AI BASED DIABETICS PREDICTON SYSTEM

AI\_PHASE : 3



1. NAME: R . Sandhiya
2. DEPT/YEAR: ECE / IIIrd
3. NM ID:au513521106032

4.GMAILID: sandyas779@gmail.com

1.Data analysis: Here one will get to know about how the data analysis part is done in a data science life cycle.

2.Exploratory data analysis: EDA is one of the most here one will need to know that how to make inferences from the visualizations and data analysis

3.Model building: Here we will be using 4 ML models and then we will choose the best performing model.

4.Saving model: Saving the best model using pickle to make the prediction from real data.

Importing libraries diabetics prediction using MI

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

sns.set()

from mlxtend.plotting import plot\_decision\_regions

import missingno as msno

from pandas.plotting import scatter\_matrix

from sklearn.preprocessing import standardscaler

from sklearn.model\_selection import train\_test\_split

from sklearn.neighbors import kneighborsclassifier

from sklearn.metrics import confusion\_matrix

from skl arn import metrics

from learn.metrics import classification\_report

import warnings.filterwarning(‘ignor’)

warnings%matplotlib in line

diabetes\_df = pd.read\_csv('diabetes.csv')

diabetes\_df.head()

diabetes\_df = pd.read\_csv('diabetes.csv')

diabetes\_df.head()

 Output:

## Exploratory Data Analysis (EDA)

**Now let’ see that what are columns available in our dataset**

diabetes\_ d f .columns

Index(['Pregnancies', 'Glucose', 'Blood Pressure', 'Skin Thickness', 'Insulin',

'BMI', 'Diabetes Pedigree Function', 'Age', 'Outcome'],

dtype='object')

**Information about the dataset**

### Deployment of the prediction system

The proposed machine learning‐based diabetes prediction system has been deployed into a website and smartphone application framework to work instantaneously on real data.

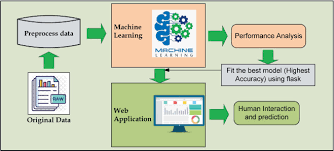
We have used HTML and CSS for the frontend part of the proposed website. After that, we finalized the machine learning model XGBoost with ADASYN, as it provided the best performance. The model deployment has been done with Spyder, a Python

platform that works with Anaconda. Figure shows the illustration of the

development process.

Development of the web application

Android smart phone application: To demonstrate the automatic diabetes forecasting system in real time, we also designed an Android smartphone application to test its performance. Android Studio is used for the front end part of this application. We employed Java as the necessary coding language. After that, the model has been implemented in Android Studio using the pickle package. While developing the API, we used Heroku to host our model on the corresponding hosting server. Figure [6](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10107388/figure/htl212039-fig-0006/) demonstrates the necessary steps in developing the proposed Android application.



1. The Supervised Learning/Predictive Models Supervised learning algorithms are used to construct predictive models. A predictive model predicts missing value using other values present in the dataset. Supervised learning algorithm has a set of input data and also a set of output, and builds a model to make realistic predictions for the response to new dataset. Supervised learning includes Decision Tree, Bayesian Method, Artificial Neural Network, Instance based learning, Ensemble Method. These are booming techniques in Machine learning.[3]

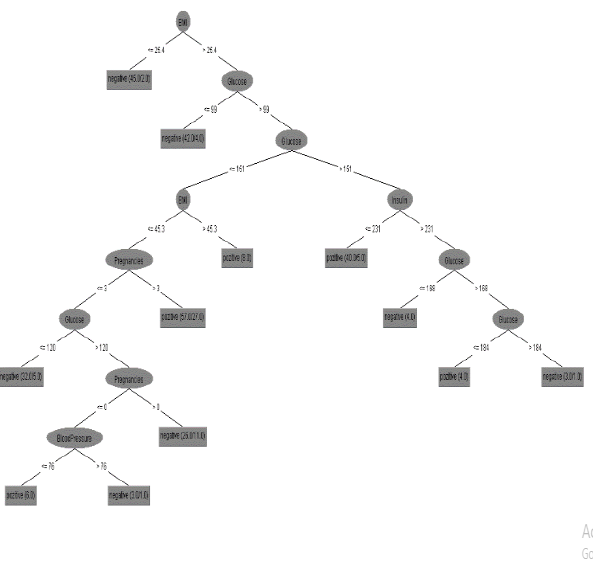
B. Unsupervised Learning / Descriptive Models Descriptive models are developed using unsupervised learning method. In this model we have known set of inputs but output is unknown. Unsupervised learning is mostly used on transactional data. This method includes clustering algorithms like k-Means clustering and k-Medians clustering.[3] C. Semi-supervised Learning Semi Supervised learning method uses both labeled and unlabeled data on training dataset. Classification, Regression techniques come under Semi Supervised Learning. Logistic Regression, Linear Regression are examples of regression techniques.[3] III. MOTIVATION There has been drastic increase in rate of people suffering from diabetes since a decade. Current human lifestyle is the main reason behind growth in diabetes. In current medical diagnosis method, there can be three different types of errors1. The false-negative type in which a patient in reality is already a diabetic patient but test results tell that the person is not having diabetes.

