DIABETES PREDICTOR USING ENSEMBLE TECHNIQUE

Prof. Pavitha N[1], Tanmay Patil[2], Mahalakshmi Phaldesai[3], Ritesh Pokarne[4], Prachi Kumar[5], Tanushri Bhuruk[6]

[1-6]*Department of Artificial Intelligence and Data Science, Vishwakarma Institute of Technology, Pune, Maharashtra, India.*

---***---

Abstract- *Diabetes mellitus, commonly referred to as chronic illness, is a group of diseases caused by high blood sugar. If early predictions are met, the risk and severity of diabetes can be greatly reduced. Reliable and accurate diabetes is extremely difficult due to the small amount of labelled data and the presence of suppliers of non-essential items other than the deficit in sugar-based data sets. We have developed a strong diabetes prediction framework for this paper, which mainly includes external rejection, data suspension, feature selection, various Machine Learning (ML) categories, Decision Trees, Random Forest, SVM, code integration, and retrospective), and Multilayer Perceptron (MLP). In this study, weighted integration of a few ML models was also proposed to improve diabetes prognosis, where weights were calculated using the ML model corresponding to the Area Under ROC Curve (AUC), F1 score, accuracy, and memory based on report classification. Using the Pima Indian Diabetes Dataset, all the studies in this book were conducted under the same screening settings.*

Keywords —. *Diabetes Prediction, SVM, Decision Tree, Ensembling Classifier, Machine Learning, Pima Indian Diabetic Dataset*

I. INTRODUCTION

Diabetes is a chronic metabolic condition characterized by elevated blood glucose levels caused by inadequate insulin use or insufficient insulin synthesis. In 2010, it was projected that 285 million individuals globally had diabetes (6.4 percent of adults). That number is predicted to climb to 552 million by 2030. Based on the present pace of disease progression, one out of every ten persons is anticipated to have diabetes by 2040. In South Korea, the prevalence of diabetes has risen substantially; according to current studies, 13.7 percent of all individuals have diabetes, and over a quarter have prediabetes.

Diabetes is usually not diagnosed because people with diabetes have no knowledge of the disease or have no symptoms; about a third of people with diabetes do not know their status. Diabetes causes serious long-term damage to many organs and systems, including the kidneys, heart, nerves, blood vessels, and eyes, if not properly controlled. As a result, early detection of the disease allows vulnerable people to take preventive measures to reduce the course of the disease and improve their quality of life.

Research has been done in many different fields, including machine learning (ML) and artificial intelligence (AI), in order to reduce the effects of diabetes and improve the quality of patient care. Many researchers have shown ML-based methods of predicting the incidence of diabetes. Current status (testing, diagnosis) and forecasting techniques are two types of these strategies. Current data classification is processed by current status recognition techniques; pre-screening methods are concerned with the classification of future data models. The aim of this activity is to create a machine learning model (ML) that can predict the onset of type 2 diabetes (T2D) in the following year (Y + 1) using the values of the current year (Y). Predictability models classify data entry conditions as normal (non-diabetic), prediabetes, or diabetes, depending on the condition. The performance of the predictive models, which included object rotation (LR), support vector support (SVM), descending order and resolution tree algorithms, were compared. We also evaluated the effectiveness of integration strategies such as the voting phase.

II. ALGORTHMS

The algorithms which have been used in this project are SVM (Support Vector Machine), MLP (Multilayer Perceptron), Decision Tree and the Ensemble classifier to integrate these algorithms.

III. LITERATURE REVIEW

[1]. TITLE: Predicting Type 2 Diabetes Using Mechanical Separation Methods

AUTHORS: Neha Prerna Tigga, Shruti Garg

Procedia Computer Science, Volume 167

YEAR:2020

In this paper six ways to categorize machine learning were available was applied and the results were compared. The trained database was then processed taken online and offline questions with 18 different questions these are the algorithms used on the PIMA website. Forest Algorithm provided accuracy which was 94.10% which was the highest among other algorithms.

[2]. TITLE: “Predictability model for type 2 diabetes based on data mining.”

