IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 11, Issue 6, June 2022

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.118

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 6, June 2022||

| DOI:10.15680/IJIRSET.2022.1106004|

Prediction of Type 2 Diabetes Using Machine Learning Techniques

Sharmin Akhter, Khaulat Al Jahwari, P.Vijaya, Joseph Mani

Department of Mathematics and Computer Science, Modern College of Business & Science, Oman

ABSTRACT - Diabetes, whether Type I or Type 2, affects the way your body converts food into energy. Even though both types have similar symptoms like hazy vision, fatigue, and severe thirst, the way they develop is different. The main objective of this research work is to build a machine learning model to predict Type 2 diabetes in the initial stages for females. For this, Pima Indians Diabetes Database dataset from Kaggle is utilized along with four classification algorithms which are Logistic Regression, SVM, Decision Tree, and Random Forest. The dataset contains 8 attributes and 768 instances. 80% of the data was used as a training set and 20% as a testing set. Data normalization was applied to reduce the range between the values. Coding was done on Python IDLE application, and by using streamlit along with anaconda prompt, our work could be displayed on a webpage. For performance evaluation of the models, confusion matrix, accuracy score, precision score, recall score, and f1 score were used. Comparative analysis was done to see where our models are standing.

KEYWORDS: Machine Learning, Diabetes, Decision Tree, Logistic Regression, Support Vector Machine, Random Forest.

I. INTRODUCTION

A. Type 1 vs Type 2

Diabetes is a disease that impairs your body's ability to convert glucose into energy[10].There are several types of the disease, but the most common are types 1 and 2. Type 2 diabetes affects approximately 95% of diabetics, while type 1 affects approximately 5%. Insulin is the hormone that aids in the absorptionof blood sugar. With type 1 diabetes, you either stop producing insulin or produce insufficient amounts. With type 2 diabetes, insulin remains in your system, but your body becomes immune to it and it ceases to function. In either case, the end result is the same: more and more glucose accumulates in your blood. This can cause organ damage and serious, even life-threatening consequences over time [10].

B. Symptoms

Type 1 and type 2 diabetes symptoms include:

- Severe thirst
- Urine frequency
- Fatigue
- Hazy vision
- Changes in mood or irritability
- Hunger at its peak
- Hand or foot numbness or tingling
- Cuts that take a long time to heal

Type 1 and type 2 diabetes have symptoms that are similar, but they develop in diverse ways. Type 2 diabetes progresses slowly, sometimes over many years, and people with the condition may not realize they have it when it is quite advanced. Symptoms begin mild and progress slowly, becoming much more noticeable only when complications develop. Type 1 diabetes, on the other hand, develops quickly and can progress from mild to severe in a matter of weeks. Because type 1 diabetes is much more likely to receive a diagnosis in childhood or adolescence, it was previously referred to as "juvenile diabetes." Although type 2 diabetes can occur at any age, it is most commonly diagnosed in middle age. Type 1 diabetes, on the other hand, can strike at any age [12].

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 6, June 2022||

| DOI:10.15680/IJIRSET.2022.1106004|

C. Cause

Diabetes type 1 is an autoimmune disorder. When your immune system works properly, it attacks pathogens in order to fight disease and infection. It malfunctions and ruins healthy cells in the pancreas called B-cells that produce insulin in type 1 diabetes. It ultimately destroys most of these cells that the pancreas no longer produces insulin at all [11].

Type 2 diabetes develops as the body's cells become increasingly resistant to insulin. As a result, they become less capable of absorbing glucose and converting it to energy. Some people may eventually lose their ability to produce insulin entirely. A person's lifestyle choices, in addition to genetics, can be a risk factor for type 2. A poor diet, insufficient exercise, and being overweight can all boost the risk of developing the disease [12].

D. Diagnosis

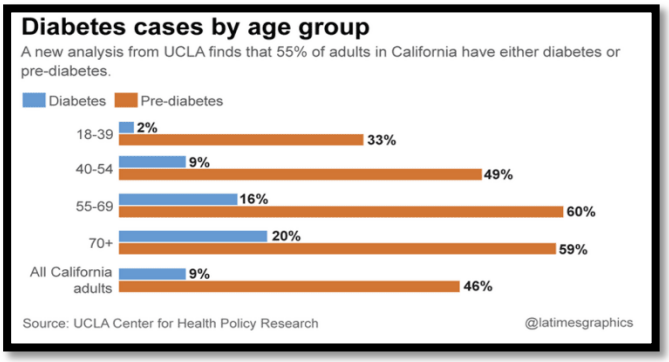
According to Diabetesresearch organization [12], diabetes can be diagnosed using a variety of tests.

- The A1C test, or glycated hemoglobin test, measures your average blood sugar levels over several months.
- A fasting blood glucose test determines your blood glucose levels after fasting overnight.
- A random blood sugar test is used to measure your blood sugar at an arbitrary time.
- An oral glucose tolerance test assesses your blood sugar levels after an overnight fast and again after drinking a sugary beverage.

E. Statistics

According to a new study from the University of California, Los Angeles (UCLA), a startling number of Californians have diabetes or are on the verge of developing it. In California, 55 percent of adults have diabetes or prediabetes, a disease in which a person's blood glucose levels are higher than normal but not high enough to be classified as diabetic [15].

