IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 8, Issue 8, August 2020

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.488

9940 572 462 | 6381 907 438 | ijircce@gmail.com | www.ijircce.com

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | Impact Factor: 7.488 |

|| Volume 8, Issue 8, August 2020 ||

Comparative Study of Supervised Learning Algorithms for Student Performance Prediction

S. B. Archana[1], S. Lalitha[2], K. Gobinathan [3] , R. Umamaheswari [4]

[1]PG Scholar, Department of Computer Science and Engineering, Gnanamani College of Technology, Namakkal, Tamilnadu, India

[2]Assistant Professor, Department of Computer Science and Engineering, Gnanamani College of Technology, Namakkal, Tamilnadu, India

[3]Assistant Professor, Department of Computer Science and Engineering, Gnanamani College of Technology, Namakkal, Tamilnadu, India

[4]Professor and Head, Department of Computer Science and Engineering, Gnanamani College of Technology, Namakkal, Tamilnadu, India

ABSTRACT: Machine learning is an area of computer science in which the computer predicts the next task to perform by analyzing the data provided to it. The data accessed by the computer can be in the form of digitized training sets or via interaction with the environment. The algorithms of machine learning are constructed in such a way as to learn and make predictions from the data unlike the static programming algorithms that need explicit human instruction. There have been different supervised and unsupervised techniques proposed in order to solve problems, such as, Rule-based techniques, Logic-based techniques, Instance-based techniques, stochastic techniques. We evaluate the effectiveness of various supervised learning algorithms to predict if a bug report would be reopened. State-of-the-art classical supervised learning algorithm in machine learning literature. Diabetes Mellitus is one of the most serious health challenges affecting children, adolescents and young adults in both developing and developed countries. To predict hidden patterns of diseases diagnostic in the healthcare sector, nowadays we use various data mining techniques. We also discuss various applications of machine learning in the field of medicine focusing on the prediction of diabetes through machine learning. Diabetes is one of the most increasing diseases in the world and it requires continuous monitoring. To check this we explore various machine learning algorithms which will help in early prediction of this disease.

KEYWORDS: Data Mining, Naive Bayes, Machine learning, rule-based techniques, Diabetes, health care, decision tree, supervised learning algorithms.

I.INTRODUCTION

Diabetes Mellitus is one of the most widespread chronic diseases of childhood, affecting children, adolescents and young adults. Diabetes is a condition in which a human body is unable to produce the required amount of insulin needed to regulate the amount of sugar in the body [1]. This Phenomenon leads to various diseases including cardiovascular diseases, blindness, kidney failure, and lower limb amputation. Maintaining blood glucose levels, blood pressure and cholesterol at normal range can help delay or prevent diabetes complications arising due to diabetes. Therefore diabetic patients need regular monitoring. In 2013, 382 million people worldwide were reported with diabetes: - 24 million in South and Central America, 37 million in North America and the Caribbean, 56 million in Europe, 35 million in the Middle east and North Africa, 20 million in Africa and 138 million in Western Pacific and it is predicted that 592 million people worldwide will live with diabetes in 2035 [2]. As the prevalence of diabetes continues to grow worldwide, related diseases like morbidity and mortality are emerging as a major health care problem. Patients with diabetes are at increased risk of developing and dying from cardiovascular diseases (CVD). People with diabetes have been shown to have twice the risk of CVD as the general population [2]. Moreover, Coronary Heart Disease (CHD) is the leading cause of death among adults with diabetes [3]. Diabetes disease diagnosis and interpretation of data is an important classification problem [4].

