IMPLEMENTATION OF PREDICTING DIABETES DISEASE USING MACHINE LEARNING BASED UNIFIED FRAMEWORK

S. Srinivas[1] , T.Veeranna[2] , C.Dastagiraiah[3] , Dr.Kanusu Srinivasa Rao[4] , Ratna Kumari Challa[5]

[1] Department of Computer Science and Engineering, CVR Engineering College, Telangana, India
Email id: s.srinivas@cvr.ac.in

[2,3] Department of Computer Science and Engineering, Sai Spurthi Institute of Technology, B.Gangaram, Khammam, Telangana, India, Email id: veeru38@gmail.com, dattu50521@gmail.com

[4] Department of Computer Science and Tecnology, Yogi Vemana University, Kadapa, Andhra Pradesh, India., Email id: kanususrinivas@gmail.com

[5] Department of Computer Science and Engineering, RGUKT, AP-IIIT, Idupulapaya Kadapa, India.
Email id:ratnamala3784@gmail.com

ABSTRACT:

One of the chronic and deadliest diseases cause an increase the sugar level in blood is diabetics. In fact, diabetes can inflict many severe effects like burning extremities, kidney & heart failures, myopia, blurred vision. When the human body can not contain enough insulin for regulating the threshold or sugar levels reaches to a certain threshold then diabetes occurs. In HCS (Healthcare Services) the Machine learning has gained a signification position because of its improving ability of disease prediction in healthcare services. The tedious identifying technique needs the patient should consult a doctor and visit a diagnostic center. This critical problem was solved by the rise of machine learning techniques. Hence there is a requirement for unified framework designing & diabetes prediction implementing by machine learning is proposed here. The three basic classifications of machine learning algorithms are Naive Bayes, Decision Tree and support vector machine (SVM) used in this proposed system for detecting the diabetics at the early stage. From the sources of UCI (University of California) machine learning respiratory the experiments are performed on PIDD (Pima Indians Diabetes Database). The three algorithms performance is evaluated by different parameters such as F-Measure, Accuracy, Recall and Precision measured through classified instances incorrectly or correctly. From the obtained results the Naïve Bayes exhibits the highest accuracy of 78% compared to remaining algorithms

KEYWORDS: diabetes, machine learning, Decision Tree, SVM (support vector machine), Naive Bayes.

DOI Number: 10.14704/nq.2022.20.13.NQ88240 Neuro Quantology 2022; 20(13):1940-1946

I.INTRODUCTION

The disease which can affect the ability of producing the insulin hormone in the human body is the diabetes, which in turn makes abnormal carbohydrate metabolism and raises the glucose levels in blood [1]. Basically, the diabetic person suffers from high blood sugar. Typical health signs include elevated hunger, repeated urination, and excessive blood pressure. A form of Diabetes cannot be treated successfully with oral medications alone, so patients require insulin therapy [2]. If the diabetes is untreated then more complications may appear. Few of the critical complications include nonketotic hyperosmolar coma & ketoacidosis [3]. When the sugar substance measurement is not controllable then the diabetes can be treated as serious health issue. Diabetes is affected by different factors such as hereditary factor, weight, height & insulin, among all of this insulin is considered as the main reason for sugar concentration. For staying away from severe complications only remedy is early identification [4].