Figure 1. Statistics showing the percentage of diabetes cases by age group



According to [37], diabetes is responsible for a large proportion of new cases of blindness, kidney failure, and amputations of the feet and legs that are not related to accidents in the United States. There is an increase in new cases of diagnosed patients with diabetes over the last thirty years with patients aged from eighteen to seventy-nine years old in U.S.

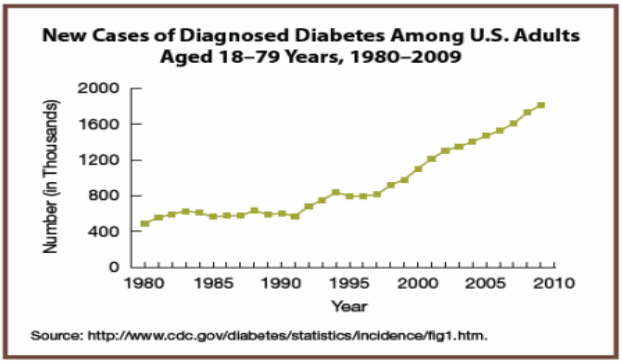


Figure 2. A curve showing new cases of diabetes from 1980 to 2009 in the US

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 6, June 2022||

| DOI:10.15680/IJIRSET.2022.1106004|

F. Machine Learning

Machine learning is a category of artificial intelligence aimed at allowing computers to "learn" without it being explicitly programmed [4]. It is all about getting computers to change their activities to enhance their accuracy, with accuracy being defined as the number of times the specified actions result in the right behaviors [3]. Arthur Samuel who was an American pioneer in the fields of computer games and artificial intelligence coined the term "machine learning" in 1959, defining machine learning as a branch of research that allows computers to learn without being explicitly programmed [4]. There are four types of machine learning paradigms which are supervised, semi-supervised, unsupervised, and reinforcement. Machine learning can be used to design and program explicit algorithms with high-performance output in a variety of computing areas, such as email spam filtering, online stock trading, face & shape detection, medical diagnosis, and so on [4]. We will be using supervised learning for our project to predict Type 2 diabetes using a machine learning model. Type 2 diabetes is widely recognized as a critical public health issue that has a significant impact on human life and healthcare costs. According to [21], Type 2 diabetes affects an estimated 462 million people worldwide, accounting for 6.28 percent of the global population. This disorder was responsible for almost one million deaths in 2017, making it the ninth highest cause of death. When compared to 1990, when Type 2 diabetes was the eighteenth largest cause of mortality, this is a concerning increase. Diabetes is the sixth most common disease in terms of human suffering (DALYs). It will be beneficial if Type 2 diabetes can be predicted for controlling the death rate due to this and this can be done using a trained machine learning model. We decided to use four classification algorithms under supervised learning which are logistic regression, decision trees, random forest and support vector machine (SVM). Logistic regression is a type of supervised learning in which probabilities are evaluated using the sigmoid function to estimate the relationship between a binary dependent variable and at least one independent variable. Logistic regression, contradictory to its name, is a type of machine learning classification problem in which the dependent variable is dichotomous (0/1, -1/1, true/false) and the independent variable can be binomial, ordinal, interval, or ratio-level [32]. In decision trees (also known as classification and regression trees (CART)), a series of decision rules are generated based on continuous and/or categorical input variables. Regression trees predict continuous results, while classification trees predict categorical results [4]. Random forest is a collection of decision trees built from randomly selected samples of the dataset. It votes on the output from every decision tree and categorizes an observation as diabetic or non-diabetic based on most of the decision trees' output [16]. Support vector machines (SVMs) are a class of supervised learning algorithms that are used to solve classification and regression problems. SVMs build an optimal boundary, known as a hyperplane, that best separates observations of distinct classes [4]. This boundary is a point in one dimension, a line in two dimensions, and a plane in three dimensions. This report is about using four different algorithms of supervised learning to decide which predicts Type 2 diabetes in females more accurately.

II. RELATED WORK

This segment will review various research papers that will be divided into six main parts which are diabetes review, machine learning review, logistic regression review, decision tree review, SVM (Support Vector Machine) review, and Random Forest review, respectively. Diabetes is quite common, and it is one of the diseases responsible for most deaths globally. It drains people financially due to its expensive and never-ending medication. As stated by Kopitar, et al. [22] 425 million people were estimated to have diabetes in 2017, among which 90% had Type 2 diabetes. Based on estimations for the coming years, for instance, in 2045, 48% of the numbers above are predicted to suffer from diabetes. Machine Learning can be used in various sectors like marketing, finance, transportation, and healthcare. Kenyon [20] states that Machine Learning algorithms in the healthcare sector can study many healthcare records and patients' data to detect patterns linked with diseases and health conditions. They can aid in the detection of tumors on scans and the detection of potential health issues as well. According to [5] there are three distinct styles of learning in machine learning algorithms: supervised learning, unsupervised learning, and semi-supervised learning.

Supervised learning uses labeled data during model training as elaborated by a research paper. Examples of problems are regression and classification. The commonly used classification algorithms especially in diabetes diagnosis are Naïve Bayes, Decision Tree, Logistic regression, and SVM. Ahamed, et al. [2] developed a model on PIMA Indian Dataset to predict type-2 diabetes using logistic regression, XGBoost, gradient boosting, decision trees, Extra Trees, random forest, and light gradient boosting machine (LGBM) classifiers. LGBM classifier was seen to overpower other classifiers with 95.20%. logistic regression gave the lowest accuracy of 75.2%.