AUTHORS: Wu, H., Yang S., Huang, Z., He, J., Wang, X.

YEAR:2018

This paper focuses on the advanced type of KNN and retrospective algorithms that have helped predict% of a person's risk of developing type 2 diabetes and found a 95.42% accuracy. the conversion and this value were collected by performing approximately 100 tests and the selected minimum number was 'within the total number of square errors'.

[3]. TITLE: "Diagnosis of Diabetes Using Classification Mining Techniques", IJDKP, vol. 5, no. 1, pp. 01-14

YEAR: 2015.

AUTHORS: A. Iyer, J. S, and R. Sumbaly,

This paper focuses on identifying solutions helpful for detecting diabetes. Mainly the classification analysis of the data is done by two algorithms are Naive Bayes and Decision Tree. Cross-validation approach and by using the PIMA dataset we got an accuracy of 74.8% for Decision Tree while Naive Bayes gave 79.5% but used a 70:30 split.

[4]. TITLE: Current Methods of Predicting Diabetes: A Review and Case Study

YEAR: 2019

AUTHORS: Souad Larabi-Marie-Sainte, Linah Aburahmah, Rana Almohaini, Tanzila Saba

This article is about multidisciplinary research conducted on a diabetes forecasting article that used algorithms such as tree cutting, Vector support (SVM), and foundation. They aimed to use dividers with rarely used categories and to evaluate their effectiveness. This article explores and evaluates diabetes guessing projects over the past 6 years using DL and ML techniques. Projects are based on the Pima Indian data set and tested using Wefa software Tool. The types of divisions used were Trees, Laziness, Laws, Jobs, and Bayes. Throughout the following categories accuracy, precision, recall, F rating, and ROC location were recorded. The highest accuracy obtained from the above dividers was 74.48% given by the REP tree (Reduce Error Pruning). He is a fast-paced and multi-tree decision-maker. His study time is short because he is a fast learner. It works on the basis of computer literacy using entropy. This process is also used to reduce the number of errors between actual output and expectations, which helps achieve a higher level of accuracy.

[5]. TITLE: Prediction of Type 2 Diabetes Based on Machine Learning Algorithm

YEAR: 2021

AUTHORS: Henock M. Deberneh and Intaek Kim

In this article, characteristics from the present year were used where they created a machine learning (ML) model to predict T2D occurrence in the following year (Y + 1) in the current year(Y). The resultant features used were g plasma glucose (FPG), HbA1c, triglycerides, etc. On the basis of the data, logistic regression, random forest, support vector machine, XGBoost, and ensemble machine learning algorithms were used to predict whether the patient was normal (non-diabetic), prediabetic, or had diabetes. The results of cross-validation (CV) showed that the ensemble models outperformed the single models. The differences in performance between the single models (LR, RF, SVM, and XGBoost methods) were insignificant. On the test dataset, the best accuracy for predicting the occurrence of diabetes was 73 percent, while the LR model had the lowest accuracy at 71 percent.

[6]. TITLE: Diabetes Predictability using Machine Learning Algorithms

YEAR: 2019

AUTHORS: Aishwarya Mujumdara, Dr. Vaidehi Vb

This paper was about the project description of Diabetic Predicting Models using the best classifications, factors used for glucose, BMI, age, insulin, etc. In their model, they used a single algorithm that combines various ML methods/modifications such as e Support Vector Classifier, Random Forest Classifier, Decision Tree Classifier, Extra Tree Classifier, Ada Boost algorithm, Perceptron, Linear Discriminant algorithm, Logistic Regression, K-Nearest Neighbor, Gaussian Naïve Bayes, Bagging algorithm, Gradient Boost Classifier.

Another algorithm used was piping. Pipes work by allowing the line sequence of data conversion to be tied together, leading to a measurable modeling process. The goal is to ensure that all the steps along the way, such as the training database or each wrap of the various validation methods, are in line with available testing data. They came out with 96 percent accuracy, Logistic Regression became the top model, while AdaBoost classifier was the best model with 98.8% accuracy.