As per the report of WHO (World Health Organization) the diabetes is the 7th prominent cause for the death in 2030 [2]. By 2040 the 642 million adults i.e. 1 in 10 adults will be projected to diabetes [3]. In 2015 the deaths of 1.6 million people are affected by diabetes completely. Due to high glucose levels of blood 2.2 million deaths happened in 2014 [5]. The diabetes is independent of age so it may happen to the people at any age. The diabetes is classified in to 3 types, they are type 1 diabetes, type 2 diabetes & type 3 diabetes. A disease of autoimmune is the type 1 diabetes. During this condition the cells which are essential for producing the insulin for absorbing the sugar for producing energy are destroyed by the body. Obesity is caused by type 1 diabetes. Due to the obesity the BMI (body mass index) is increased than the normal BMI level of a person [6]. This type of diabetes can be occurring at adolescence age or childhood [7]. The adults who are having obese are usually affected by type 2 diabetes [8]. Due to this the human body failed to produce the insulin or observing insulin is resisted. This type of diabetes can occurs at aged people or middle aged groups. Type 3 or Gestational diabetes means hyperglycemia which can occur during pregnancy due to the hormones change [9]. Viral or bacterial infection, bad diet, obesity, chemical or toxic contents in food, change in lifestyles, reaction of autoimmune, pollution in environment, eating habits, etc are the reasons for diabetes. Various diseases like renal issues, foot ulcers, cardiovascular complications and retinopathy are caused by diabetes [10].

Over the years large amount of EHRS (Electronic Health Records) are collected for providing a base for the prediction & risk analysis [11]. With the massive knowledge on analytics technology, detection of disease has been paid more attention in the big data analysis perspective, variant researches has been conducted by the selection of characteristics mechanically from the variety of outsized information for the risk classification accuracy instead of characteristics chosen in past. But the existing system mainly considered structured data. For such cases the semantics may differ from healthy person to severely ill person or it cannot predict the disease in particular. For healthy cases there is no ground truth available. The alive set of cases are simply treated as negative class and it can be a highly noise majority class. On the other hand this large alive set unlabeled genuinely, as opposes to cases with known labels are removed, it would became a multi-class learning problem having large unlabeled data. Most of the existing methods of health care data don't consider the multimodal disease data prediction issues.

II. MACHINE LEARNING BASED PREDICTIVE MODELS

Ensemble methods are statistical and computational learning procedures. They are in sync with human social learning and trying different opinions before making any final decision. Set of learning machines are used to combine choices and provide more robust and accurate predictions on controlled and unattended learning problems. There is no single, theoretically sound explanation for classifier ensemble methods [12]. The Machine learning approach suggested and changing the SVM rules for prediction. Comparison analysis between Naïve Bayes, Decision Tree and K-NN algorithms has been performed [13].

The algorithms of ML are classified into 3 categories; they are reinforcement learning, unsupervised learning and supervised learning. For accuracy testing the supervised learning algorithm is used instead of other machine learning algorithms. From the pre existing data the supervised learning algorithm learns the patterns and based on previous learning try to predict the new results. For identifying the existing data such as, tree – based, rule – based, probability – based, instance – based and function – based etc. the ML algorithms are used. Using various algorithms of data mining the different ML algorithms are introduced to assist the medical experts. By the accuracy of the decision support system the effectiveness is recognized. For predicting & diagnose a specific disease with high degree of accuracy is the main goal of decision support system building.

An open source data set which is used to evaluate and train the proposed system uses preexisting data sets called PIDD [14]. So many researches has been conducting experiments to diagnose the disease using different classifications of ML

algorithms approaches are Naive Bayes, Decision Table, SVM, Decision Tree, J48 etc. the researchers have proved these ML algorithms work better in diagnosing the various diseases.

The classifiers should be used for diabetes prediction, and they are recommended to improve them through the production of hybrid models. In short, the deep learning studies the features corresponded to outcomes of diabetes are compared with traditional ML models for extracting the useful data from EHR. The key issues with a filter - based collection of features were (i) the bulk of the features do not accept consistency, (ii) the limitation of a particular filter-based system in the chosen function subset and (iii) poor prediction accuracy during classification [15].

In this presented system the classification of ML algorithms: SVM, Decision Tree, Naive Bayes are utilized for evaluating over the data set PIDD to find the diabetes prediction for a patient. By using different measures, all these three algorithms experimental performance are compared and these algorithms achieve good accuracy. The merits of the Naïve Bayes algorithm being an intuitive technique, there was no need to set values of parameters before proceeding. The probabilities returned by this algorithm can be easily applied to further experiments or analyses. Its learning rate is fast, and classification starts with few datasets. It is also computationally fast when making decisions.