Mujumdar, et al. [25] proposed a diabetes prediction model using a couple of supervised learning algorithms like Gaussian Naïve Bayes, K-Nearest Neighbor, and Logistic Regression for better diabetes classification and reduction of misdiagnosis chances so that a person does not get unnecessary treatments or no treatments at all when required. Jian, et al. [17] came up with a model to lessen the risks of developing significant complications related to diabetes like

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 6, June 2022||

| DOI:10.15680/IJIRSET.2022.1106004|

neuropathy and dyslipidemia which uses supervised learning algorithms. The models, AdaBoost, XGBoost, decision tree (DT CART), support vector machine (SVM), and logistic regression (LR), will be used to predict various complications linked to diabetes.

Chang, et al. [6] proposed an e-diagnosis system based on Machine Learning algorithms and implementation will be on the Internet of Medical Things (IoMT) environment. They have used three supervised learning models which are Naïve Bayes classifier, Random Forest classifier, and J48 decision tree model. These models were trained on Pima Indians Diabetes Database using R programming language. Precision, specificity, accuracy, and sensitivity were used to analyze the performance of the models. They used 70/30 split for training and testing set. They played around with the number of features (3, 5, then all of them) and checked the accuracy, precision, specificity, and sensitivity every time. It showed that when the features were reduced, Naïve Bayes algorithm provided the best accuracy and Random Forest provided the least. As all the features were used, the accuracy increased to 80% for Random Forest and it was the highest accuracy score. This made Chang, et al conclude that Naïve Bayes model does not work well with many correlated features and Random Forest will give best results when there are more features.

Deberneh [9] developed a machine learning model to predict Type 2 Diabetes occurrence in the following year. The dataset was taken from a private medical institute. The algorithms employed were random forest, logistic regression, XGBoost, support vector machine, and ensemble techniques. The accuracy of all these algorithms is almost the same. The highest accuracy is 73% from Random Forest, SVM, and the other ensembles techniques, and the lowest is 71% from Logistic Regression. Saxena, et al. [28] proposed a study to draw a comparative study of different classifiers and feature selection methods to predict diabetes with greater accuracy. The classifiers used were decision trees, K-nearest neighbor, and random forest with few feature selection techniques. Experiments were done on PIMA Indians diabetes dataset using Weka 3.9. Decision trees, K-nearest neighbor, and random forest obtained the accuracy of 76.07%, 78.58%, and 79.8% respectively.

In unsupervised learning, unlabeled data is used, and prediction is made by deducing structures that are found in the unlabeled data. Examples of problems can be dimensionality reduction and clustering. Semi-supervised is a blend of supervised and unsupervised learning, using labeled data in lesser amounts and unlabeled data in enormous amounts [7]. Sivashankari, et al. [30] proposed a heterogeneous ensemble model (stacked ensemble model) which uses different techniques like bagging, boosting, and stacking to predict whether a person is diabetes positively or negatively. This proposed model obtained the highest accuracy of 93.1% among other models which were LR, NB, and LDA which obtained an accuracy of 72%, 74.4%, and 81% respectively.

Logistic regression is a model that maps predicted values to probabilities ranging from 0 to 1 with the help of the sigmoid function [36]. Since logistic regression gives outcome in binary form, whenever probability is above 0.5, it will be allocated to one, and if the predicted probability is below 0.5, it will be allocated to zero. Research was done on a dataset from the UCI Machine Learning repository to figure out whether a person has diabetes based on various variables. The overall precision and accuracy of the model were good; however, it did not do well in terms of recall which means that the performance of the model in true positives is good but struggles in false negatives. Tigga, et al. [33] proposed a study to predict the risk of Type 2 diabetes among people. The model that was used is logistic regression and the dataset is Pima Indians Diabetes. Their model gave an accuracy of 75.32% which they considered to be a good accuracy. Another research was done on Pima Indians Diabetes Dataset by Joshi, et al. [18] using Logistic regression and decision tree as the models. Their Logistic Regression model yielded an accuracy of 78.26%.

The decision tree model has a sequential hierarchical structure including nodes where splitting of data occurs and leaves which represent final outcomes and leaves. Research was done on Pima Indians Diabetes DataBase using a decision tree model to diagnose diabetes where 30% of data was taken for testing and 70% for training. Overall accuracy, precision, and recall came out good. As the testing data and training data were kept the same (50%), the overall accuracy increased to which Dudkina, et al. [14] concluded that the more the training data, the greater the accuracy estimate will be obtained. Sadhasivam, et al. [27] built a model to predict diabetes disease using decision tree and three different feature selection models. The dataset that was used is Pima Indians Diabetes Dataset. The accuracy was evaluated in two different cases: when using decision tree only without feature selection techniques, and when using decision tree along with feature selection techniques. The accuracy obtained was higher (74.48%) when feature selection techniques weren't used.