[7] TITLE: Diabetes Mellitus Predictability using Machine Learning Algorithms

YEAR: 2019

AUTHORS: Hansen & Schnell

A WHO study of pregnant women revealed that 2 to 17.8% had diabetes during pregnancy. Diabetes mellitus is one of the major concerns in medical science research due to the extreme social effect of a particular disease, which produces large amounts of data. Therefore, without a doubt, when it comes to diagnostics, management and other related clinical management aspects, machine learning and data mining techniques in diabetes mellitus are of great concern. . Various methods have been developed based on the context of this study and therefore, a combination of machine learning and data mining has been proposed to differentiate diabetes. Obesity and gestational diabetes are complications that occur during pregnancy and ultimately affect the baby's generation during childbirth and fatal health (World Health Organization, 2013).

[8] TITLE: Prediction of Diabetes

YEAR: 2020

AUTHORS: (Goyal, Malik, Kumar, Rathore & Arora, 2020; Kumar et al., Mittal, Arora, Pandey & Goyal

Much research has been done on predicting diseases such as diagnosis, prognosis, isolation, treatment etc. Recent research shows that various ML (Machine Learning) algorithms have been used to diagnose and predict disease. They have resulted in spectacular performance and development in deep and familiar ML routes. ML has demonstrated its efficiency and effectiveness with high flexibility numbers while making solid speculation models. Monitored ML methods focus on word-dependent verification in the form of standalone / variable names. Predictable modelling is widely used in many areas of data mining and health care such as brain tumour detection and detection of significant, usually 85% of those identified by supervised learning methods and 15% by unsupervised, and especially, organizational rules.

[9] TITLE: Accuracy in Diabetes Prediction using ensemble

YEAR: 2008

AUTHORS: Grudzinski, Husain & Khan

It is an important way of transforming biological data sets into useful information, advanced clinical research, and improving health care. The need for segregation of diabetic patients as mentioned above has led to many improvements in ML strategies. The National Centre for Diabetes and Digestive and Kidney Diseases originally provided this Pima Indian diabetes data. It contains 769 data points of which 500 are diabetic and 268 are diabetic (data on Pima Indians Diabetes May 2008). Research history in this database shows that various machine learning algorithms and methods of integration have been used to classify diseases but none of them have been able to achieve more than 76% accuracy.

[10] TITLE: Ensemble Diabetes Prediction

YEAR: 2017

AUTHORS Smith Everhart, Dickson & Johannes

The proposed neural network algorithm ADAP algorithm to create an integrated model where they randomly selected data for training and achieved accuracy was 76%. Quinlan used the C4.5 learning model and the model performed well with 71.1% accuracy. Naive Bayes, J48, Radial basis function, Artificial neural network used for diagnosis of type 2 diabetes. 76.52%, 74.34% respectively.

[11] TITLE: Voting classifier Diabetes Prediction

YEAR: 2014

AUTHORS: Vijayan & Ravikumar

In 2014, Vijayan et al. used a variety of diabetes data mining techniques in 2017, the author shared the importance of AdaBoost and how to incorporate machine learning bags using J48 as a basis for forecasting diabetes. It surprisingly puts patients with diabetes and non-diabetics based on risk factors for diabetes. It was noted that the AdaBoost learning algorithm exceeds the package as well as the J48 algorithm.

IV. METHODOLOGY

This section of the literature provides a detailed description of the project's flow in the following sections: 1,2,3,4,5, which are Data set description, Data prepossessing, machine learning model with hyper parameter, Ensemble voting classifier, and project framework.

i) Data set description

ii) Data Preprocessing

iii) Models used

iv) Project framework

A. Data set description

This project's machine learning classifiers are trained on the publicly available PIMA Indian Diabetes dataset, which contains information about 768 females. As shown in table 1.1, the data set contains 268 diabetic patients and 500 non-diabetic patients with eight different attributes.