III. DIABETES PREDICTION USING MACHINE LEARNING

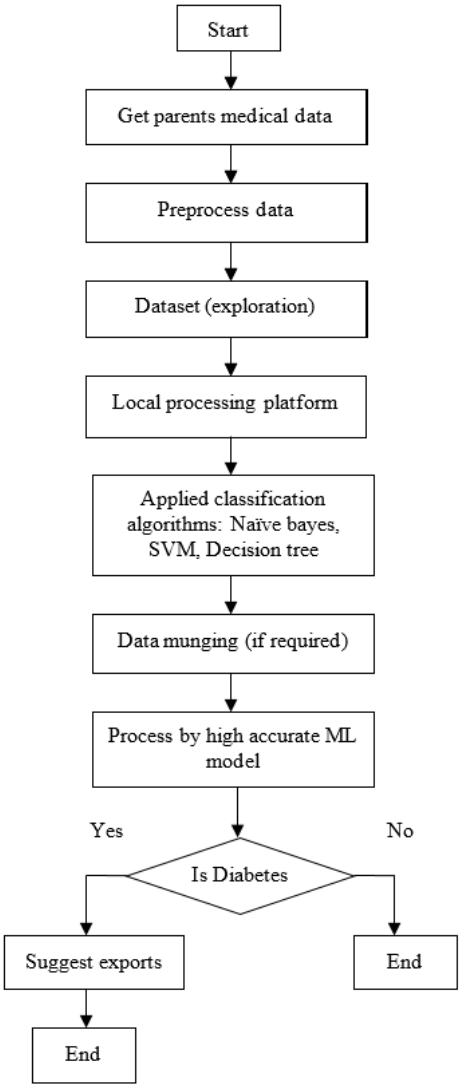

Fig. 1: FRAMEWORK OF DIABETES PREDICTION USING MACHINE LEARNING

A unified framework of implementation of predicting disease diabetes using ML is represented in the above Fig. 1. The methodology or working process is described below.

3.1 Dataset:

In this proposed model the patients needs to give their medical data for their successful diagnosis diabetes test. With the help of PIDD this proposed system can predict the diabetes of individual. The PIDD consists of a 768 instances having 8 attributes. Here if one class value is 1 treated as tested positive diabetes & the class value 0 is

considered as tested diabetes negative. Dataset includes glucose concentration found in glucose level (oral glucose tolerance test), number of times pregnant, thickness of skin, BMI, age, diabetes pedigree functions and blood pressure.

3.2 Preprocess data:

Higher prediction accuracy is produced by preprocessing of PIDD. Thus, it can also calculate the proposed system accuracy, prediction accuracy is also improved.

3.3 Local processing platform:

In python environment dataset is explored with the involved attributes of data dictionary. Local processing platform (LPP) is able to process collected data. The data is processed once by a Local processing platform, it will be sent to the machine learning classification stage.

3.4 Machine learning algorithms:

For finding the suitable algorithm of machine learning which has capability to predict the diabetes more accurately, the power full ML models like Decision tree, SVM, Naïve Bayes are tested.

3.4.1 Naive Bayes Classifier:

A classification model with a notion and denotes all features as unrelated to each other & independent is called as Naïve Bayes. While denoting a specific feature status of one class doesn't affect another status feature. For the classification purpose Naive Bayes is taken as powerful algorithm which is based on the conditional probability. For missing values and data unbalancing problems case the Naïve Bayes work well. The ML classifier works on theorem of Bayes. The posterior probability P(C/X) is calculated by using Bayes theorem, calculated from P(X), P(X|C) & P(C).

$$P\left(\frac{C}{X}\right) = \frac{P\left(\frac{X}{C}\right)P(C)}{P(X)}$$

Where,

P (C|X) = posterior probability of target class's

P (X|C) = probability of predictor class's

P (C) = probability being true of class C's

P (X) = prior probability of predictor's

3.4.2 SVM (Support Vector Machine):

The supervised learning model standard set is the support vector machine. Two-class training sample is given to SVM; the main objective of this is to find out the separation of hyperplane between two classes with highest - margin. For better hyperplane generalization the plane should not closer to data points which belong to other class and the hyperplane must be chosen far from the data points from each type. The points which lie closer to the margin are the support vectors.