SVM model creates the best hyperplane which can separate n-dimensional space into classes to place the new data point into a precise category. SVMs have effectively performed in classification as well as regression problems. A dataset from a public hospital in Colombia was taken and research was done on it. Eighty percent of the data were used for training the model, and 20% was used for testing. This model obtained the highest accuracy out of the other two models mentioned above (95.36%). Viloria, et al. [34] concluded that the accuracy could be increased by combining SVM with added computational techniques like genetic algorithms. Abbas, et al, [1] developed a predictive model based on support vector machine. Data used was generated from an oral glucose tolerance test (OGTT). The model

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 6, June 2022||

| DOI:10.15680/IJIRSET.2022.1106004|

obtained an accuracy of 96.8%; however, the results showed that for an efficient prediction, parsimonious clinical data needed to be collected.

Random Forest is an algorithm that uses the Ensemble technique. Ensemble means incorporating more than one model. As a result, rather than using a single model to make predictions, a collection of models is used. A study was done on the classification of diabetes using a random forest algorithm. The result showed an accuracy of 89% and AUC of 94.8% to which Wang, et al. [35] concluded that when a Random Forest classifier is combined with LASSO and SVM-SMOTE feature reduction method, it provides the best identification of people with higher Diabetes risk from individuals. Another research was done on a dataset from a hospital in Luzhou, China using three algorithms: Decision tree, Neural network, and Random Forest. Random Forest gave the highest accuracy of 80.8%. When Pima Indians dataset is used, Neural Network is shown to give the highest accuracy of 76.68% while Random Forest gave 76% and 72.75% with Decision Tree.

III. METHODOLOGY

In this project, we are going to "Pima Indians Diabetes Database" that is from Kaggle which has seven hundred and sixty-eight female patients and has eight attributes. The attributes or features present are:

- Pregnancies
- Glucose
- Blood pressure
- Skin thickness
- Insulin
- BMI (Body mass index): measures body fat based on the weight and height of a person
- Diabetes pedigree function: this function gives a score of the likelihood of diabetes based on family history
- Age
- Outcome

Our target variable is the outcome. This dataset is used in our classification algorithms namely, Logistic Regression, Decision Tree, SVM, and Random Forest to predict diabetes.

A. *Logistic Regression*

This is a sort of statistical analysis that is used in assessing the relationship between a dependent variable and one or more independent variables. The nature of the dependent variable is usually binary; however, there are cases when it is not binary. The output given will be discrete that ranges in the interval of 0 and 1.

1) Types of Logistic Regression:

Logistic regression, in general, refers to binary logistic regression with target variables that are binary. There are three more types of logistic regression that are classified based on response variables.

a) Binary logistic regression: This is the most basic type of logistic regression. In this type, the target variable can be in one of two classes, either 0 or 1.
b) Multinomial logistic regression: The target variable can be in one of three or more classes that are unordered, like class A, class B, or class C.
c) Ordinal logistic regression: The target variable can be in one of three or more classes that are ordered, like a score from 1 to 10.

A logistic regression model is derived from a linear regression model using the sigmoid function. This means that the outcome of the linear regression model is passed through the sigmoid function to compute the probability of an event then the logistic regression model maps the probability to a binary output. A sigmoid function maps any actual value into probability ranging in the intervals 0 and 1. This function generates an S-shaped graph which gives probability 1 when input is nearing positive infinity, and 0 when nearing negative infinity, giving a binary output.

Linear model:

$y = mx+c$

Sigmoid function:

$F(x) = \frac{1}{1+ e^{-x}}$

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 6, June 2022||

| DOI:10.15680/IJIRSET.2022.1106004|

Logistic Regression model:

$$y = \frac{1}{1+e^{-(mx+c)}}$$

B. SVM (SUPPORT VECTOR MACHINE)

This is a machine learning algorithm that uses supervised learning for classification as well as regression problems. The goal of this algorithm is to identify a hyperplane in an N-dimensional space that clearly categorizes the data points. After taking data as an input, SVM gives a line, hyperplane as an output which separates two classes. The hyperplane's dimension is determined by the number of input features present. The hyperplane is a line when there are 2 input features. It becomes a two-dimensional plane when the input features are 3.

C. DECISION TREE

It is a graphical depiction for obtaining all feasible solutions to a problem depending on given parameters. A decision treehas internal nodes representing the dataset's features, branches representing the rules for making decisions, and leaf nodes representing the result. In a decision tree, the prediction of a record's class label starts from the root node. A comparison is made, and if the condition is true, we move to the next node by following the branch that corresponds to that condition. Classification of a particular example is provided in the leaf node. Deciding which attribute to be kept in the root node and sub-nodes is not easy because selecting nodes randomly will give less accurate results. So, decision trees follow an approach known as ASM (Attribute Selection Measure). The most popular methods under ASM are Gini Index and Information Gain.

1) Gini Index:

It assesses the amount of impurity in a node. This method only works with target variables that are categorical like true/false. The only split it performs is binary. The higher the value of Gini Index an attribute has the less likely it will be preferred to be in a particular node.

$$1-\sum_{i=1}^{C}(p_i)^2$$

Figure 3. A formula for Gini Index

Where C refers to the total number of classes, i refers to a class, and p refers to the probability of picking a class.

2)Information Gain:

Before understanding informationgain is, we need to know the meaning of entropy. Entropy tells how random our data is (node's impurity). The greater the entropy, the more difficult it is to make any conclusions from the data. Since the objective of machine learning is to reduce the dataset's uncertainty, just by using entropy, we will not know whether the entropy of a certain node has been reduced. For this, we use Information gain which shows how the parent entropy has fallen after performing a split with an attribute. Node is split and a decision tree is built based on information gain's value. A node that is split first is the one with the greatest information gain.