S No.	Attributes (F)	Description	Mean +_ Std
1	Pregnant	Total number of times got pregnant	3.85 ± 3.37
2	Glucose	Fasting glucose level	120.90 ± 31.97
3	Pressure	Diastolic Blood Pressure (mm Hg)	69.11 ± 19.36
4	Triceps	Skin thickness in triceps area (mm)	20.54 ± 15.95
5	Insulin	Serum Insulin	79.81 ± 115.24
6	BMI	Body mass index	32.00 ± 7.88
7	Pedigree	Diabetes Pedigree function	0.47 ± 0.33
8	Age	Age of patient	33.24 ± 11.76

Table 1.1

Figure a: The population distribution of all attributes in the data set, where the blue and violet colours denote non-diabetic and diabetic classes, respectively.

B. *Data preprocessing*

The framework for this project is shown in diagram fig: X where the first step is collecting raw data and pre-processing the data so that the data is used efficiently by the machine learning classifier for better results.

In the proposed framework, the pre-processing step includes outlier rejection (P), filling missing values (Q), standardization (R), and feature selection of the attribute which are briefly described as follows:

Outlier rejection(Q):

The outlier [23] is a markedly deviated observation from other observations. It requires to be rejected from data distribution as the classifiers are very much sensitive to the data range and distribution of the attributes The mathematical formulation for the outlier rejection in this literature can be written as in (2)

$P(x) = (x, \text{ if } Q1 - 1.5 \times IQR \leq x \leq Q3 + 1.5 \times IQR$ (2)

reject, otherwise

hear, x is the instance of feature vector that lies in n-dimensional space, $x \in Rn$. Q1, Q3, and IQR is the first quartile, third quartile, and interquartile range of the attributes respectively, where Q1, Q3, IQR $\in R^n$

Filling the missing value(Q):

The missing or null value in dataset can lead to false prediction by the classifier. In this type of data pre-processing the missing or null values are substituted with mean value of the attribute, rather than dropping it this can be represented by mathematical formulation as written in (3):

$Q(x) = (\text{mean}(x) \text{ if } x = \text{null/missed}$

x, otherwise (3)

here x is the instances of the feature vector that lies in n-dimensional space, $x \in R^n$

Standardization (R):

Standardization is a technique to rescale the attributes for the for achieving standard normal distribution with zero mean and unit variance and can be represented by mathematical formulation as written in (4)

$R(x) = x - \bar{x} / \sigma$ (4)

where x is the n-dimensional instances of the feature vector, $x \in R^n$. $\bar{x} \in R^n$ and $\sigma \in R^n$ are the mean and standard deviation of the attributes. However, in many ML models such as tree-based models are probably the models, where feature standardization can't provide a guarantee for significant improvement.

The accuracy of the classifiers increases with the increment of the attribute's dimension. However, the performance of the classifiers will tend to reduce when the attribute's dimension increases without increasing the samples.

C. Model used in this project

To carry out the task of classification we have used different machine learning models like K-nearest neighbour (KNN), Support vector machine (SVM), Decision tree (DT)and multi-layer perceptron (MLP) a brief description about the model is given in the Table 1.2. Ensembling of machine learning model is an approach to boost performance recall and the precision of the prediction.

Table 1.2

Machine learning model	Hyper parameter
K-nn	• N number of neighbour • Algorithms used for finding nearest neighbour • Ball Tree (BT): Node defines a D-dimensional hypersphere or ball • KD Tree (KDT): Leaf node is a D-dimensional point • Brute: Based on the brute-force search • Leaf size for BT or KDT which depends on the nature of problem • Metric (Manhattan distance (L1-norm) or Euclidean distance (L2-norm))
DT	• Measuring function: Entropy The strategy used for splitting at each node • Minimum sample for an internal node and leaf node

Support vector machine:

Support Vector Machine (SVM) is a set of related monitoring methods that are often used in medical diagnostics to differentiate and reverse. Simultaneously SVM reduces the error of empirical separation and enlarges geometric margins. SVMs can therefore be also called Maximum Margin Classifiers. SVM is a standard algorithm based on the proven responsibility for the study of mathematical learning theory, called the structural risk reduction principle. SVMs can optimize non-linear partitions using a so-called kernel trick, which clearly lays out their input into high-resolution feature areas. The kernel trick allows you to create a separator without clearly knowing the feature space.

