International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 05 Issue: 09 | Sept - 2021 ISSN: 2582-3930

Diabetes Prediction using Ensembling Methods in Supervised learning

S C YALLA REDDY[1], B SAI PRAVALIKA REDDY[2], M YOGESH[3], G KARTHIKA[4].

[1]*Student, Department of Computer Science Engineering, Gitam University ,Visakhapatnam.*
[2]*Student, Department of Information Technology, MIT Anna University ,Chennai.*
[3]*Student, Department of Computer Science Engineering, Gitam University ,Visakhapatnam.*
[4]*Assistant Professor, Department of Computer Science Engineering, Gitam University ,Visakhapatnam.*

ABSTRACT: Diabetes is a serious health problem in which the health state of diabetes patients must be monitored continuously because any sudden change in the blood sugar level can end up in serious risks in this paper we focus on a diabetic prediction model which can monitor the diabetic state of a patient by knowing minimum details about the patient through many research and forecast methods exist for the same this paper focuses on providing a user-friendly model. it also differs from the previous methods in that it ensembles k nearest neighbors convolutional neural network decision tree models which are expected to give high accuracy

Keywords- Blood sugar, k nearest neighbors, convolutional neural network, decision tree.

INTRODUCTION

Now-a-days there has been a drastic growth in medical field that is generated by electronic devices which opened on to effort less production of data according to 2017 statistics nearly 425 million mankind are suffering with diabetics if we come to India nearly 74 million humans are suffering with diabetics due to which many people are dead it is a serious issue in recent times if this not considered seriously then there will be many consequences to be faced in future.

There are different diabetes prediction methods that use machine learning algorithms and models to forecast diabetes states to understand the novelty of this paper let us take a look into various pre-existing models looking at the previously existing models it is found that the input details that all the previous models require for prediction can't be calculated at home so it is not useful for continuous monitoring. this paper focuses on building a model that will forecast diabetes based on data that can be calculated at home most of the papers focus on comparing different machine learning methods to find which technique is most suited for diabetes forecast based on the previous researches this paper tries to the ensemble k-nearest neighbors, convolutional neural network, decision tree and Random forest to build a model that provides very high accuracy most of the papers mentioned above focuses on the building of a model with high accuracy the deployment of such models with an easy user-friendly interface is essential for reaching common people this paper not only focuses on the collection of details from the users and displaying the results but also provides diet tips and exercises for users based on the predicted results

As diabetics is a serious disease which is commonly observed in the people aged above 18 years due to several reasons like obesity, family heredity, increasing of age, improper diet, no physical exercise. If diabetics is not diagnosed in time there will be so many consequences to be faced further.

SIDE EFFECTS OF NOT MONITORING DIABETES PROPERLY:

Long term effects of not monitoring diabetes can result in adverse health condition affecting many vital parts or one's body like eyes, brain, heart, kidneys, nerves.

The only solution to avoid such health issues is to monitor the health condition regularly. There are several techniques to predict diabetes condition. Problems with existing techniques:

- The laboratory methods provide a very high accuracy but is not easily available for daily use, so continuous monitoring is difficult.
- Training of machine learning models and using them to predict diabetes is an efficient solution. Already existing models require skin thickness, pedigree function for prediction. These details again can't be calculated at home, so continuous monitoring is not possible.

LITERATURE SURVEY

Priyanka Sonar et al.[1] focuses on the PIMA Indians Diabetes (PID). The author uses SVM, ANN, DT, Naive Bayes strategy to predict diabetes. A Comparison is done between various machine learning algorithms. A Machine learning Matrix is used to compare the models.

Muhammad Azeem Sarwar et al. [2] focuses on the PIMA Indians Diabetes(PID). The author uses KNN,

NB, SVM, DT, LR and RF. The author computes a machine learning matrix where accuracy is given by : Accuracy=(TP+TN) / (P + N)

MD.Kamrul Hasan et al. [3] focuses on the PIMA Indians Diabetes(PID). The author uses KNN, RF, DT, NB, Ada Boost, XGBoost (XB) and Multilayer Perceptron. The author tries to ensemble different models for higher accuracy. As a result author finds AB and XB is the best fusion for diabetes prediction.

Amani Yahyaoui et al. [4] focuses on the PIMA Indians Diabetes (PID). The author uses DSS, ML, DL, SVM, RF, CNN.The outcomes exhibit that RF was more successful in all rounds of examinations.

