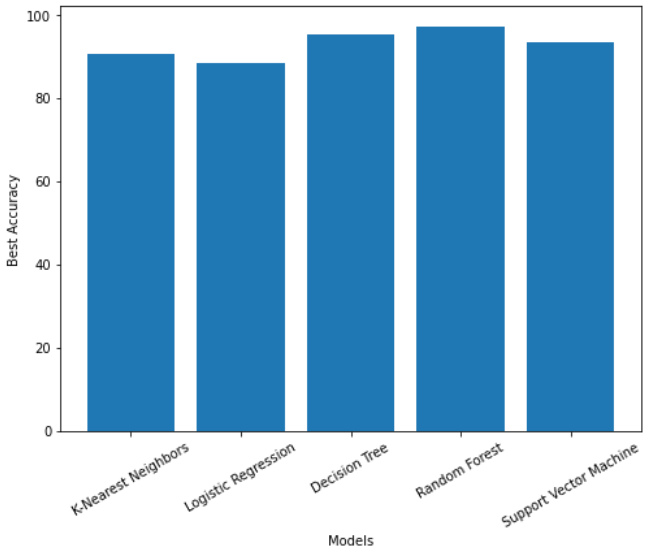


Fig 7. Accuracy Comparison for the algorithms on Early Stage Diabetes Risk Prediction Dataset

CONCLUSION

Predictive analytics in healthcare can change the way medical researchers and practitioners gain insights from medical data to make decisions. In this study, different machine learning algorithms were applied to both datasets, and classification was done using different algorithms, some of which had promising results implemented in the project. We have compared the accuracy of machine learning algorithms with two different datasets. The algorithms definitely works differently on both the datasets as there parameters or attributes are different although the outcome may be the same. Both the datasets make the algorithms learn differently, we can same one with the medical attributes and other with symptoms of the person. It is clear that this implementation gives us two different methodology to predict if a person has diabetes or not based on the dataset used.Both the ways are different of its kind but Early Risk Prediction Dataset gives better prediction with greater accuracy as compared to PIMA Diabetes Dataset. In addition, you can extend this work to find ways to use this model to predict and know if you have diabetes through the interface.

REFERENCES

[1] Chen N, Zhou M, Dong X, et al. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. Lancet. 2020;395(10223):507-513.
[2] Knapp S. Diabetes and infection: is there a link?—a mini-review. Gerontology. 2013;59(2):99-104.
[3] Shah BR, Hux JE. Quantifying the risk of infectious diseases for people with diabetes. Diabetes Care. 2003;26(2):510-513.
[4] Joshi N, Caputo GM, Weitekamp MR, Karchmer AW. Infections in patients with diabetes mellitus. N Engl J Med. 1999;341(25):1906-1912.
[5] Hodgson K, Morris J, Bridson T, Govan B, Rush C, Ketheesan N. Immunological mechanisms contributing to the double burden of diabetes and intracellular bacterial infections. Immunology. 2015;144(2):171-185.
[6] Yang JK, Feng Y, Yuan MY, et al. Plasma glucose levels and diabetes are independent predictors for mortality and morbidity in patients with SARS. Diabetic Med. 2006;23(6):623-628.
[7] Wang D, Hu B, Hu C, et al. Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus-infected pneumonia in Wuhan, China. JAMA. 2020;323:1061.
[8] International Diabetes Federation IDF Diabetes Atlas (8th ed) (2017)
[9] American Diabetes Association2. Classification and diagnosis of diabetesDiabetes Care, 40 (2017), pp. S11-S24
[10] J.A. Beckman, M.A. CreagerVascular complications of diabetesCirculation Research, 118 (11) (2016), pp. 1771-1785

