PREMIER UNIVERSITY, CHATTOGRAM

Department of Computer Science & Engineering

NNFLL(CSE 452) Final Project Report

On

DIABETES PREDICTION USING ANN

SUBMITTED BY

Name: Bonnhi Shikha Parna

ID: 1903610201820

Name: Md Hafizul Islam

ID: 1903610201822

SUBMITTED TO

MR. FAISAL AHMED

Assistant Professor

Department of Computer Science & Engineering

Premier University, Chattogram

14 March, 2023

Table of Contents

Abstract

Diabetes Mellitus is among critical diseases and lots of people are suffering from this disease. Age, obesity, lack of exercise, hereditary diabetes, living style, bad diet, high blood pressure, etc. can cause Diabetes Mellitus. People having diabetes have high risk of diseases like heart disease, kidney disease, stroke, eye problem, nerve damage, etc. Current practice in hospital is to collect required information for diabetes diagnosis through various tests and appropriate treatment is provided based on diagnosis. Big Data Analytics plays an significant role in healthcare industries. Healthcare industries have large volume databases. Using big data analytics one can study huge datasets and find hidden information, hidden patterns to discover knowledge from the data and predict outcomes accordingly. In existing method, the classification and prediction accuracy is not so high. In this paper, we have proposed a diabetes prediction model for better classification of diabetes which includes few external factors responsible for diabetes along with regular factors like Glucose, BMI, Age, Insulin, etc. Classification accuracy is boosted with new dataset compared to existing dataset. Further with imposed a pipeline model for diabetes prediction intended towards improving the accuracy of classification.

Keywords: Diabetes,Prediction,Dataset,Ensemble,Diabetes Mellitus,Big Data Analytics,Healthcare.

CHAPTER 1

INTRODUCTION

1.1 Introduction

Diabetes is noxious diseases in the world. Diabetes caused because of obesity or high blood glucose level, and so forth. It affects the hormone insulin, resulting in abnormal metabolism of crabs and improves level of sugar in the blood. Diabetes occurs when body does not make enough insulin. According to (WHO) World Health Organization about 422 million people suffering from diabetes particu- larly from low or idle income countries. And this could be increased to 490 billion up to the year of 2030.
Diabetes Mellitus (DM) is classified as-
Type-1 known as Insulin-Dependent Diabetes Mellitus (IDDM). Inability of human's body to generate sufficient insulin is the reason behind this type of DM and hence it is required to inject insulin to a patient. Type-2 also known as Non-Insulin-Dependent Diabetes Mellitus (NIDDM). This type of Diabetes is seen when body cells are not able to use insulin properly.Type-3 Gestational Diabetes, increase in blood sugar level in pregnant woman where diabetes is not detected earlier results in this type of diabetes. DM has long term complications associated with it. Also, there are high risks of various health problems for a diabetic person. A technique called, Predictive Analysis, incorporates a variety of machine learning algorithms, data mining techniques and statistical methods that uses current and past data to find knowledge and predict future events. By applying predictive analysis on healthcare data, significant decisions can be taken and predictions can be made. Predictive analytics can be done using machine learning and regression technique. Predictive analytics aims at diagnosing the disease with best possible accuracy, enhancing

patient care, optimizing resources along with improving clinical outcomes.[1] Machine learning is considered to be one of the most important artificial intelligence features supports development of computer systems having the ability to acquire knowledge from past experiences with no need of programming for every case. Machine learning is considered to be a dire need of today's situation in order to eliminate human efforts by supporting automation with minimum flaws. Existing method for diabetes detection is uses lab tests such as fasting blood glucose and oral glucose tolerance. However, this method is time consuming. This paper focuses on building predictive model using machine learning algorithms and data mining techniques for diabetes prediction.

1.2 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 errors-
1. 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.

1.3 Objective

The paper is organized as follows- Chapter II-gives literature review of the work done on diabetes prediction earlier. Chapter III-presents methodology behind working on this topic. Chapter IV gives results and analysis. Chapter V gives conclusion.