Ms. K Sowjanya et al. [5] focuses on training data like general body factors and family heredity. J48, NB, SVM and MLP. The results indicated that j48 was the best for the prediction.

Legila Alic et al. [6] focuses on the SAHS dataset with a total of 1496 participants. It used the Linear SVM to construct a prediction model of future growth of type-2 diabetes. The outcomes of the research show that high values of glucose noticed at the 2h mark through the OGTT may strongly indicate the dangerous future growth of type-2 diabetes.

MATERIALS AND METHODS

Data Collection:

This paper centers around PIMA diabetes dataset. The models are intended to
train on information that can be determined at home. Consequently the information dataset incorporates details, for example,

- Number of pregnancies
- Glucose
- Pulse
- Insulin
- BMI – Weight List
- Age
- Result (Diabetes condition)

Preprocessing:

This paper centers around building three models one dependent on KNN method, one dependent on CNN method and other dependent on DT. Pre handling is fundamental stage in ML. The preprocessing like bringing in libraries, understanding information, checking for missing qualities, the information, information parting are done on the obtained dataset for productive learning.

Training Of The Models:

Four ML models were to constructed. The primary model is to be assemble in view of KNN strategy which is a basic supervised learning method what's more, is appropriate for characterization issues. The subsequent model is to fabricate on CNN strategy which is a deep learning method which gives a very high precision with negligible preparing. The third model is Decision tree strategy which utilizes a tree-like model of decisions to outwardly and unequivocally represent decisions and decision making. Next is progressed rendition of decision tree known as Random forest. Each of the four models go through directed learning on the gathered dataset.

Testing Of The Models:

The four models which have gone through managed learning are tried for precision utilizing the test information. In the event that adequate exactness is arrived at the preparation interaction is finished effectively. In the event that adequate exactness isn't reached, a few changes are made in the model development and by and by the preparation is done.

Ensembling Of The Models:

After the three models are affirmed to give adequate precision, the three models are ensembled to give high precision. Casting a ballot and Averaging Based Ensemble Methods can be utilized to raise the precision of the expectation.

Evaluation:

This is the last stage of the model. Here, we assess the expectation results utilizing different assessment measurements like confusion matrix, accuracy.

MODEL ANALYSIS:

Let us take a look at the models that has been developed.

Fetching the Dataset and spliting data:

- Dataset: The dataset used for training is modified version of PIMA
dataset ie. Features such as age, Body Mass Index, Insulin, BP,
Pregnancy and Glucose alone are taken from standard PIMA
dataset along with the result (Presence/Absence of diabetes)
- X: Input variables of the model which include:
  - Age
  - Body Mass Index
  - Insulin
  - BP
  - Pregnancy
  - Glucose
- Y: Output variable of the model which is the result
- This is a binary classifier
(0: Presence of diabetes, 1:Absence of diabetes)
- test_train_split: Here we are using a ratio of 80:20 for training data:
testing data ratio.

K-Nearest Neighbor:

KNN is one of the least complex Machine Learning algorithm in view of Supervised Learning strategy.

- KNN algorithm expects the likeness between the new case/information also, accessible cases and put

the new case into the class that is generally like the accessible classifications.

• KNN algorithm stores every one of the accessible information and characterizes another information point dependent on the closeness. This implies when new information shows up then, at that point it very well may be effectively grouped into a well suite classification by utilizing KNN algorithm.

IMPLEMENTATION OF KNN OVER DATASET:

Prepropessing of data:

Normalize the information utilizing scaler function:

• StandardScaler adheres to Standard Normal Distribution (SND).

In this manner, it makes mean = 0 and scales the information to unit variance.

• MinMaxScaler scales every one of the information highlights in the reach [0, 1] or disaster will be imminent in the reach [-1, 1] in case there are negative qualities in the dataset. This packs all the inlier in the thin reach [0, 0.005]. Within the sight of anomalies, StandardScaler doesn't ensure adjusted element scales, because of the impact of the anomalies while figuring the observational standard deviation and mean. This prompts the shrinkage in the scope of the component esteems.

• By utilizing RobustScaler(), we can eliminate the exceptions and afterward use either StandardScaler or MinMaxScaler for preprocessing the dataset. By utilizing StandardScaler and RobustScaler precision could be expanded.

