Diabetes Prediction using Machine Learning

Naga Chetan Kumar Reddy
Computer Science
University of Central Missouri
Lee's Summit
cxn34080@ucmo.edu

Sridhar Seepana
Computer Science
University of Central Missouri
Lee's Summit
sxs29730@ucmo.edu

Saikuslu Gullapalli
Computer Science
University of Central Missouri
Lee's Summit
sxg32910@ucmo.edu

Yalavarthi rohith
Computer Science
University of Central Missouri
Lee's Summit
rxy55810@ucmo.edu

***Abstract*— One of the most fatal diseases in the world is diabetes. Inspite of this, Most of the people facing different types of problems like blindness, chronic kidney disease, hearing, cardiac failure, etc. In this case, the patient must visit a diagnostic facility to get their reports after consultation. With the advanced Machine Learning techniques, we now are planning a tool to have the flexibility to look for a solution to the current problem rather investing time and money.**

***Index Terms*—Machine Learning, Diabetes, Prediction.**

I. INTRODUCTION

The "Diabetes Prediction using Machine Learning" project is a predictive modeling project that utilizes machine learning techniques to predict the onset of diabetes in individuals. This project has significant importance as diabetes is a chronic disease that affects millions of people worldwide and can lead to serious complications if left untreated.

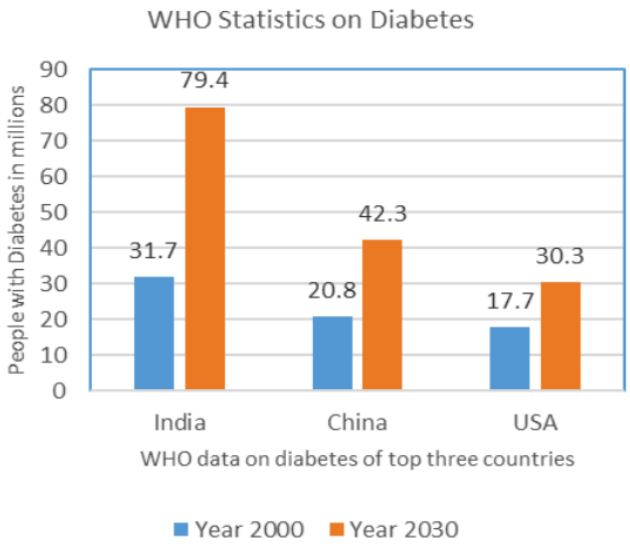

Fig. 1. WHO report on diabetes

In India, diabetes is becoming a fairly frequent condition. In India, there were approximately 31.7 million diabetic patients in 2000, and by 2030, there may be 79.4 million. The WHO figures for diabetes patients in India are displayed in Figure 1 [5]. This disease needs to be under control in India.

II. MOTIVATION

The primary motivation behind this project is to provide a tool that can accurately predict the likelihood of an individual developing diabetes. Early prediction can help individuals take preventive measures, such as lifestyle changes or medical treatment, to reduce the risk of developing the disease or manage the symptoms.

Three major sorts of errors can occur in the current medical diagnosis process:

1. The false-negative type, in which a patient is already diabetic but test results indicate otherwise.
2. False-positive kind, second. In this case, despite test results indicating otherwise, the patient is not actually diabetic.
3. The third type is unclassifiable, in which a system is unable to identify a specific case. The reason for this is that an unclassified form of prediction for a certain patient may result from insufficient knowledge extraction from historical data.

The patient must actually determine whether they fall into the diabetes or non-diabetic categories. Such diagnostic mistakes may result in therapies that are either unneeded or not provided at all when they are needed. It is necessary to develop a system using machine learning algorithms and data mining techniques that will offer accurate results and minimize human efforts in order to avoid or lessen the severity of such an impact.