Information Gain = E(Y)-E(Y|X) , where E(Y) is entropy of full dataset and E(Y|X) is entropy of dataset given some feature.

Decision trees mostly face an overfitting problem when it fits all the samples present in the training set affecting the accuracy when making a prediction. The most common solution used is pruning where unwanted branches are removed from the tree. It is done by removing decision nodes starting from down (leaf node) so as not to disturb the accuracy.

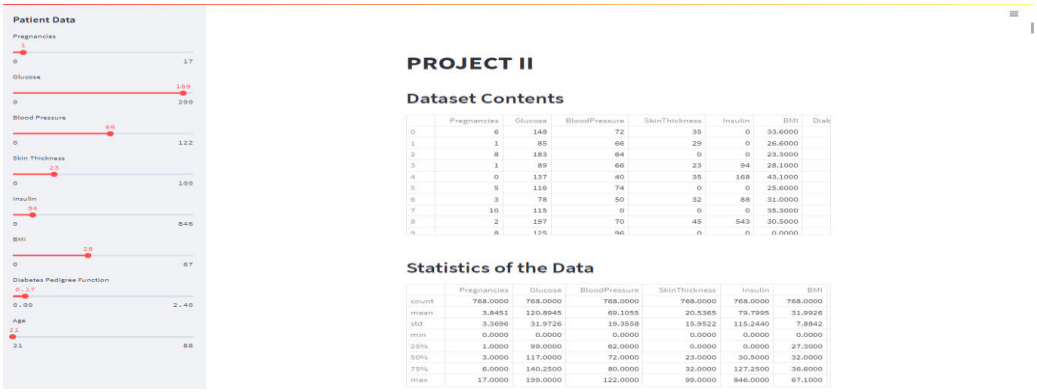
D.RANDOM FOREST

This is a supervised learning algorithm that uses the ensemble technique. The Ensemble technique combines various classifiers to deal with complex problems and enhance the model's performance. The random forest contains multiple decision trees which are collectively known as a forest. The forest is trained using a bootstrap aggregating or bagging algorithm. In this, different sets of training data are created by choosing data points randomly and replacing them. These subsets are trained individually. Then the random forest algorithm takes each tree's prediction. The final prediction is found by taking the majority of predictions. For instance, there are five decision trees and three of them gave a prediction of 1. The final output will be 1 because most of the trees gave this prediction. The greater the number of decision trees, the better the accuracy the model will give.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 6, June 2022||

| DOI:10.15680/IJIRSET.2022.1106004|



IV. EXPERIMENTAL RESULTS

Implementation is done on Python IDLE and the output is shown on a webpage using streamlit and anaconda prompt. Streamlit is used to create a webpage. This webpage can be accessed by running a code on anaconda prompt (streamlit run pjt2.py). On the webpage in figure 4, there is a side panel on the left side that shows all the 8 features as can be seen above. A patient can change the value of any feature accordingly and the result will be given as shown in figure 5. The model which provides this prediction will be chosen based on the accuracy score.

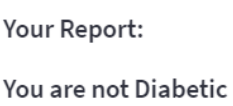


Figure 5. An example of the output

We made use of Python's libraries like NumPy which is used to execute mathematical operations on arrays, pandas which is used for manipulating and analyzing data and to provide flexible, expressive, and fast data structures so that working with labeled or relational data becomes easy, and matplotlib which is used for visualizing data and plotting graphs. The data in the dataset is split as follows: 80% as training set and 20% as testing set. Figure 6 shows a workflow of how our implementation will be done.

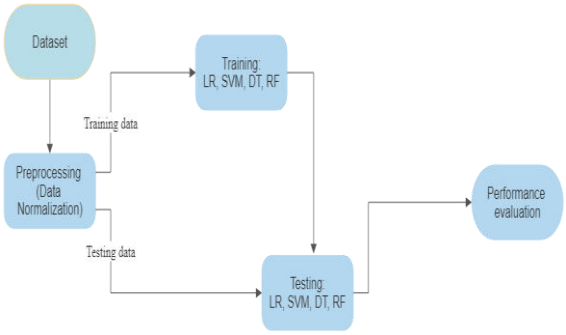


Figure 6. A diagram showing the steps of our project's implementation

To evaluate the performance of the models, we used confusion matrix, accuracy, precision, recall, and f1 score.

- Accuracy is obtained by: (number of predictions that are correct) / (total number of predictions made).
- Precision is obtained by: True Positive / (True Positive + False Positive)
- Recall is obtained by: True Positive / (True Positive + False Negative)
- f1 score is obtained by: 2 * ((precision * recall) / (precision + recall))

We will then compare the accuracy of our models to see which is the best in detecting diabetes.The Figure.7 shows how the patients are distributed in the dataset where 0 represents a nondiabetic and 1 represents a diabetic patient. Around 65.1 % of the patients are diabetic and 34.9 % are non-diabetic.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 6, June 2022||

| DOI:10.15680/IJIRSET.2022.1106004|

Algorithm	Precision		Recall		f1 score		Accuracy
	0	1	0	1	0	1	
SVM	0.84	0.76	0.92	0.60	0.88	0.67	0.8182
Random Forest	0.85	0.71	0.89	0.64	0.87	0.67	0.8117
Logistic Regression	0.84	0.76	0.92	0.60	0.88	0.67	0.8182
Decision Tree	0.88	0.65	0.82	0.74	0.85	0.69	0.7987