CHAPTER 2

LITERATURE REVIEW

2.1 Previous Works

In this section, some closely related works are discussed briefly. In most of the reseach works, Pima Indians Diabetes Dataset (PIDD) have been used by many researchers for diabetes prediction. Various supervised machine learning algorithms were used to predict diabetes (Kaur Kumari, 2020). Radial basis function (RBF) kernel SVM, artificial neural network (ANN), multifactor dimensionality reduction (MDR), linear SVM and k-NN are some of them to mention. Based on p value and odds ratio (OR), Logistic Regression (LR) has been used to recognize the risk factors for diabetes (Maniruzzaman et al., 2020). Four classifiers have been adopted to predict diabetic patients, such as NB, DT, Adaboost, and RF. Partition protocols like- K2, K5, and K10 were also adopted, repeating these protocols into 20 trails. For the performance measurement of the classifiers, accuracy (ACC) and area under the curve (AUC) were analyzed.

Kopitar et al. (2020) showed a comparison of widely utilized regression models such as Glmnet, RF, XGBoost, LightGBM for predicting type 2 diabetes mellitus. The goal of this work was to examine if innovative machine learning methodologies gave any advantages in early prediction of impaired fast glucose and fasting plasma glucose (FPGL) levels compared to classic regression techniques [1].

For the prediction of diabetic patients, Maniruzzaman et al. (2020) have chosen four classifications such as naive bays (NB), decision tree (DT), adaboost and random forest. These methods were also implemented by three types of partition protocols (K2, K5, and K10). These classifers' performances are measured with precision (ACC) and curve surface (AUC) [2].

A hybrid model to detect type 2 diabetes was suggested by Albahli (2020). In order to extract unknown, hidden property from the dataset and to obtain more exact results, we use K-mean clustering, which is followed by the execution of a Random Forest and XGBoost classifier [3].

Yahyaoui et al. (2019) suggested a Machine Learning Techniques (ML) DSS for anticipating diabetes. They compared traditional machine learning with approaches to the deep learning. The authors applied the classifiers most typically used for a standard machine learning method: SVM and the Random Forest (RF). In contrast, they used a full-scale neural network (CNN) for Deep Learning (DL) to forecast and identify patients who suffer from diabetes [4].

Zou et al. (2018) predicted diabetes using the decision tree, random forests, and neural network. The dataset is collected from the Luzhou physical exams in China. The PCA was applied to reduce the dimension of the dataset. They selected several ML approaches to execute independent test to verify the universal applicability of method [5].

Supervised machine learning models which explore data-driven approaches were used to identify patients with diabetes diseases (Dinh et al., 2019). A complete research was conducted based on the National Health and Nutrition Examination Survey (NHANES) dataset. To develop models for cardiovascular, prediabetes, and diabetes detection, they have used all available feature variables within the data. Using various time frames and set of features within the data, different machine learning models, namely Support Vector Machines, logistic regression, gradient boosting and random forest were evaluated for the classification [6].

In Choubey et al. (2017) the authors used NBs for the classification on all the attributes. Afterwards GA was used as an attribute selector and NBs used the selected attributes for classification. The experimental results show the performance of this work on PIDD and provide better classification for diagnosis. Three specific supervised machine learning methods are used by Joshi and Chawan (2018), namely SVM, Logistic regression and ANN. His goal for research was to predict diabetes patients and he has also proposed an effective model for the prior detection of diabetes disease. Rajeswari and Prabhu (2019) focused on machine learning classification algorithms for predicting diabetes disease with more accuracy. Their study in SVM classification algorithm achieved highest accuracy. Various measures have been used to calculate the performance of classification algorithms [7].

An intelligent model using machine learning practices is developed (Nilashi et al., 2017) to identify diabetes disease. This model is constructed using approaches like clustering, removal of noise and classification, each of which made use of SOM, PCA and NN, respectively. The adaboost and bagging ensemble techniques are used to detect diabetes (Perveen et al., 2016). Along with standalone data mining technique, a base learner is used to identify patients with diabetes mellitus, namely J48 (c4.5) decision tree that makes use of multiple diabetes risk factors. In the Canadian Primary Care Sentinel Surveillance Network, three different ordinal adult groups are selected for classification. Experimental result shows that, the adaboost ensemble method shows better performance than both bagging and standalone J48 decision tree. For diagnosing T2DM, Kazerouni et al. (2020) has taken in consideration four different classification models, namely SVM, K-NN, ANN and LR. A comparison is done among these algorithms to measure the diagnostic power of this algorithms. The algorithms are performed on six LncRNA variables and demographic data [8].

CHAPTER 3

METHODOLOGY

3.1 Data Description

This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict whether a patient has diabetes, based on certain diagnostic measurements included in the dataset. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage.From the data set in the (.csv) File We can find several variables, some of them are independent (several medical predictor variables) and only one target dependent variable (Outcome) [9].