III. RELATED WORK

Examining alternative component choice estimates for predicting diabetes serves as the introduction to a study. This drafted paper examines the fundamental motives for developing the diabetes illness. This information-mining-

https://github.com/700742973-SRIDHAR-SEEPANA/MACHINE-LEARNING_FINAL-PROJECT_DIABETES

based timetable and associated similar objects serve as the foundation for the illness forecasting framework. The order and foresight study for estimating the work that is actualized in the cloud are activated by this project work. The suggested idea uses artificial neural networks to forecast the onset of the diabetes illness. The factors altered as a result of considering information from a recent competition for system improvement, and the system built became successful. Where the current approaches and the results have been failing, the expectation system is diverging [2]. Here, we evaluate the prediction results using various evaluation metrics like classification accuracy, confusion matrix and f1-score.
Classification Accuracy- It is the ratio of number of correct predictions to the total number of input samples.
It is given as
$Accuracy = \text{Number of Correct Predictions} / \text{Total number of predictions Made.}$

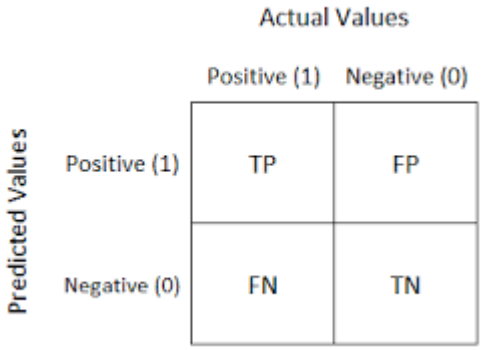

Fig. 2. Confusion Matrix

Confusion Matrix- It gives us gives us a matrix as output and describes the complete performance of the model.
Where, TP: True Positive FP: False Positive FN: False Negative TN: True Negative [4].
Accuracy for the matrix can be calculated by taking average of the values lying across the main diagonal.
It is given as-
$Accuracy = (TP+FN)/N$, Where, N:Total number of samples
F1 score-It is used to measure a test's accuracy. F1 Score is the Harmonic Mean between precision and recall. The range for F1 Score is [0, 1]. It tells you how precise your classifier is as well as how robust it is. Mathematically, it is given as-
$F1 = 2 * 1/ ((1/\text{precision}) + (1/\text{recall}))$
F1 Score tries to find the balance between precision and recall.

IV. PROPOSED METHODOLOGY

A relatively well-liked machine learning model is the Support Vector Machine (SVM) [20]. It uses a supervised machine learning model to function. In supervised machine learning, the model is trained using the recommendations of a teacher or critic. It is highly helpful in resolving classification-related issues. Numerous additional writers [5–14] have used machine learning algorithms for disease prediction and detection as well. Many studies use support vector machines to forecast a variety of diseases [11–14]. However, there is room for research to improve the SVM algorithm's effectiveness in predicting diabetes patients in the setting of India. SVM is a highly helpful tool for forecasting and monitoring health care, according to Harimoorthy et al. Additionally, SVD was used in this study to predict diabetes. The proposed work for diabetes detection using SVM is discussed in the previous section.

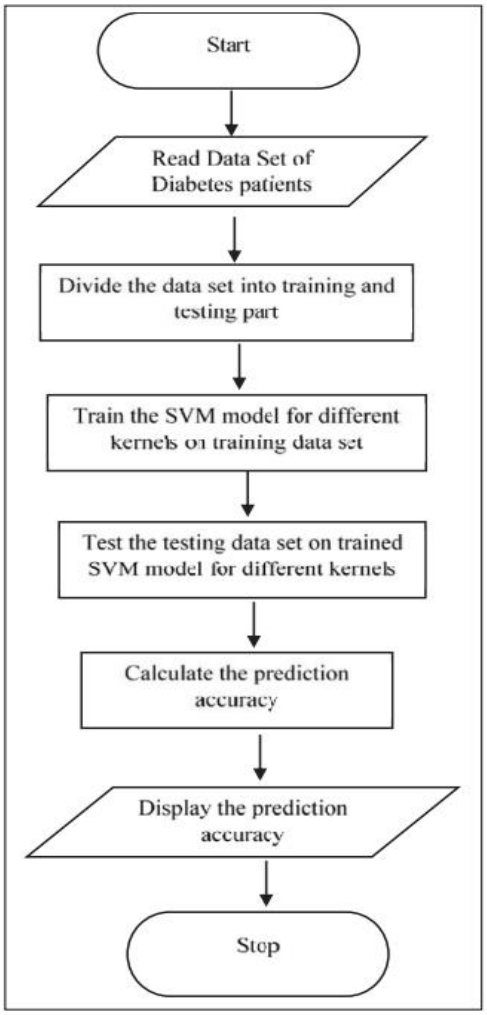

Fig. 3. Proposed Methodology

In order to predict diabetes, this article used the SVM machine learning technique. Python programming is used to implement and test the SVM algorithm on a collection of data. Python is a computer language that is used to generate the SVM model. Training and testing portions of the dataset are separated. The SVM model is then trained appropriately.