A classifier is required to be designed in an efficient way, cost effective and accurate manner. A medical diagnosis is a classification process where a physician has to analyze a lot of factors before diagnosing a patient with diabetes, which is generally, seems to be a difficult problem. In the past, various statistical methods have been used for modeling in the area of diseases diagnosis. However most of the methods requires prior assumptions and are less capable of dealing with

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | Impact Factor: 7.488 |

|| Volume 8, Issue 8, August 2020 ||

massive and complicated nonlinear and dependent data [5]. Data mining has been proven to be more powerful and effective approach which provides process for discovering useful patterns from large data sets [6]. Recently, many methods and algorithms including Neural networks (NNs), Decision Trees (DTs), Fuzzy Logic Systems, Naive Bayes, SVM, Categorization, Logistic Regression are proposed by different researchers to mine biomedical datasets for identification of hidden patterns embedded within them [7,8,9]. These algorithms minimize time for the medical treatment, providing safe health care treatment and providing various healthcare treatments based on patients' needs, symptoms and preferences. A brief description of the general symptoms of diabetes is given in Table 1. Machine learning empowers system with the ability to learn automatically and get better with experience without being explicitly programmed. The algorithms of machine learning are useful in areas where deploying explicitly written algorithms with high speed performance are unfeasible.

A simple task such as sorting of numbers is easy and can be performed by giving some numbers as input and getting an ordered list as an output. Here we know what to give as input and what procedure has to be followed to obtain the desired output. But certain tasks are not easy to comprehend such as filtering of emails to differentiate between legitimate emails and spam mails. Here we know the input to be provided and the output is in the form of true or false, but the instructions that need to be given to the program to perform these actions are not clear. Such unique situations where there is no specific algorithm to achieve success, we take the help of data and instruct the machine to analyze the data and make an intelligent sense of this data

[1]. Few Applications of Machine Learning:
[2] Classification of texts or documents. e.g: Filtering spam messages
[3]. Speech recognition
[4]. Computer vision tasks such as image recognition and face detection
[5]. Self-driving vehicles
[6]. Web page ranking like for search purposes
[7]. Collaborative filtering
[8]. Medical diagnosis
[9]. Computation biology application
[10]. Recommendation systems, search engines, information extraction systems

Multiple opportunities for healthcare are created because machine learning models have potential for advanced predictive analytics. There are already existing models in machine learning which can predict the chronic illness like heart disorder, infections and intestinal diseases. There are also few upcoming models of machine learning to predict non-communicable diseases, which is adding more and more benefit to the field of healthcare. Researchers are working on machine learning models that will offer very early prediction of specific disease in a patient which will produce effective methods for the prevention of the diseases. This will also reduce the hospitalization of patients. This transformation will be very much beneficial to the healthcare organisations.

The most explored area is the healthcare system which uses modern computing techniques is in healthcare research. As mentioned above the researchers in the related fields are already working with the healthcare organization to come up with more technology ready systems. Diabetes is a disease which reduces the body's capability to produce insulin. In other words the body can not retaliate to the hormone insulin production. This results in anomalous metabolism of carbohydrates and increased blood glucose levels. Early detection of diabetes becomes very important because of the reasons mentioned above. Many people in the world are getting affected by diabetes and this number is increasing day by day. This disease can damage many vital organs hence the early detection will help the medical organisation in treatment of it. As the number of diabetic patients is more there is an excessive important medical information which has to be maintained. With the support of increasing technology the researchers have to build a structure that store, maintain and examine these diabetic information and further see feasible dangers.

The blood glucose levels become too high in the body when there is diabetes. Glucose is created in the body after eating food. The hormone insulin produced in the body helps balance the glucose levels and regulate blood sugar levels, deficiency of insulin causes Diabetes. Type 1 diabetes is a scenario where the body does not produce insulin at all to balance the sugar levels in blood. Type 2 is a diabetes type where the body produces insulin but does not utilize this hormone completely to balance blood sugar levels. If a woman gets this disease during pregnancy then it is known as gestational diabetes. By managing our weight, meal plan and exercise we can control diabetes. One should always keep a check on its blood sugar levels.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | Impact Factor: 7.488 |

|| Volume 8, Issue 8, August 2020 ||

2. LEARNING STRATEGIES

Machine learning employs the following strategies

1. Supervised learning:

A. Regression:

i. Linear Regression:

In simplest terms we can say that in linear regression we add the inputs multiplied by some constants to obtain the output. It creates a correlation between Y, a dependent variable, and X, which can be multiple independent variables, using a straight line (regression line).