Multilayer perceptron:

The neural network consists of neuron as small processing unit connected by unidirectional weighted connection. The D dimensional input vector is converted in n dimensional by the MLP. The output of each processing unit can be expressed as in (7).

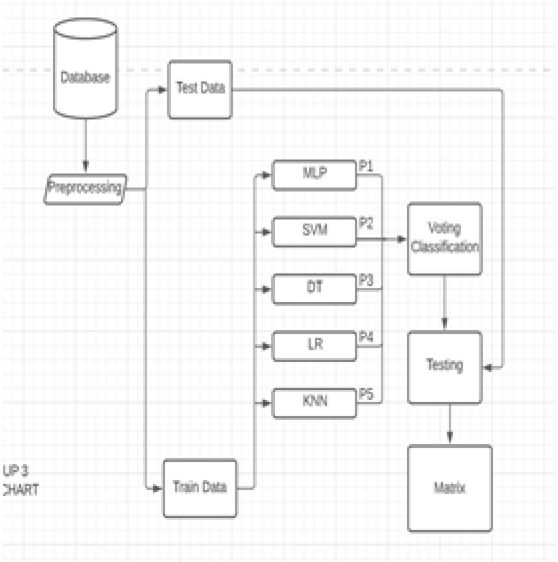
$$f(x) = \Phi\left(\sum_j w_j x_j + b\right) \tag{7}$$

where the xj, wj, b and Φ are the inputs, weights, bias to the neuron and the non-linear activation function respectively. The parameters of the neurons are updated as in (8) during the training using backpropagation to minimize the errors.

Voting classifier:

A voting classifier is a machine learning model that trains on ensemble of numerous models which have been used in the project namely DT, MLP, SVM, K-NN, LR. The ensemble model predicts output based on the best performance of the models which improves the prediction.

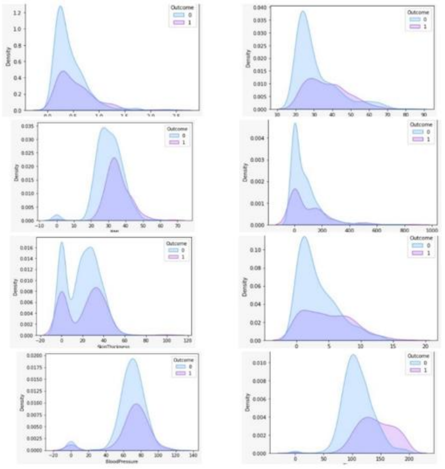
D. Project framework:



V. RESULTS AND DISCUSSIONS

Diabetes mellitus is a disease that affects many persons nowadays. As a result, early detection of this disease is critical. The major goal of this study is to find the most accurate and efficient algorithm for predicting diabetes patients. The accuracy of machine learning algorithms that were used in the preceding five years was investigated. As a result, the authors developed a soft voting classifier model based on a combination of four machine learning algorithms: decision tree, logistic regression, SVM, and MLP. The proposed model was first tested on the Pima Indians diabetes dataset, following which it was used to the breast cancer dataset. On the Pima Indians diabetes dataset, the ensemble soft voting classifier produced 79.08 percent accurate results and 97.02 percent accurate results on the breast cancer dataset. Using alternative deep learning models in the future, this accuracy could be improved.

a. Figures and Tables



These were the results we obtained from our model which were the different part of the confusion matrix.

ALGORITHM	ACCURACY
SVM	78%
MLP	78%
Ensemble Coding	75%
Logistic Regression	78%
Decision Tree	67%

In terms of recall, SVM shows a less of a reliable result set hence in terms of recall, I would suggest that our MLP and logistic regression models showed better results.

ALGORITHM	RECALL
SVM	0: 90% 1: 54%
MLP	0: 85% 1: 63%
Ensemble Coding	0: 81% 1:58 %
Logistic Regression	0: 88% 1:59 %
Decision Tree	0: 71% 1:58 %

For F1 score we see that best algorithm is logistic regression in terms of having the best F1 score. The best F1 score is when it is 1.