Table I.RESULTS OF THE MODEL'S PERFORMANCE

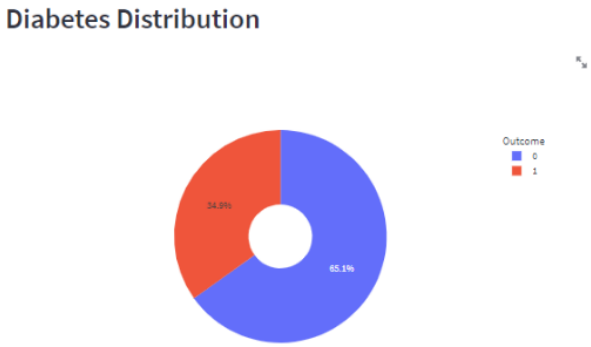


Figure 7. A pie chart showing diabetes distribution

Among all the 8 features, glucose had shown the most correlation to the outcome of value more than 0.4 and blood pressure the least of value less than 0.1 as can be seen in the below figure.

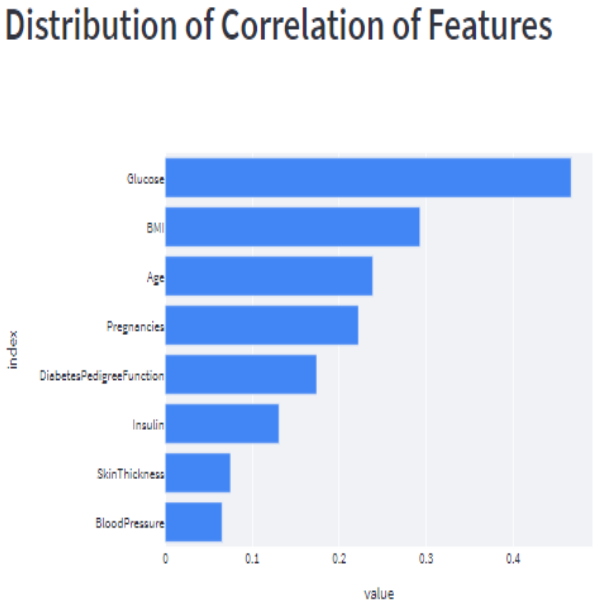


Figure 8. A bar chart showing the correlation of the features to the outcome

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 6, June 2022||

| DOI:10.15680/IJIRSET.2022.1106004|

In the Table I., decision tree has the highest precision of 0.88 for non-diabetic and SVM and logistic regression has least precision of 0.84 for non-diabetic. SVM and logistic regression has highest precision value of 0.76 for diabetic and decision tree has least value of 0.65 for diabetic. SVM and logistic regression has the highest recall value of 0.92 for non-diabetic and decision tree has the least value 0.82 for non-diabetic. Decision tree has the highest value of recall 0.74 for non-diabetic and SVM and logistic regression has the least value 0.60 for diabetic. SVM and logistic regression have the highest value of 0.88 for f1 score for non-diabetic and decision tree has the least value of 0.85 for non-diabetic. Decision tree has the highest value of 0.69 for f1 score for diabetic and SVM, logistic regression and random forest has least value 0.67 for f1 score for diabetic. It can be seen that SVM and Logistic regression gave the highest accuracy of 81.8% and Decision Tree gave the lowest, 79.87%. We notice that SVM and logistic regression has same values fordiabetic and non-diabetic for precision, recall, f1score and accuracy.

Table II- Comparison of Different Model Accuracy

S. No	Paper title	Author	Accuracy for LR in %	Accuracy for SVM in %	Accuracy for RF in %	Accuracy for DT in %
1	Prediction of Type-2 Diabetes Mellitus Disease Using ML Classifiers and Techniques	Ahamed, et al.	75.20		94.8%	94.40
2	Pima Indians diabetes mellitus classification based on ML algorithms	Chang, et al.			79.57	74.78
3	A Novel Approach for Feature Selection and Classification of Diabetes Mellitus: Machine Learning Methods	Saxena, et al.			79.83	76.07
4	Prediction of Type 2 Diabetes Based on ML Algorithm	Deberneh, H	71.0	73.0	73.0	
5	An Empirical Model to Predict the Diabetic Positive Using Stacked Ensemble Approach	Sivashankari, et al.	71	73.1	68.5	

Since Logistic Regression and SVM gave the highest accuracy as can be seen it Table 1, we will use Logistic Regression as the model to predict results in real-time for patients.Table II shows the comparison of different literature review accuracy.All of our models have provided an accuracy that is higher than most existing models. Therefore, the prediction obtained from our models will be pretty much accurate.

IV. CONCLUSION

Type 2 Diabetes is happening nowadays to even teenagers as well. It is important to have a balanced diet, regular exercise, and a limited intake of excessive sugar to reduce the risk of diabetes. Predicting Type 2 diabetes in the initial stages can limit the amount the damage is done to the body. This research paper writes about how we used machine learning techniques to predict diabetes in females using four different classification algorithms of supervised learning: logistic regression, decision trees, support vector machines, and random forest on the model. A diabetes dataset from "Pima Indians Diabetes Database" that is from Kaggle has been used to compare different machine learning algorithms and various steps were written to predict diabetes. The model SVM and logistic regression gave the highest model

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 6, June 2022||

| DOI:10.15680/IJIRSET.2022.1106004|

accuracy of 81.8% for predicting diabetes in females. All the results were displayed through a webpage created using Streamlit.