3.4.3 Decision Tree Classifier:

For solving the classification problems a supervised ML algorithm of Decision tree is used. The main aim of utilizing the decision tree in this proposed model is to predict the target class taken from the prior data using decision rule. For the classification & prediction decision tree uses internodes & nodes. The classifications of instances with various features are done by root nodes. If the leaf nodes represent the classification, the root nodes may have two or more branches. In each stage, DT selects every node by the evaluation of highest information gain among the attributes.

3.5 Data Munging:

Data munging means the estimation of missing data values in few variables & it is needed because many interpretations will not be performed when data is missing. If it is continuous variable then the missing value is replaced by mean value, in categorical variable the missing data is replaced by mode value.

3.6 Evaluation of accurate ML:

The performance of proposed diabetes disease prediction model is calculated by using parameters performance as precision, accuracy,

F-measure & recall. If the accuracy is high then the proposed method was efficiently detects the diabetes patients. Then the diabetes patients are suggested to experts for further medication. If not, the test sampled data of patient is declared as non-diabetes.

IV. EXPERIMENTAL RESULTS

All the classified algorithms performance is assessed with various aspects of statistical measurement like precision, F1-score, accuracy & recall. The measurement factors of these classified algorithms are calculated by the below terms: FP (False Positive), FN (False Negative), TP (True Positive) & TN (True Negative). Here

FP: Results of the prediction are yes and the person doesn't have diabetes actually (type 1 error)

FN: results of the prediction are no but the patient had diabetes (type 2 error).

TP: Results of prediction are yes & the patient has diabetes.

TN: Results of prediction are no & the patient doesn't have diabetes.

For measuring the factors as computational formulae is given below.

Accuracy:

Accuracy is defined as the ratio of predictions made by model correctly to all types of completed predictions in the problem classification.

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)}$$

Precision:

Positive precision or prediction is defined as the ratio of number of accurate positive score to the number of positive scores predicted by the classification algorithm.

$$Precision = \frac{TP}{(TP + FP)}$$

Recall:

Sensitivity or recall or TP rate is defined as a measure which is the ratio of actual positive instances having diabetes to the positive instances actually (both TP & FN).

$$Recall = \frac{TP}{(TP + FN)}$$

F1-score:

Weighted average of precision & recall is the measure of F1. F1 must be 1 for good performance of classification algorithm, 0 for bad performance.

$$F1 - Score = 2 \times \frac{Precision * Recall}{Precision + Recall}$$

Different classifiers performance parameters comparisons are described in below Table 1:

Table 1: DIFFERENT CLASSIFIERS PERFORMANCE PARAMETERS COMPARISON

Parameters	Naïve bayes	SVM	Decision tree
Accuracy	78	66	74
Precision	75	45	72
Recall	77	66	74
F1-score	78	54	75

The graphical representation of accuracy and precision is represented in below Fig. 2.