This study is innovative in that it proposes using SVM kernel performance on a set of medical data to forecast diseases. The model is trained using four of the available SVM kernels, and the testing set is used to calculate the prediction accuracy.

https://github.com/700742973-SRIDHAR-SEEPANA/MACHINE-LEARNING_FINAL-PROJECT_DIABETES

Four kernels—the linear, polynomial, sigmoid, and RBF kernels—are used to test the SVM. The optimal SVM kernel is chosen and applied for predicting diabetes. The proposed model's flowchart is shown in Figure 3.
The suggested concept is put into practice using the Python programming language and examined using a collection of 768 patient records. The data set was gathered from the N. Inst. of Diabetes & Diges. & Kidney Dis., which is freely accessible on the Kaggle platform. The data collection is accessible as a CSV file and is best used using the Python programming language. The SVM ML algorithm's ability to predict outcomes accurately depends on the model it uses. The models that are offered are the linear, polynomial, RBF, and sigmoid models. The SVM is initially trained and tested on several models. The attributes used in the algorithm are shown in the Figure 4.

```
In [7]: df.columns.values

Out[7]: array(['Pregnancies', 'Glucose', 'BloodPressure', 'SkinThickness',
               'Insulin', 'BMI', 'DiabetesPedigreeFunction', 'Age', 'Outcome'],
              dtype=object)

In [8]: df.dtypes

Out[8]: Pregnancies                int64
        Glucose                    int64
        BloodPressure              int64
        SkinThickness              int64
        Insulin                    int64
        BMI                      float64
        DiabetesPedigreeFunction float64
        Age                        int64
        Outcome                    int64
        dtype: object
```

Fig. 4. Attributes of the dataset

V. ALGORITHMS USED

A. Logistic Regression

Instead of being a regression model, logistic regression is a classification model. This algorithm categorises observations into a limited number of classes. Logistic regression changes its output using the logistic sigmoid function to generate a probability value that may then be translated to two or more discrete classes, in contrast to linear regression, which produces continuous numeric values. It's a classification model that's incredibly simple to implement and performs admirably with linearly separable classes. It is a widely used algorithm in industry for classification. A method for binary classification that can be expanded to multiclass or multi-attribute classification is the logistic regression model.

B. Random Forest

Supervised machine learning algorithms like random forest are frequently employed in classification and regression issues. On various samples, it constructs decision trees and uses their majority vote for classification and average in regress- sion cases. The Random Forest Algorithm's ability to handle data sets with both continuous variables, as in the case of regression, and categorical variables, as in the case of classification, is one of its most crucial qualities. In terms of classification issues, it delivers superior outcomes.

C. K-nearest neighbors

The k-nearest neighbours (KNN) technique is a fairly straightforward, easily implementable supervised machine learning algorithm that may be used to address both classification and regression problems. The K Nearest Neighbour method, as its name suggests, uses K Nearest Neighbours to forecast the class or continuous value for a given data point.

D. Adaboost Classififer

Even though we may temporarily alter the parameters or add more data, we ultimately continue to use the same model. Even if we create an ensemble, each trained model must be used on its own set of data. Boosting employs a more iterative methodology. Although it employs a cleverer strategy, it is still an ensemble technique in that any models are pooled to perform the final one.

E. Support Vector Machine(SVM)

A supervised machine learning model called the support vector machine (SVM) employs classification techniques to solve two-group classification issues. They can classify new text by providing SVM model sets with labelled training data for each category.