Building model:

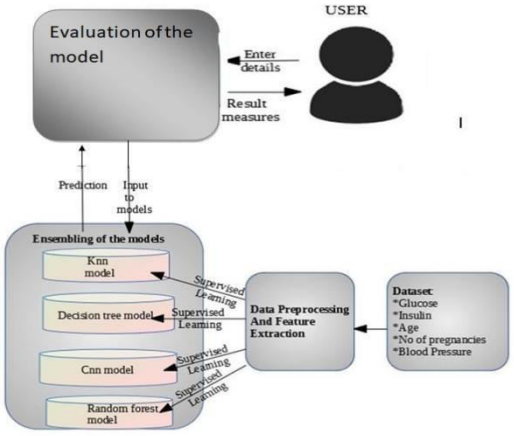

Fig 1: Basic Architecture

Parameter tuning:

• Fixing parameters is a significant piece of increasing accuracy of the model.

• Some significant parameters include:

- leaf_size
- p
- n_neighbors
- loads
- metric

• Grid object is prepared to do 10-overlap cross validation on a KNN model utilizing classification accuracy as the assessment metric

• GridSearchCV is a comprehensive pursuit over indicated boundary values for an assessor.

• We note that best leaf size is 1, best p is 2, best k value is 19, best metric is minkowski.

• But fixing k value we have to consider a certain equation to obtain maximum accuracy: square root of number of data samples

• So let us fix k=11

Decision Tree :

Decision Trees apply a succession of decisions or decides that regularly rely upon a solitary variable at a time. These trees segment our contribution to districts, refining the degree of detail at every cycle/level until we arrive at the finish of our tree, likewise called leaf node, which gives the last anticipated mark

Attribute Selection Measures

It is a heuristic way to deal with select the best parting basis that isolates guaranteed
information parcel, D, of class-marked preparing tuples into individual classes.

• Splitting measure is known as the best when subsequent to parting, each parcel will be unadulterated.

• A parcel is considered unadulterated when all the tuples that fall into the segment has a place with a similar class.

• ASM are otherwise called dividing rules since they decide how the tuples at a given node are to be parted.

• First, a position is accommodated each trait that depicts the preparation tuples. Also, the trait having the best score for the action is picked as the parting property for the given tuples.

• If the parting trait is nonstop esteemed or then again in case we are confined to parallel trees, then, at that point separately either a split point or a parting subset should likewise be not set in stone as a component of the parting model. In ASM there are two famous strategies, which are:

1.Information Gain
2.Gini Index

1. Information Gain:

We can characterize information gain as a proportion of how much data an element gives about a class. Data acquire assists with deciding the request for ascribes in the nodes of a DT. The fundamental node is known to as the parent node, though sub-nodes are known as child nodes.

Information Gain = Eparent − Echild

Where,

Eparent=Entropy of parent node

Echild= Avg Entropy of child nodes

2. Gini Index:

Gini list is a proportion of contamination or virtue utilized while making a DT in the CART algorithm. A characteristic having the low Gini list ought to be liked when contrasted with the high Gini file. It as it were makes double parts, and the CART utilizes the Gini record to make parallel parts. Gini list can be determined utilizing the below formula:

Gini Index= 1-ΣjPj2

Implementation of decision tree over dataset:

Test precision of the outcome:

Firstly,the accuracy was 0.66 and subsequent to changing the boundaries as per dataset. The accuracy has been expanded.

clf=DecisionTreeClassifier(random_state=10,criterion='gini',splitter='random',max_depth=6,max_leaf_nodes=17,min_samples_leaf=1,min_weight_fraction_leaf=0.0,min_samples_split=25,min_impurity_decreas e=0.0,ccp_alpha=0.0)

Convolutional Neural Network:

Convolutional neural networks are prepared out of different layers of artificial neurons. Artificial neurons are numerical capacities that ascertain the weighted amount of various data sources and yields an actuation esteem. The activity of duplicating values by loads and adding them is called "convolution". In view of the activation map the last convolution layer, the order layer results a bunch of certainty scores (values somewhere in the range of 0 and 1) that determine how possible the picture is to have a place with a "class."

Some significant parameters:

Dense:

Dense layer is the standard profoundly associated neural network layer. It is
generally and regularly utilized layer.

output = activation(dot(input, kernel) + bias)

{ input - address the information
activation - address the actuation work kernal – address weight information dot information and its relating loads address numpy dot product of all bias – address a one-sided esteem utilized in ML to enhance the model}

Conv2d:

Conv2D channels stretch out through the three directs in picture. After the convolutions are performed separately for each channels, they are amounted to get the last tangled picture. The yield of a channel after a convolution activity is known as featured map

Relu:

Relu does not initiate all the neurons at one particular time, which is the main advantage of the Relu. This will result the input straightforwardly if it is +ve, else, it will result 0.