REFERENCES

[1] Abbas, H., Alic, L., Erraguntla, M., Ji, J., Abdul-Ghani, M., Abbasi, Q., and Qaraqe, M. Predicting long-term type 2 diabetes with support vector machine using oral glucose tolerance test. 11 December 2019. Available online: https://pubmed.ncbi.nlm.nih.gov/31826018/ (accessed on 20 June 2022).

[2] Ahamed, B., Arya, M., and Nancy, A. Prediction of Type-2 Diabetes Mellitus Disease Using Machine Learning Classifiers and Techniques. 10 May 2022. Available online: https://www.frontiersin.org/articles/10.3389/fcomp.2022.835242/full (accessed on 20 June 2022).

[3] Alzubi, J., Nayyar, A. and Kumar, A. Machine Learning from Theory to Algorithms: An Overview. 2018. Available online: https://iopscience.iop.org/article/10.1088/1742-6596/1142/1/012012/pdf (accessed on 31 March 2022).

[4] Bi, Q., Goodman, K., Kaminsky, J. and Lessler, J. What is Machine Learning? A Primer for the Epidemiologist. 2019. Available online: https://academic.oup.com/aje/article/188/12/2222/5567515?login=true (accessed on 31 March 2022).

[5] Brownlee, J. A Tour of Machine Learning Algorithm. 2021. Available online: https://machinelearningmastery.com/a-tour-of-machine-learning-algorithms/ (accessed on 22 March 2022).

[6] Chang, V., Bailey, J., Xu, Q., and Sun, Z. Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms. 2022. Available online: https://link.springer.com/article/10.1007/s00521-022-07049-z#Sec12 (accessed on 20 June 2022).

[7] DataRobot. Semi-Supervised Learning. 2020. Available online: https://www.datarobot.com/blog/semi-supervised-learning/ (accessed on 22 March 2022).

[8] Debereh, H. and Kim, I. Prediction of Type 2 Diabetes Based on Machine Learning Algorithm. 2021.Available online: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8004981/ (accessed on 22 March 2022).

[9] Deberneh, H. Prediction of Type 2 Diabetes Based on Machine Learning Algorithm. 2021. Available online: https://www.researchgate.net/publication/350335041_Prediction_of_Type_2_Diabetes_Based_on_Machine_Learning_Algorithm (accessed on 20 June 2022).

[10] Diabetesresearch.org. Type 1 vs Type 2 Diabetes. Available online: https://www.diabetesresearch.org/type-1-vs-type-2-diabetes (accessed on 13 June 2022).

[11] Diabetesresearch.org. Type 1 Diabetes | Diabetes Research. Available online: https://www.diabetesresearch.org/type-1-diabetes (accessed on 13 June 2022).

[12] Diabetesresearch.org. Diabetes Type 2 - Symptoms, Treatment & Prevention | DRIF. Available online: https://www.diabetesresearch.org/type-2-diabetes (accessed on 13 June 2022).

[13] Diabetes UK. What is HbA1c?. 2022. Available online: https://www.diabetes.org.uk/guide-to-diabetes/managing-your-diabetes/hba1c (accessed on 15 June 2022).

[14] Dudkina, T., Meniailov, I., Bazilevych, K., Krivtsov, S. and Tkanchenko, A. Classification and Prediction of Diabetes Disease using Decision Tree Method. Available online: http://ceur-ws.org/Vol-2824/paper16.pdf (accessed on 22 March 2022).

[15] Fidler, J. Diabetes and Prediabetes Rates in California: 'A Storm is Coming'.2021. Available online: https://naturalsociety.com/diabetes-prediabetes-rates-california-6159/(accessed on 15 June 2022).

[16] Fraser, A., Twombly, J., Zhu, M., Long, Q., Hanfelt, J., Narayan, K., Wilson, P. and Philips, S. Delay in the diagnosis of diabetes is not the patient's fault. 2010. Available online: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3971419/ (accessed on 22 March 2022).

[17] Ismail, I., Materwala, H., Tayefi, M., Ngo, P. and Karduck, A. Type 2 Diabetes with Artificial Intelligence Machine Learning: Methods and Evaluation. 07 October 2021. Available online: https://link.springer.com/article/10.1007/s11831-021-09582-x (accessed on 22 March 2022).

[18] Jian, Y., Pasquier, M., Sagahyroon, A., and Aloul, F. A Machine Learning Approach to Predicting Diabetes Complications. (2021). Available online: https://www.mdpi.com/2227-9032/9/12/1712/pdf (accessed on 14 June 2022).

[19] Joshi, R., and Dhakal, C. predicting Type 2 Diabetes Using Logistic Regression and Machine Learning Approaches. 9 July 2021. Available online: https://www.mdpi.com/1660-4601/18/14/7346 (accessed on 22 June 2022).

[20] Kaggle. Pima Indians Diabetes Database. 2016. Available online: https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database?resource=download (accessed on 26 April 2022).

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 6, June 2022||

| DOI:10.15680/IJIRSET.2022.1106004|

[21] Kenyon, T. Top 10 sectors for machine learning. 2021. Available online: https://aimagazine.com/top10/top-10-sectors-machine-learning (accessed on 14 June 2022).