3.2 Preprocessing

Data preprocessing is the process of transforming raw data into a useful, understandable format. Real-world or raw data usually has inconsistent formatting, human errors, and can also be incomplete. Data preprocessing resolves such issues and makes datasets more complete and efficient to perform data analysis.This phase of model handles inconsistent data in order to get more accurate and precise results. This dataset contains missing values. So we imputed missing values for few selected attributes like Glucose level, Blood Pressure, Skin Thickness, BMI and Age because these attributes cannot have values zero. Then we scale the dataset to normalize all values.

One neuron on output layer

We used minmax scalling as preproccess.MinMax scaling is a data preprocessing technique that transforms the features of a dataset to a fixed range of values. Specifically, it scales the data so that it falls between a specified minimum and maximum value, usually 0 and 1. The formula for MinMax scaling is:
X_scaled = (X - X_min) / (X_max - X_min)
where X is the original feature value, X_min is the minimum value of that feature in the dataset, and X_max is the maximum value of that feature in the dataset. MinMax scaling is often used in machine learning algorithms that require data to be on the same scale, such as k-nearest neighbors (KNN) and support vector machines (SVM). It can also help to improve the performance of gradient descent-based algorithms, such as neural networks.

Two neuron on output layer

We used One-Hot encoding as preproccess to Concatenated the encoded dataframe with the original dataframe for two neuron on output layer.Concatenation is the process of joining two or more datasets together along a specific axis, either horizontally or vertically, to create a single dataset. The resulting dataset has the same number of rows or columns as the original datasets, depending on whether the concatenation was performed vertically or horizontally.

Horizontal concatenation is used when the datasets have the same number of rows, and we want to combine them to increase the number of columns. This operation is performed using the concatenate function along the axis of columns (axis=1). It is often used when we have datasets with different variables, and we want to combine them to create a single dataset with all the variables. Vertical concatenation is used when the datasets have the same number of columns, and we want to combine them to increase the number of rows. This operation is performed using the concatenate function along the axis of rows (axis=0). It is often used when we have datasets with the same variables, but different samples or observations, and we want to combine them to create a single dataset with more data points.We used it.

Concatenation is a useful technique for combining datasets with complementary information or for preparing data for machine learning tasks. However, it is important to ensure that the resulting dataset makes sense and that the concatenated features are not redundant or conflicting.

3.3 Classification

a. Decision Tree: Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome.

In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches. The decisions or the test are performed on the basis of features of the given dataset. It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions. It is called a decision tree because, similar to a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure. In order to build a tree, we use the CART algorithm, which stands for Classification and Regression Tree algorithm. A decision tree simply asks a question, and based on the answer (Yes/No), it further split the tree into subtrees.

b. K-Nearest Neighbor: K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique. K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories. K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm. K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems. K-NN is a non-parametric algorithm, which means it does not make any assumption on underlying data. It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.

KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data. Example: Suppose, we have an image of a creature that looks similar to cat and dog, but we want to know either it is a cat or dog. So for this identification, we can use the KNN algorithm, as it works on a similarity measure. Our KNN model will find the similar features of the new data set to the cats and dogs images and based on the most similar features it will put it in either cat or dog category.

CHAPTER 4

RESULTS AND ANALYSIS

4.1 Performance Measure

a. Accuracy: The accuracy metric is one of the simplest Classification metrics to implement, and it can be determined as the number of correct predictions to the total number of predictions.
It can be formulated as:
Accuracy = TP+TN/TP+FP+FN+TN
To implement an accuracy metric, we can compare ground truth and predicted values in a loop, or we can also use the scikit-learn module for this. We can use accuracy_score function of sklearn.metrics to compute accuracy of our classification model.

b. Precision: Precision, used in document retrievals, may be defined as the number of correct documents returned by our ML model. We can easily calculate it by confusion matrix with the help of following formula
Precision = TP/TP+FP
The precision metric is used to overcome the limitation of Accuracy. The precision determines the proportion of positive prediction that was actually correct. It can be calculated as the True Positive or predictions that are actually true to the total positive predictions (True Positive and False Positive).

c. Recall: Recall may be defined as the number of positives returned by our ML model. We can easily calculate it by confusion matrix with the help of following

formula
Recall = TP/TP+FN
It is also similar to the Precision metric; however, it aims to calculate the proportion of actual positive that was identified incorrectly. It can be calculated as True Positive or predictions that are actually true to the total number of positives, either correctly predicted as positive or incorrectly predicted as negative (true Positive and false negative).