[11] K. Iglay, H. Hannachi, P. Joseph Howie, J. Xu, X. Li, S.S. Engel, *et al.*Prevalence and co-prevalence of comorbidities among patients with type 2 diabetes mellitusCurrent Medical Research and Opinion, 32 (7) (2016), pp. 1243-1252
[12] A.N. Long, S. Dagogo-JackComorbidities of diabetes and hypertension: mechanisms and approach to target organ protectionJournal of Clinical Hypertension, 13 (4) (2011), pp. 244-251
[13] S. Fazeli Farsani, P.C. Souverein, M.M. van der Vorst, C.A. Knibbe, A. de Boer, A.K. Mantel-TeeuwisseChronic comorbidities in children with type 1 diabetes: a population-based cohort studyArchives of Disease in Childhood, 100 (8) (2015), pp. 763-768
[14] S.P. Laing, A.J. Swerdlow, L.M. Carpenter, S.D. Slater, A.C. Burden, J.L. Botha, *et al.*Mortality from cerebrovascular disease in a cohort of 23 000 patients with insulin-treated diabetesStroke, 34 (2) (2003), pp. 418-421
[15] J.A. Critchley, I.M. Carey, T. Harris, S. DeWilde, F.J. Hosking, D.G. CookGlycemic control and risk of infections among people with type 1 or type 2 diabetes in a large primary care cohort studyDiabetes Care, 41 (10) (2018), pp. 2127-2135
[16] J.L. Hine, S. deLusignan, D. Burleigh, S. Pathirannehelage, A. McGovern, P. Gatenby, *et al.*Association between glycaemic control and common infections in people with Type 2 diabetes: a cohort studyDiabetic Medicine, 34 (4) (2017), pp. 551-557
[17] J.B. Kornum, R.W. Thomsen, A. Riis, H.H. Lervang, H.C. Schonheyder, H.T. SorensenType 2 diabetes and pneumonia outcomes: a population-based cohort studyDiabetes Care, 30 (9) (2007), pp. 2251-2257
[18] A. Mor, O.M. Dekkers, J.S. Nielsen, H. BeckNielsen, H.T. Sorensen, R.W. ThomsenImpact of glycemic control on risk of infections in patients with type 2 diabetes: a population-based cohort studyAmerican Journal of Epidemiology, 186 (2) (2017), pp. 227-236
[19] Gauri D. Kalyankar, Shivananda R. Poojara and Nagaraj V. Dharwadkar," Predictive Analysis of Diabetic Patient Data Using Machine Learning and Hadoop", International Conference On I-SMAC,978-1-5090-3243-3,2017
[20] Veena Vijayan V. And Anjali C, Prediction and Diagnosis of Diabetes Mellitus, "A Machine Learning Approach" ,2015 IEEE Recent Advances in Intelligent Computational Systems (RAICS) | 10- 12 December 2015 | Trivandrum
[21] Song, Yunsheng, Jiye Liang, Jing Lu, and Xingwang Zhao. "An efficient instance selection algorithm for k nearest neighbor regression." Neurocomputing 251 (2017): 26-34.
[22] P. Suresh Kumar and V. Umatejaswi, "Diagnosing Diabetes using Data Mining Techniques", International Journal of Scientific and Research Publications, Volume 7, Issue 6, June 2017 705 ISSN 2250-3153.
[23] Meng, X. H., Huang, Y. X., Rao, D. P., Zhang, Q., & Liu, Q. (2013). Comparison of three data mining models for predicting diabetes or prediabetes by risk factors. The Kaohsiung journal of medical sciences, 29(2), 93-99.
[24] Zheng, Tao, Wei Xie, LilingXu, Xiaoying He, Ya Zhang, Mingrong You, Gong Yang, and You Chen. "A machine learning-based framework to identify type 2 diabetes through electronic health records." International journal of medical informatics 97 (2017): 120- 127.
[25] Mani Butwall and Shraddha Kumar," A Data Mining Approach for the Diagnosis of Diabetes Mellitus using Random Forest Classifier", International Journal of Computer Applications, Volume 120 - Number 8,2015.
[26] Dost Muhammad Khan1, Nawaz Mohamudally2, "An Integration of K-means and Decision Tree (ID3) towards a more Efficient Data Mining Algorithm ", Journal Of Computing, Volume 3, Issue 12, December 2011.
[27] Xu, Weifeng, Jianxin Zhang, Qiang Zhang, and Xiaopeng Wei. "Risk prediction of type II diabetes based on random forest model." In Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB), 2017 Third International Conference on, pp. 382-386. IEEE, 2017.
[28] Rani, A. Swarupa, and S. Jyothi. "Performance analysis of classification algorithms under different datasets." In Computing for Sustainable Global Development (INDIACom), 2016 3rd International Conference on, pp. 1584-1589. IEEE, 2016.
[29] Kavakiotis, Ioannis, Olga Tsave, AthanasiosSalifoglou, NicosMaglaveras, IoannisVlahavas, and IoannaChouvarda. "Machine learning and data mining methods in diabetes research." Computational and structural biotechnology journal (2017).

[30] Smith, J.W., Everhart, J.E., Dickson, W.C., Knowler, W.C., & Johannes, R.S. (1988). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In Proceedings of the Symposium on Computer Applications and Medical Care (pp. 261--265). IEEE Computer Society Press.

[31] Islam, MM Faniqul, et al. 'Likelihood prediction of diabetes at early stage using data mining techniques.' Computer Vision and Machine Intelligence in Medical Image Analysis. Springer, Singapore, 2020. 113-125.

[32] Uttara Gogate, Harshita Bhagwat "Review of Machine Learning approaches in the Field of Healthcare",Cyber Security Threats and Challenges facing Human Life,Taylor & Francis Group, LLC 2021