VI. ANALYSIS OF THE DATA WITH THE FEATURES AVAILABLE/ DATA DESCRIPTION

To be used in calculations by a machine learning model, the categorical features must be transformed into integers. Our data collection contains categorical variables that are not ordinal—that is, they have no inherent order. For instance, "DSL" internet service is not better than "Fibre optic" internet service. Ratings on a scale of one to five or a variable with the categories "bad," "average," and "good" are examples of ordinal categorical variables. Each category will be given a number when the categorical variables are encoded. The category with the greater number will be given more weight or influence over the model. Therefore, we must encrypt the variables so that each category is represented by a column, and the value of that column is either 0 or 1. Scaling continuous variables is also necessary. In the absence of this, variables with higher values will be given more weight, which affects the model's accuracy. For machine learning models, target variables having an unbalanced class distribution are undesirable. By randomly choosing rows from the class, we use up sampling, which means we increase the amount of samples from the class with less samples. We collected data from various sources, including lifestyle factors, and medical history. We conducted exploratory data analysis (EDA) to understand the relationships between the variables and identify any trends or patterns. We found that Pregnancies, Glucose,

Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function and Age were strong predictors of diabetes incidence.

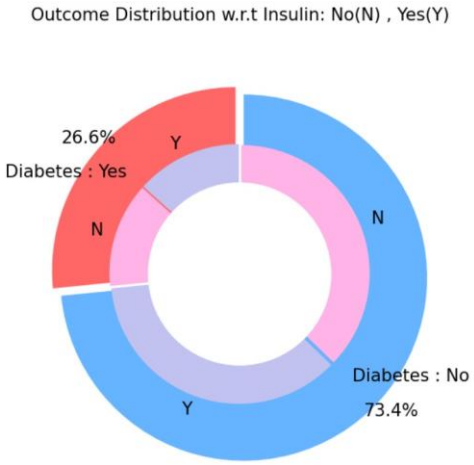

Fig. 5. Distribution of Diabetes prediction w.r.t Insulin

VII. RESULTS AND ANALYSIS

Here is the Fig. 6. we can observe the correlation between the each attribute that is the dataset that effects the diabetes prediction rate. Here, we present the individual graphs and confusion matrices for the machine learning algorithms that we utilised, allowing us to examine which method has created or predicted diabetes with the highest degree of accuracy.

Fig. 6. Correlation co efficient matrix.

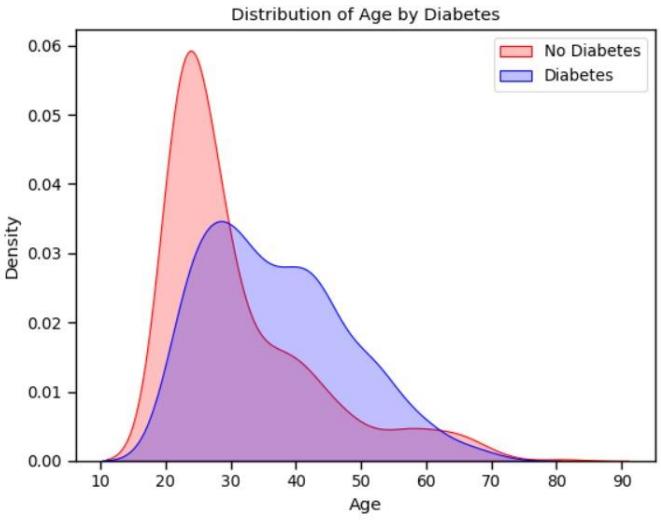

Fig. 7. Distribution of age by diabetes

A. *Random Forest*

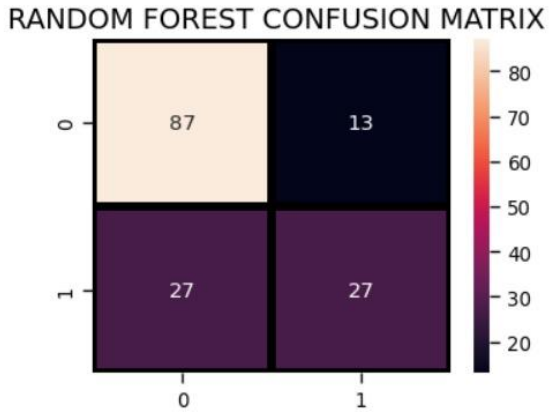

Fig. 8. Random Forest Confusion matrix.