* 2. The false-positive type. In this type, patient in reality is not a diabetic patient but test reports say that he/she is a diabetic patient.
* 3. The third type is unclassifiable type in which a system cannot diagnose a given case. This happens due to insufficient knowledge extraction from past data, a given patient may get predicted in an unclassified type. However, in reality, the patient must predict either to be in diabetic category or non-diabetic category. Such errors in diagnosis may lead to unnecessary treatments or no treatments at all when required. In order to avoid or reduce severity of such impact, there is a need to create a system using machine learning algorithm and data mining techniques which will provide accurate results and reduce human efforts:
* . Experimental Results To conduct this study we used WEKA [7] software based on the approach and familiarity with its use. WEKA is an open source tool for data mining, which allows users to apply preprocessing algorithms but it does not provide assistance in terms of which one to apply.
* However, since different data mining algorithms have different requirements regarding the dataset, some preprocessing is applied by default inside some of the algorithms. Data preprocessing includes cleaning, instance selection, normalization, transformation, feature extraction, selection, etc. Data preprocessing affects the way in which outcomes of the final data processing can be interpreted.
* WEKA software package has different programs for different techniques and algorithms. Experiments are done by using Crossvalidation on default option folds= 10.
* Cross validation helps to improve the model results. The 10-fold cross validation technique has been used for better predictions. We have divided our dataset in to 10 samples
* . Each sample had to go from the process of retained as a validation data, where the rest 9 samples acted as a training data. This was a 10 times vice versa process. That's why it is call 10-fold cross validation. The advantage gained by this process step is that it cuts down the bias association with random sampling methods. Different classification algorithms
* slightly different as the working criteria of each algorithm is different. The results were evaluated on the basis of correctly classified instances, accuracy, precision, recall and fmeasureThis algorithm is clear and easy when we use it to interpret the It selects the attribute value of the data that most effectively separates the tested data into subset data which enriches the class.

The model construction is done by modifying the parameter values and this algorithm classifies diabetes disease data with a higher accuracy than other algorithms of data mining methods. This is shown in Table 3, it is the comparison of Accuracy of models after the implementation of algorithms.

1. racy, precision, recall and fmeasure

DECISION TABLE :

**7. Conclusion**

The purpose of this article was to create a

decision-making structure for diagnosing

diabetes.

* This structure was realized through

the study of classification data mining methods

such as Naive Bayes, Decision Tree, Support

Vector Machine (SVM), Logistic Regression

and their evaluation to show the highest

performing method on the dataset.

* The results of experiments conducted in this research by

implementing algorithms of data mining

methods have revealed that these methods are

applicable in the process of diabetes prediction.

* The decision tree as a data mining classification

method has classified diabetes data at an accuracy rate of 79%. This method has shown

promising results for the problem of diabetes

prediction as the accuracy rate is high in the

experiments performed. Furthermore, the

decision tree seems more viable due to the fact

that in contrast to other algorithms, it expresses

the rules explicitly. These rules can be

expressed in human language so that anyone

can understand them. Decision trees are easy to

interpret and understand.

* The use of machine

learning in analysis diabetes is important

because data mining methods and machine

learning can be used in the decision making

* In the future extension of this study

some models will be created for predicting the

diabetes that will help health centers, hospitals,

etc. to create policies or make decisions about

diabetes by preventing it.

* Algorithms’ behavior changes will be looked at when more data is

added. In the future we plan to do the same

study but this time not only on women but on

all persons regardless of gender.

* We also intend to implement this study to an integrated

Diabetes Decision Support System (DDSS)