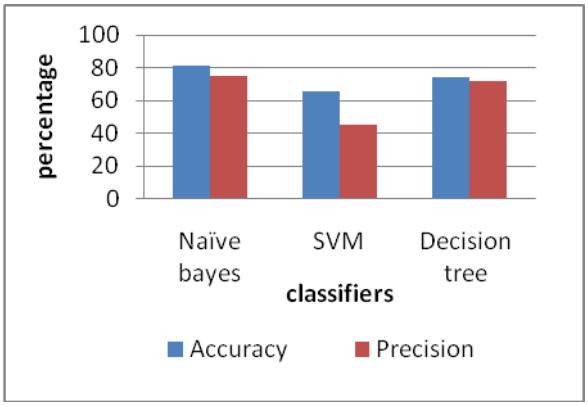

Fig. 2: COMPARISON BETWEEN DIFFERENT CLASSIFIERS ACCURACY AND PRECISION PARAMETERS

The graphical representation of Recall and F1-score are represented in below Fig. 3.

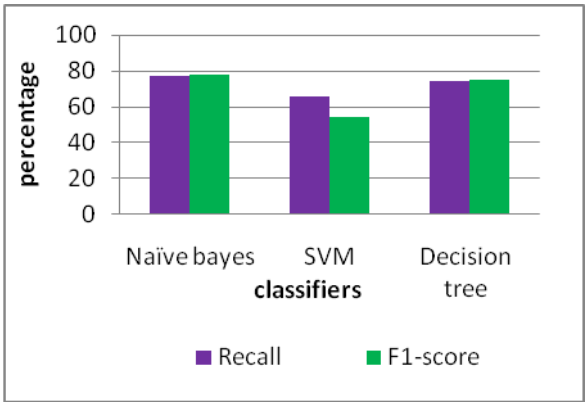

Fig. 3: COMPARISON BETWEEN DIFFERENT CLASSIFIERS RECALL AND F1-SCORE PARAMETERS

Table 2 determines performance of classifiers based on the classification of instances. In accordance with the classified instances the accuracy was analyzed & calculated. By using the incorrectly classified instances & correctly classified instances the individual algorithm performance is evaluated. The total numbers of instances are 768.

Table 2: CLASSIFIER PERFORMANCE BASED ON CLASSIFER INSTANCES

Classifiers	Correctly classified instances	In-correctly classified instances
Naïve bayes	590	178
SVM	505	263
Decision tree	574	194

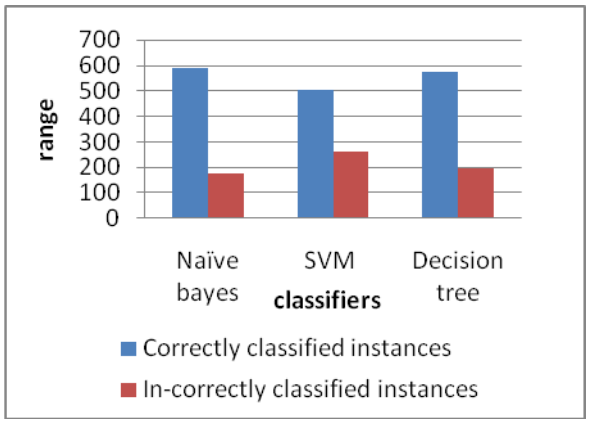

Fig. 4: CLASSIFIED INSTANCES

All the classification of algorithms is based on the classification of instances is shown in Fig. 4. From the Tables 1 & 2 it is clear that the Naïve Bayes algorithm performs greatly compared to other classified algorithms.

V. CONCLUSION

A unified framework for implementation of predicting diabetes disease using ML is proposed in this paper. This gives fast and accurate diabetes predictions. Based on PIDD the experiments are performed. The presented system used 768 instances in 8 attributes. In order to remove unwanted data and to speed up processing time, the used data are preprocessed. During this work, 3 ML algorithms classified as Naive Bayes, SVM, and Decision Tree are evaluated & studied over different measures such as accuracy, precision, recall and F1-score. Highest accuracy is obtained at naïve bayes classifier as 78%. In this paper diabetes disease prediction using ML algorithm is presented & developed. The presented system plays a

significant role in the medical & science fields to detect the variant medical data with high degree of accuracy. In future, this presented system with ML classified algorithms can be utilized for predicting the diabetes or other diseases.