Softmax:

Sigmoid function is the last actuation function of CNN network which result 0 or 1.

Binary crossentropy:

Binary crossentropy is a misfortune function that is utilized in binary classification
tasks. These are undertakings that answer an inquiry with just two decisions (yes or no, An or B, 0 or 1, left or right).

Adam:

It is a swap streamlining algorithm for SGD . This is truly effective when working with huge issue including a great deal of data or boundaries. It needs less memory and is effective

Two different approaches:

As we picked a csv document as our dataset we proposed 2 models.

• One model simply acknowledges contributions of 8 dimensions(8 input highlights)

• Other one reshapes it as a picture type info and trains on the include.

First approach:

Model summary:

Layer (type)	Output Shape	Param #
dense_10 (Dense)	(None, 500)	4500
dense_11 (Dense)	(None, 100)	50100
dense_12 (Dense)	(None, 2)	202

Total params: 54,802
Trainable params: 54,802
Non-trainable params: 0

Second approach:

CNN model where we change the input dimension alike an image and
perform training:

Here takes place the reshaping and building of model:
This is the model summary:

Layer (type)	Output Shape	Param #
dense_10 (Dense)	(None, 500)	4500
dense_11 (Dense)	(None, 100)	50100
dense_12 (Dense)	(None, 2)	202

Total params: 54,802
Trainable params: 54,802
Non-trainable params: 0

Random forest:
Random forest, in any case, it mainly utilized for Classification issue. RF settles on DT on data tests and subsequently gets the forecast from all of them in conclusion chooses the best game plan through a voting form. It is a social affair system which is superior to a single DT since it diminishes the overfitting by minimizing the outcome.

Working:
• Initially the determination of arbitrary examples from a dataset.
• After that it will develop a DT for each sample.Then, at this stage it gets the expected outcome from every DT.
• Then, casting a ballot will be performed for each predicted outcome.
• Finally, choose the most casted votes forecast result as the last expected outcome.

RESULT:
These are the outcomes of the machine learning models we got from the dataset

ALGORITHM	ACCURACY
KNN	83.7%
Decision tree	84.4%
CNN	74%
Random forest	77.9%

By ensembling all the Machine learning models the Final and overall accuracy of the Final model we have built is : **82.46%**

CONCLUSION:

Prediction of diabetes is done by ensembling various machine learning models like KNN, CNN, DECISION TREE, RANDOM FOREST with accuracies of 83.7%, 74%, 84.4%,77.9% respectively, And by comparing with various machine learning algorithms, the most elevated precision of 82.46% is accomplished form this model.

REFERENCES:
[1] P. Sonar and K. JayaMalini, "Diabetes Prediction Using Different Machine Learning Approaches," 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), 2019, pp. 367-371, doi: 10.1109/ICCMC.2019.8819841.
[2] M. A. Sarwar, N. Kamal, W. Hamid and M. A. Shah, "Prediction of Diabetes Using Machine Learning Algorithms in Healthcare," 2018 24th International Conference on Automation and Computing (ICAC), 2018, pp. 1-6, doi: 10.23919/IConAC.2018.8748992.
[3] M. K. Hasan, M. A. Alam, D. Das, E. Hossain and M. Hasan, "Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers," in IEEE Access, vol. 8, pp. 76516-76531, 2020, doi: 10.1109/ACCESS.2020.2989857.
[4] A. Yahyaoui, A. Jamil, J. Rasheed and M. Yesiltepe, "A Decision Support System for Diabetes Prediction Using Machine Learning and Deep Learning Techniques," 2019 1st International Informatics and Software Engineering Conference (UBMYK), 2019, pp. 1-4, doi: 10.1109/UBMYK48245.2019.8965556.
[5] K. Sowjanya, A. Singhal and C. Choudhary, "MobDBTest: A machine learning based system for predicting diabetes risk using mobile devices," 2015 IEEE International Advance Computing Conference (IACC), 2015, pp. 397-402, doi: 10.1109/IADCC.2015.7154738.
[6] H. Abbas, L. Alic, M. Rios, M. Abdul-Ghani and K. Qaraqe, "Predicting Diabetes in Healthy Population through Machine Learning," 2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS), 2019, pp. 567-570, doi: 10.1109/CBMS.2019.00117.