[22] Khan, M., Hashim, M., King, J., Govender, R., Mustafa, H. and Al Kaabi, J. Epidemiology of Type 2 Diabetes – Global Burden of Disease and Forecasted Trends. 2020. Available online: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7310804/ (accessed on 31 March 2022).

[23] Kopitar, L., Kocbek, P., Cilar, L., Sheikh, A. and Stiglic, G. Early detection of type 2 diabetes mellitus using machine learning-based prediction models. 2020. Available online: https://www.nature.com/articles/s41598-020-68771-z (accessed on 22 March 2022).

[24] Lama, L., Wilhelmsson, O., Norlander, E., Gustafsson, L., Lager, A., Tynelius, P. and Ostenson, C. Machine learning for prediction of diabetes risk in middle-aged Swedish people. 2021. Available online: https://www.cell.com/heliyon/pdf/S2405-8440(21)01522-X.pdf (accessed on 21 March 2022).

[25] Mayo clinic. A1C test - Mayo Clinic. 2021. Available online: https://www.mayoclinic.org/tests-procedures/a1c-test/about/pac-20384643 (accessed on 26 April 2022).

[26] Mujumdar, A., and Vaidehi, V. Diabetes Prediction using Machine Learning Algorithms. 2020. Available online: https://www.sciencedirect.com/science/article/pii/S1877050920300557 (accessed on 14 June 2022).

[27] Naeem, A. What is sigmoid and its role in logistic regression?. Available online: https://www.educative.io/answers/what-is-sigmoid-and-its-role-in-logistic-regression (accessed on 15 June 2022).

[28] Sadhasivam, J., Muthukumaran, V., Raja, J., Joseph, R., Munirathanam, M., and Balajee, J. Diabetes disease prediction using decision tree for feature selection. 2021. Available online: https://iopscience.iop.org/article/10.1088/1742-6596/1964/6/062116/meta (accessed on 20 June 2022).

[29] Saxena, R., Sharma, S., Gupta, M., and Sampada, G. A Novel Approach for Feature Selection and Classification of Diabetes Mellitus: Machine Learning Methods. (2022). Available online: https://www.hindawi.com/journals/cin/2022/3820360/ (accessed on 20 June 2022).

[30] Shankaracharya, Odedra, D., Samanta, S., and Vidyarthi, A. Computational Intelligence in Early Diabetes Diagnosis: A Review. 2011. Available online: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3143540/ (accessed on 21 March 2022).

[31] Sivashankari, R., Sudha, M., Hasan, M., Saeed, R., Alsuhibany, S., and Abdel-Khalek, S. An Empirical Model to Predict the Diabetic Positive Using Stacked Ensemble Approach. 2022. Available online: Frontiers | An Empirical Model to Predict the Diabetic Positive Using Stacked Ensemble Approach | Public Health (frontiersin.org) (accessed on 14 June 2022).

[32] Soo, C., Kim, W., Yoo, T., Park, J., Chung, J., Lee, Y., Kang, E. and Kim, D. Screening for Prediabetes Using Machine Learning Models. 2014. Available online: https://www.hindawi.com/journals/cmmm/2014/618976/ (accessed on 20 March 2022).

[33] Tigga, N., and Garg, S. Prediction of Type 2 Diabetes using Machine Learning Classification Methods. 2020. Available online: https://www.sciencedirect.com/science/article/pii/S1877050920308024 (accessed on 31 March 2022).

[34] Tigga, N., and Garg, S. Predicting Type 2 Diabetes Using Logistic Regression. (2020). Available online: https://link.springer.com/chapter/10.1007/978-981-15-5546-6_42#:~:text=The%20aim%20of%20this%20study%20was%20to%20use%20logistic%20regression,a%20good%20predictor%20of%20diabetes. (accessed on 20 June 2022).

[35] Viloria, A., Beltran, Y., Cabrera, D. and Pineda, O. Diabetes Diagnostic Prediction Using Vector Support Machines. 2020. Available online: https://www.sciencedirect.com/science/article/pii/S1877050920305020 (accessed on 17 March 2022.

[36] Wang, X., Zhai, M., Ren, Z., Ren, H., Li, M., Quan, D., Chen, L., and Qiu, L. Exploratory study on classification of diabetes mellitus through a combined Random Forest Classifier. (2021). Available online: https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-021-01471-4 (accessed on 14 June 2022).

[37] Williams, S. 2013. The Top 3 Causes of Diabetes | The Motley Fool. [online] Available online: https://www.fool.com/investing/general/2013/05/25/the-top-3-causes-of-diabetes.aspx (accessed on 15 June 2022).

[38] Wilkinson, P. Introduction to Logistic Regression: Predicting Diabetes. 19 January. Available online: https://towardsdatascience.com/introduction-to-logistic-regression-predicting-diabetes-bc3a88f1e60e (accessed on 22 June 2022).

[39] Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y., and Tang, H. Predicting Diabetes Mellitus With Machine Learning Techniques. 2018. Available online: https://www.frontiersin.org/articles/10.3389/fgene.2018.00515/full (accessed on 13 June 2022).

INNO SPACE
SJIF Scientific Journal Impact Factor
Impact Factor: 8.118

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

Scan to save the contact details