VI. REFERENCES

[1] Tuan Minh Le, Thanh Minh Vo, Tan Nhat Pham, Son Vu Truong Dao, "A Novel Wrapper–Based Feature Selection for Early Diabetes Prediction Enhanced With a Metaheuristic", IEEE Access, Volume: 9, Year: 2021

[2] Kirill V. Pozhar, Nikolay A. Bazaev, Evgeniia L. Litinskaia, "In Silico Testing of a Control Algorithm for a Personalized Insulin Therapy System", 2021 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (ElConRus), Year: 2021

[3] Jorge Javier Mendoza Montoya, Germán Torregrosa Penalva, Ernesto Àvila Navarro, José Chilo, "Monitoring Diabetic Ketoacidosis by Urine Ketones Tracing Using an E-Nose", 2020 IEEE International Instrumentation and Measurement Technology Conference (I2MTC), Year: 2020

[4] Sari Ayu Wulandari, Ratih Pramitasari, Sutikno Madnasri, Susilo, "Electronic Noses for Diabetes Mellitus Detection: A Review", 2020 International Seminar on Application for Technology of Information and Communication (iSemantic), Year: 2020

[5] Ivan Volkov, Gleb Radchenko, "DiaMeter: a Mobile Application and Web Service for Monitoring Diabetes Mellitus", 2020 Ural Symposium on Biomedical Engineering, Radioelectronics and Information Technology (USBEREIT), Year: 2020

[6] Chukwuemeka Obasi, Ijeoma Ndu, Ogechukwu Iloanusi, "A Framework for Internet of Things-Based Body Mass Index Estimation and Obesity Prediction", 2020 International Conference on e-Health and Bioengineering (EHB), Year: 2020

[7] Vadim V. Klimontov, Julia F. Semenova, Alla K. Vigel, "Glucose variability in subjects with type 1 diabetes: the relationships with non-enzymatic glycation, albuminuria and renal function", 2020 Cognitive Sciences, Genomics and Bioinformatics (CSGB), Year: 2020

[8] Prabhu S. Madhava, Seema Verma, "A Systematic Literature Review for Early Detection of Type II Diabetes", 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS), Year: 2019

[9] Evgenii Pustozerov, Polina Popova, "Mobile-based decision support system for gestational diabetes mellitus", 2018 Ural Symposium on Biomedical Engineering, Radioelectronics and Information Technology (USBEREIT), Year: 2018

[10] Karkhanis Apurva Anant, Tushar Ghorpade, Vimla Jethani, "Diabetic retinopathy detection through image mining for type 2 diabetes", 2017 International Conference on Computer Communication and Informatics (ICCCI), Year: 2017

[11] Yu Wang, Peng-Fei Li, Yu Tian, Jing-Jing Ren, Jing-Song Li, "A Shared Decision-Making System for Diabetes Medication Choice Utilizing Electronic Health Record Data", IEEE Journal of Biomedical and Health Informatics, Volume: 21, Issue: 5, Year: 2017

[12] K. Sumangali, B. S. R. Geetika, Harshitha Ambarkar, "A classifier based approach for early detection of diabetes mellitus", 2016 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT), Year: 2016

[13] V Veena Vijayan, C Anjali, "Decision support systems for predicting diabetes mellitus — A Review", 2015 Global Conference on Communication Technologies (GCCT), Year: 2015

[14] Savvas Karatsiolis, Christos N. Schizas, "Region based Support Vector Machine algorithm for medical diagnosis on Pima Indian Diabetes dataset", 2012 IEEE 12th International Conference on Bioinformatics & Bioengineering (BIBE), Year: 2012

[15] Hasan Akan, Uğur Demirok, Niyazi Kılıç, "The effect of feature reduction techniques on diagnosis of diabetes", 2012 20th Signal Processing and Communications Applications Conference (SIU), Year: 2012

eISSN 1303-5150