ALGORITHM	F1 SCORE
SVM	0: 84% 1: 62%
MLP	0: 83% 1: 66%
Ensemble Coding	0: 82% 1:56 %
Logistic Regression	0: 84% 1:65 %
Decision Tree	0: 75% 1:53 %

In terms of precision our SVM model showed the best results. Best score to get for precision is 1.

ALGORITHM	PRECISION
SVM	0: 79% 1: 73%
MLP	0: 82% 1: 68%
Ensemble Coding	0: 83% 1:54 %
Logistic Regression	0: 80% 1:72 %
Decision Tree	0: 63% 1:69 %

VI. FUTURE SCOPE

This research has certain drawbacks. For starters, there might be other risk variables in the diabetes dataset that the data collecting did not address. Other key variables, according to include gestational diabetes, family history, metabolic syndrome, smoking, sedentary lifestyles, particular food habits, and so on. To improve the accuracy of the prediction model, additional data would be needed. This may be accomplished by assembling diabetes datasets from many sources and creating a model from each one. In the future, a fuzzy set technique will be used to improve Bayes Network prediction, considering the unclear elements of specific diabetes variables. Other machine learning approaches, such as Neural Network, will also be examined to compare the forecasting outcomes to determine the best prediction model.

VII. CONCLUSION

As we can see, the accuracy was high for SVM, MLP, LOGISTIC REGRESSION models. So, these models will be used to predict our diabetes conditions in recent years better compared to others. However, I would like to add that our code compilation model, although showing only 75% accuracy, will be the best option in the long run and will work best in real-time data. Diagnosis of diabetes is made using a combined vote Pima class data class dividers for diabetes, by comparison with different classification algorithms, 80% maximum accuracy and 81% access to a set of data using 10 times the opposite verification and pronunciation data in 30% test and70% training.

REFERENCES

1. WHO. Diabetes. Available online: https://www.who.int/news-room/fact-sheets/detail/diabetes (accessed on 20 May 2020).

2. Shaw, J.; Sicree, R.; Zimmet, P. Global estimates of the prevalence of diabetes for 2010 and 2030. Diabetes Res. Clin. Pract. 2010, 87, 4–14. [CrossRef] [PubMed]

3. Zou, Q.; Qu, K.; Luo, Y.; Yin, D.; Ju, Y.; Tang, H. Predicting diabetes mellitus with machine learning techniques. Front. Genet. 2018, 9, 515. [CrossRef] [PubMed]

4. Won, J.C.; Lee, J.H.; Kim, J.H.; Kang, E.S.; Won, K.C.; Kim, D.J.; Lee, M.-K. Diabetes fact sheet in Korea, 2016: An

appraisal of current status. Diabetes Metab. J. 2018, 42, 415–424. [CrossRef] [PubMed]

5. Choi, S.B.; Kim, W.J.; Yoo, T.K.; Park, J.S.; Chung, J.W.; Lee, Y.-H.; Kang, E.S.; Kim, D.W. Screening for prediabetes using machine learning models. Comput. Math. Methods Med. 2014, 2014, 1–8. [CrossRef] [PubMed]

6. Deberneh, H.M.; Kim, I.; Park, J.H.; Cha, E.; Joung, K.H.; Lee, J.S.; Lim, D.S. 1233-P: Prediction of type 2 diabetes occurrence using machine learning model. Am. Diabetes Assoc. 2020, 69, 1233. [CrossRef]

7. Buch, V.; Varughese, G.; Maruthappu, M. Artificial intelligence in diabetes care. Diabet. Med. 2018, 35, 495–497. [CrossRef] [PubMed]

8. Dankwa-Mullan, I.; Rivo, M.; Sepulveda, M.; Park, Y.; Snowdon, J.; Rhee, K. Transforming diabetes care through artificial intelligence: The future is here. Popul. Health Manag. 2019, 22, 229–242. [CrossRef] [PubMed]

9. Woldaregay, A.Z.; Årsand, E.; Botsis, T.; Albers, D.; Mamykina, L.; Hartvigsen, G. Data-driven blood glucose pattern classification and anomalies detection: Machine-learning applications in type 1 diabetes. J. Med. Internet Res. 2019, 21, e11030. [CrossRef]