The general equation can be written as -

Y = a + bX, Where,

Y – Dependent Variable,

X – Explanatory Variable

ii. Support Vector Machine Regression:

If we are given a particular training set, say {(x1, y1),….,(xi, yi)} _ X × R, here X -> space of input patterns. Our goal in SV regression is to search for a fitting function f(x), having deviation less than _ from the target (yi) acquired for the relating training data set. The function should be reasonably flat. Or it can be said that any error less than _ is fine.

The linear function (f)

f (x) = (w, x) + b with, w _ X , b _ R

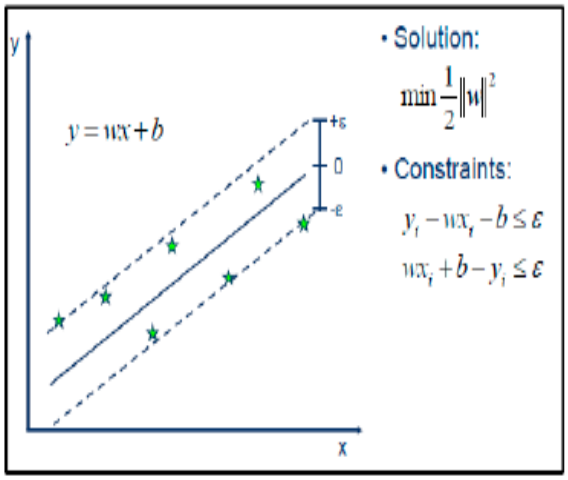


Figure 1: Support vector machine (SVM)

iii. Decision Tree Regression:

This regression mechanism works by breaking down a dataset in smaller sub datasets and subsequently related decision tree is developed in an incremental way. Finally a tree having decision nodes and leaf nodes is obtained. The tree has a root node which is the topmost decision node corresponding to the best predictor .

The ID3 (Iterative Dichotomiser 3) is the basic algorithm used to build the decision tree. The ID3 algorithm uses Standard Deviation Reduction (SDR) to construct the decision tree. Steps involved in SDR:

i. First, we calculate the standard deviation of the target.

ii. After this, we split the datasets on the different attributes.

The resulting standard deviation for each branch is than subtracted from the standard deviation before the split. This is SDR.

SDR(T,X) = S(T) - S(T,X)

iii. The attribute having the largest SDR as the decision node is to be selected.

iv. Dataset has to be divided based on the values of selected attributes. We further split a branch set if the standard deviation is greater than 0.

v. The process keeps on running in recursion until all the data is processed.

iv. Random Forest Regression:

It is an extra cover of randomness to bagging. Unlike a normal tree, random forest splits each node using the best among a subset of predictors randomly chosen at that node. It is an easy to operate technique because it has very few parameters – the number of trees in the forest and variables at each node in the random subset.

The algorithm steps of RF Regression:

i. Using the actual data, take nt bootstrap sample.

ii. For every bootstrap sample, a regression tree has to be grown with some alterations: sample mt of the predictors randomly at each node and choose the best split amongst the variables.

iii. Estimate the latest data by cumulating the predictions of the nt trees (average for regression).

v. LASSO Regression:

LASSO is short for 'least absolute shrinkage and selection operator'. By shrinking some coefficients and setting others coefficients to zero, it keeps the quality features of subset selection and also ridge regression. LASSO regression has the ability to reduce the variability and improve the precision of linear regression models and also it penalizes the entire size of regression coefficients .

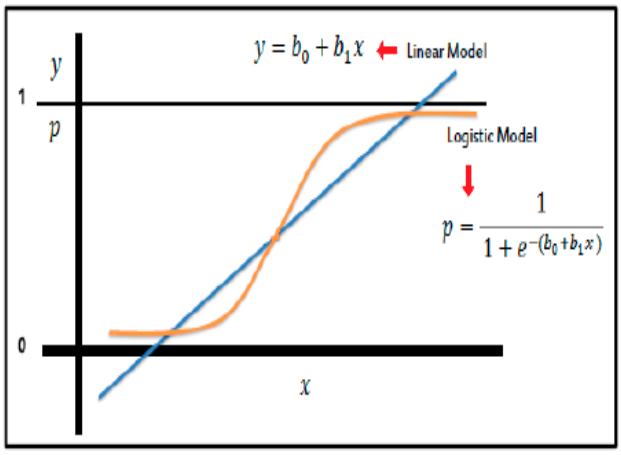
$$\sum_{i=1}^{N}\left\{y_i-\sum_{j=0}^{M}w_j x_{ij}\right\}^2+\lambda\sum_{j=0}^{M}|w_j|$$

This regression uses absolute values in the penalty function, instead of squares which makes some of the parameter predictions to be precisely zero.