d. F1-Score: This score will give us the harmonic mean of precision and recall. Mathematically, F1 score is the weighted average of the precision and recall. The best value of F1 would be 1 and worst would be 0. We can calculate F1 score with the help of following formula
F1 Score = 2 * Precision * Recall / (Precision + Recall)
F1 score is having equal relative contribution of precision and recall. We can use classification_report function of sklearn.metrics to get the classification report of our classification model. F-score or F1 Score is a metric to evaluate a binary classification model on the basis of predictions that are made for the positive class. It is calculated with the help of Precision and Recall. It is a type of single score that represents both Precision and Recall. So, the F1 Score can be calculated as the harmonic mean of both precision and Recall, assigning equal weight to each of them.

4.2 Results

We split the dataset to training data and test data to test the accuracy of the classifiers [10].

4.2.1 One neuron on output layer

The model accuracy shown in Fig 4.1.

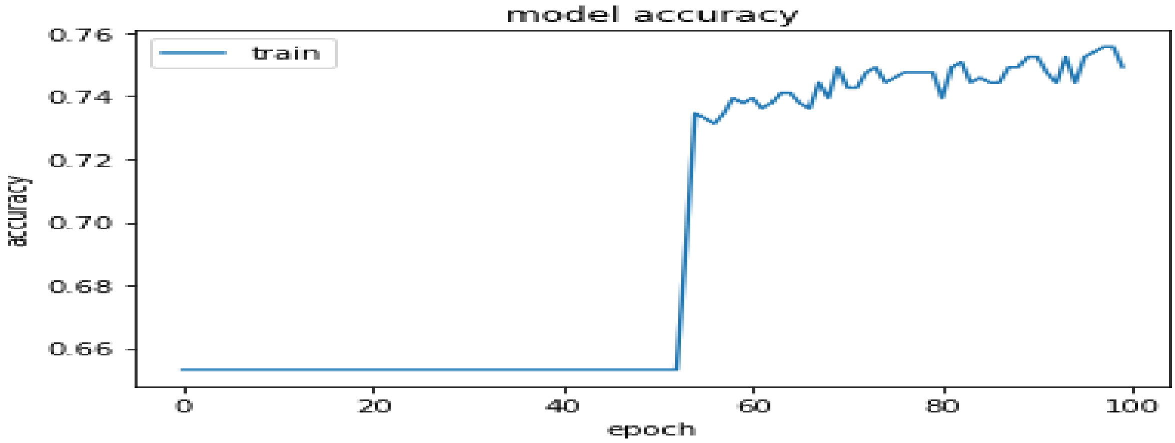

Figure 4.1. Model Accuracy

The model loss shown in Fig 4.2.

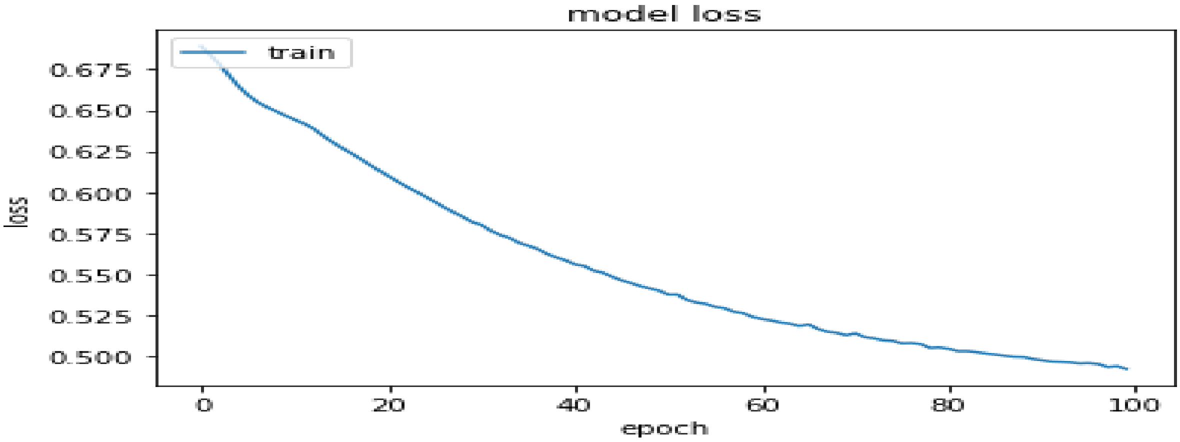

Figure 4.2. Model Accuracy

The table showing the classification report in Table 4.1

Classifier	Precision	Recall	F1-score	support
0	0.81	0.82	0.81	99
1	0.67	0.65	0.66	55
accuracy			0.76	154
macro avg	0.74	0.74	0.74	154
weighted avg	0.76	0.76	0.76	154

Table 4.1. Classification report.