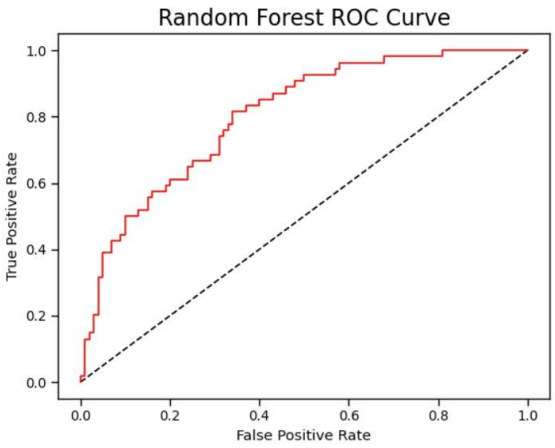

Fig. 9. Random Forest ROC curve.

https://github.com/700742973-SRIDHAR-SEEPANA/MACHINE-LEARNING_FINAL-PROJECT_DIABETES

B. Voting Classifier

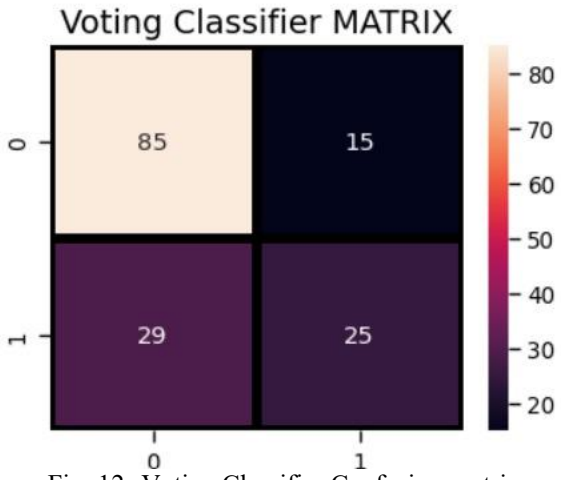

Fig. 12. Voting Classifier Confusion matrix.

C. Logistic Regression

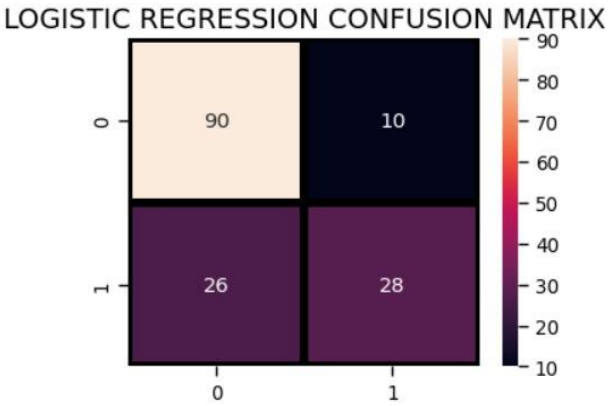

Fig. 10. Logistic Regression Confusion matrix.

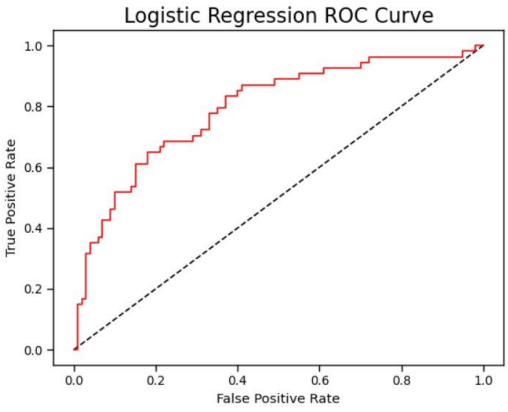

Fig. 11. Logistic Regression ROC curve.

	precision	recall	f1-score	support
0	0.75	0.85	0.79	100
1	0.62	0.46	0.53	54
accuracy			0.71	154
macro avg	0.69	0.66	0.66	154
weighted avg	0.70	0.71	0.70	154

Fig. 13. Final Classification Report.

VIII. CONCLUSION

In this study, different machine learning algorithms are applied to the dataset, and classification is done using different techniques, with the maximum accuracy being achieved by Logistic Regression which can be observed from the Figure 13. With the dataset used for the experiment, it is obvious that the model enhances diabetes prediction precision and accuracy. This research can be expanded further to determine the likelihood that non-diabetics will get diabetes during the following few years.