The table showing the results of Algorithm we used in Table 4.2

Method used	Accuracy	Precision	Recall	F1-score
Decision Tree	0.75	0.76	0.75	0.76
KNN	0.69	0.68	0.69	0.68

Table 4.2. Result of different Algorithm.

4.2.2 Two neuron on output layer

The model accuracy shown in Fig 4.3.

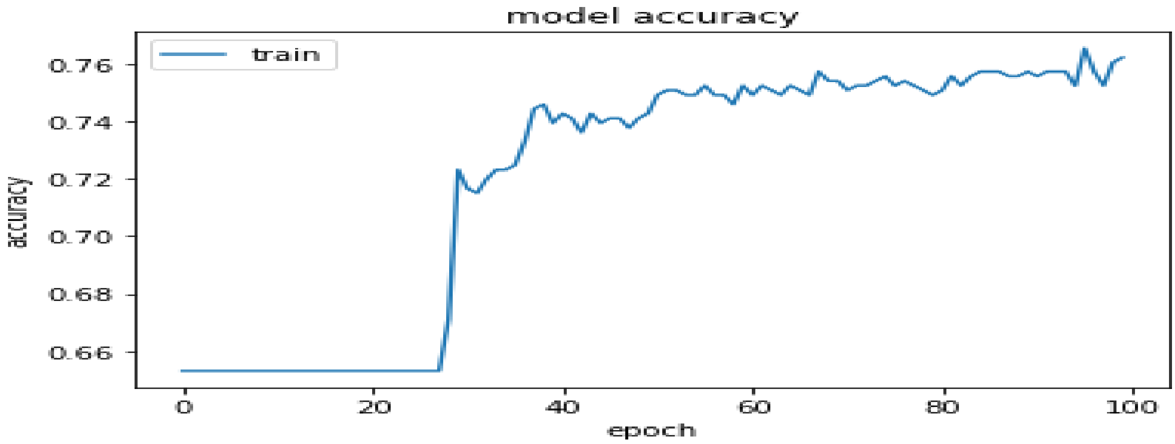

Figure 4.3. Model Accuracy

The model loss shown in Fig 4.4.

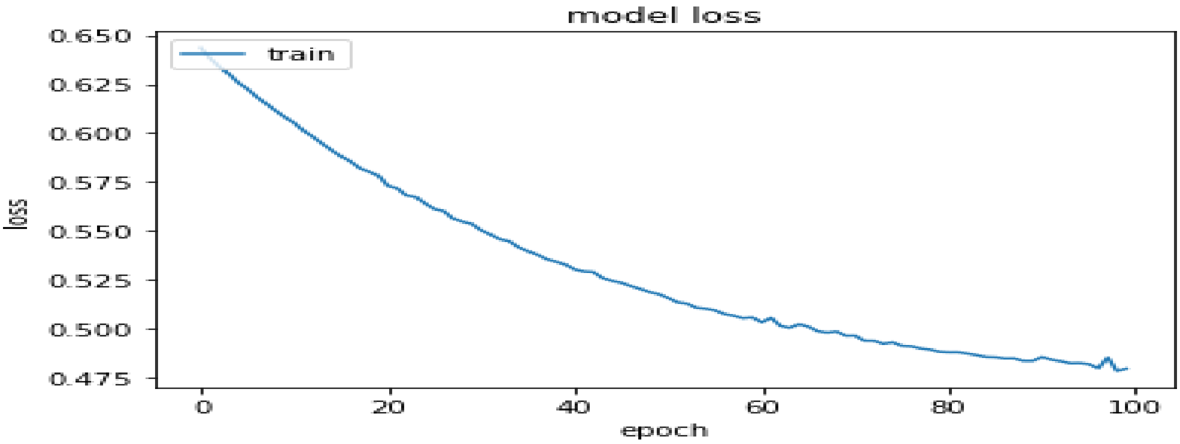

Figure 4.4. Model Accuracy

The table showing the classification report in Table 4.3

Classifier	Precision	Recall	F1-score	support
0	0.79	0.80	0.79	99
1	0.63	0.62	0.62	55
accuracy			0.73	154
macro avg	0.71	0.71	0.71	154
weighted avg	0.73	0.73	0.73	154

Table 4.3. Classification report.

The table showing the results of Algorithm we used in Table 4.4

Method used	Accuracy	Precision	Recall	F1-score
Decision Tree	0.77	0.78	0.77	0.77
KNN	0.68	0.69	0.68	0.69

Table 4.4. Result of different Algorithm.

4.3 Analysis

Various machine learning algorithms are applied on the dataset and the classification has been done using various algorithms of which Decision Tree gives highest accuracy of 75%. We have seen comparison of machine learning algorithm accuracies with two different output layer.In first we used one neuron on output layer,then we used two neuron on output layer.We observed that we get highest accuracy with two neuron on output layer.It is clear that the model improves accuracy and precision of diabetes prediction with this dataset compared to existing dataset. Further this work can be extended to find how likely nondiabetic people can have diabetes in next few years.

4.4 Discussion

The entire outcomes of the experiment in terms of accuracy, precision, recall, and f1-score are presented. For DT and KNN, the accuracy of these models is 77% and 68% respectively. This table illustrates that DT provide the highest level of accuracy and exceed the other approaches.Thus, all the classifier algorithms can detect the diabetes patients more accurately with the normalized and preprocessed dataset.

CHAPTER 5

CONCLUTION

One of the important real-world medical problems is the detection of diabetes at its early stage. In this study, systematic efforts are made in designing a system which results in the prediction of diabetes. During this work, five machine learning classification algorithms are studied and evaluated on various measures. Experiments are performed on john Diabetes Database. Experimental results determine the adequacy of the designed system with an achieved accuracy of 75% using Decision Tree algorithm. In future, the designed system with the used machine learning classification algorithms can be used to predict or diagnose other diseases. The work can be extended and improved for the automation of diabetes analysis including some other machine learning algorithms. In our project the result is classified into Yes or No. If the result is classified into No then we use time prediction module. Time Prediction - here we predict the ”time” of getting the diabetes disease. We analyze the result of the diabetes prediction and check the accuracy of the diabetes prediction, time taken to compute the accuracy of the diabetes prediction, correctly classification and incorrectly classification of result of the diabetes prediction. We have used KNN Algorithm and Decision Tree to predict the diabetes where result is classified into Yes or No . We compared the testing data and actual data to get the accuracy of our project.

BIBLIOGRAPHY

[1] *Comparison of widely utilized regression models*, Available online: https://www.sciencedirect.com/science/article/pii/S2666307421000279#bib0016 (accessed on: 26 January 2023).

[2] *Research on diabetes prediction*, Available online: https://www.sciencedirect.com/science/article/pii/S2666307421000279#bib0017 (accessed on: 30 January 2023).

[3] *A hybrid model to detect type 2 diabetes*, Available online: https://www.sciencedirect.com/science/article/pii/S2666307421000279#bib0002 (accessed on:31 January 2023).

[4] *A machine learning techniques (ml) dss for anticipating diabetes*, Available online: https://www.sciencedirect.com/science/article/pii/S2666307421000279#bib0027 (accessed on: 31 January 2023).

[5] *Predicted diabetes using the decision tree, random forests, and neural network*, Available online: https://www.sciencedirect.com/science/article/pii/S2666307421000279#bib0030 (accessed on: 03 February 2023).

[6] *Predicted diabetes*, Available online: https://www.sciencedirect.com/science/article/pii/S2666307421000279#bib0007 (accessed on: 10 February 2023).

[7] *Predicted diabetes using nbs for the classification*, Available online: https://www.sciencedirect.com/science/article/pii/S2666307421000279#bib0006 (accessed on: 15 February 2023).

[8] *Identify diabetes disease*, Available online: https://www.sciencedirect.com/science/article/pii/S2666307421000279#bib0018 (accessed on: 15 February 2023).

[9] *Diabetes dataset*, Available online: https://www.kaggle.com/datasets/akshaydattatraykhare/diabetes-dataset?select=diabetes.csv (accessed on: 26 January 2023).

[10] *Code implement*, Available online: https://colab.research.google.com/drive/1QvNiC1BLc11UN3hu4mr_ebm-O56IKfyo?usp=sharing (accessed on: 02 March 2023).