IX. REFERENCES

[1] S. Perveen, M. Shahbaz, K. Keshavjee and A. Guergachi, "Metabolic Syndrome and Development of Diabetes Mellitus: Predictive Modeling Based on Machine Learning Techniques," in *IEEE Access*, vol. 7, pp. 1365-1375, 2019

[2] K. L. Priya, M. S. Charan Reddy Kypa, M. M. Sudhan Reddy and G. R. Mohan Reddy, "A Novel Approach to Predict Diabetes by Using Naive Bayes Classifier," *2020 4th International Conference on Trends in Electronics and Informatics (ICOEI)(48184)*, Tirunelveli, India, 2020,pp. 603-607.

[3] P. Sonar and K. JayaMalini, "Diabetes Prediction Using Different Machine Learning Approaches," *2019 3rd International Conference on Computing Methodologies and Communication (ICCMC)*, Erode, India, 2019, pp. 367-371.

[4] Aishwarya Mujumdara, Dr. Vaidehi Vb," Diabetes Prediction using Machine Learning Algorithms", INTERNATIONAL CONFERENCE ON RECENT TRENDS IN ADVANCED COMPUTING 2019, ICRTAC 2019.

[5] Mohan, N., & Jain, V. (2020). *Performance Analysis of Support Vector Machine in Diabetes Prediction. 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA).* https://doi:10.1109/iceca49313.2020.9297411

[6] R. J. P. Princy, S. Parthasarathy, P. S. Hency Jose, A. Raj Lakshminarayanan and S. Jeganathan, "Prediction of Cardiac Disease using Supervised Machine Learning Algorithms," 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2020, pp. 570-575.

[7] R. Atallah and A. Al-Mousa, "Heart Disease Detection Using Machine Learning Majority Voting Ensemble Method," 2019 2nd International Conference on new Trends in Computing Sciences (ICTCS), Amman, Jordan, 2019, pp. 1-6.

[8] S. Mohan, C. Thirumalai and G. Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques," in IEEE Access, vol. 7, pp. 81542-81554, 2019.

[9] M. A. Alim, S. Habib, Y. Farooq and A. Rafay, "Robust Heart Disease Prediction: A Novel Approach based on Significant Feature and Ensemble learning Model," 2020 3rd International Conference on Computing, Mathematics and Engineering Technologies (iCoMET), Sukkur, Pakistan, 2020, pp. 1-5.

https://github.com/700742973-SRIDHAR-SEEPANA/MACHINE-LEARNING_FINAL-PROJECT_DIABETES

[10] Pouriyeh, S. Vahid, G. Sannino, G. De Pietro, H. Arabnia and J. Gutierrez, "A comprehensive investigation and comparison of Machine Learning Techniques in the domain of heart disease," 2017 IEEE Symposium on Computers and Communications (ISCC), Heraklion, 2017, pp. 204-207.
[11] Mir and S. N. Dhage, "Diabetes Disease Prediction Using Machine Learning on Big Data of Healthcare," 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), Pune, India, 2018, pp. 1-6.
[12] M. Patil, V. B. Lobo, P. Puranik, A. Pawaskar, A. Pai and R. Mishra, "A Proposed Model for Lifestyle Disease Prediction Using Support Vector Machine," 2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT), angalore, 2018, pp. 1-6.
[13] S. R. Alty, S. C. Millasseau, P. J. Chowienczyc and A. Jakobsson, "Cardiovascular disease prediction using support vector machines," 2003 46th Midwest Symposium on Circuits and Systems, Cairo, 2003, pp. 376-379 Vol. 1.
[14] S. Kaur and S. Kalra, "Disease prediction using hybrid K-means and support vector machine," 2016 1st India International Conference on Information Processing (IICIP), Delhi, 2016, pp. 1-6.
[15] R. S. Raj, D. S. Sanjay, M. Kusuma and S. Sampath, "Comparison of Support Vector Machine and Naïve Bayes Classifiers for Predicting Diabetes," 2019 1st International Conference on Advanced Technologies in Intelligent Control, Environment, Computing & Communication Engineering (ICATIECE), Bangalore, India, 2019, pp. 41-45.

https://github.com/700742973-SRIDHAR-SEEPANA/MACHINE-LEARNING_FINAL-PROJECT_DIABETES