B. Classification:

i. Logistic regression:

It is a reliable procedure to solve binary classification problem. Logistic regression is used to predict the probability of an outcome having only two values. The core of logistic regression is the logistic function - a S-shaped curve taking any real-valued number and mapping that number in a value between 0 and 1, though never precisely at 0 and 1.



Euclidean - $\sqrt{\sum_{i=1}^{k}(x_i - y_i)}$

Manhattan - $\sum_{i=1}^{k}|x_i - y_i|$

Minkowski - $\left(\sum_{i=1}^{k}(|x_i - y_i|)^q\right)^{1/q}$

Figure 2: The steepness of the curve.

ii. K Nearest Neighbors:

This method is known for its simplicity because of the factors such as the ease of interpreting and the low calculation time. It basically stores the cases that are available and categorizes new cases based on the homogeneity basis such as distance function. The object is categorized by a majority vote of its neighbors and the result is usually class

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | Impact Factor: 7.488 |

|| Volume 8, Issue 8, August 2020 ||

integration. After this, object is allotted to a class which has the greatest similarity amongst the K nearest neighbors. Some functions for distance are,

In the case of categorical variables Hamming distance must be used.

iii. Naïve Bayesian:

It is based on the probabilistic model of Bayes theorem, and easy to set up as complex iterative parameter estimation is almost none, making it viable to use for large sets of data [20]. Given the class variables, the value of a certain characteristic is assumed to be independent of the value of any other characteristic by the naïve Bayesian classifiers. We can calculate the Posterior probability P (A|B)

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

P (A|B) - Posterior probability of the class when predictor is given (attribute),
P (A) - Prior probability of the class,
P (B|A) - Probability of the predictor when class is given,
P (B) - Prior probability of the predictor

iv. Decision Tree (DT) Classification:

In the decision tree classification, the ID3 algorithm uses Entropy and Information Gain to construct a decision tree instead of Standard Deviation Reduction method. Entropy is used to calculate the homogeneity of the sample. For entropy to be zero, the sample has to be totally homogeneous and this happens if the sample is divided in equal parts.

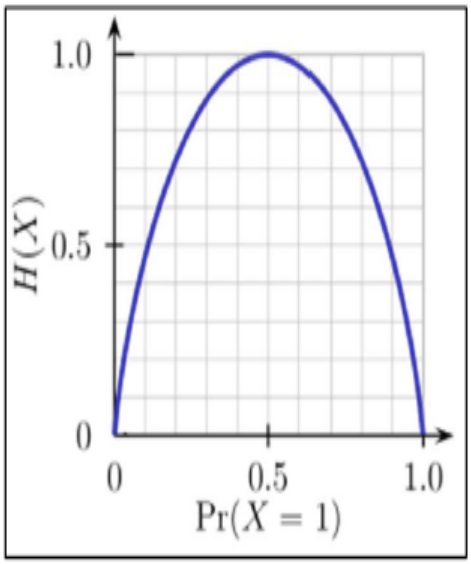


Figure 3: Entropy of a decision tree

Algorithm for Information Gain:

i. Calculate Entropy of the target.
ii. We split the dataset based on different attributes and then calculate the entropy for each branch. After this we add it proportionally which gives us total entropy after the split. The total entropy after the split is subtracted from the total entropy before the split and the result is Information Gain.
iii. Attribute with largest Information Gain is the decision node.
iv. A branch with entropy 0 is a leaf node.
v. A branch with entropy greater than 0 needs further splitting.

III.METHODOLOGY

In this section we shall learn about the various classifiers used in machine learning to predict diabetes. We shall also explain our proposed methodology to improve the accuracy. In section A we shall explain various classifiers and in section B we shall explain our proposed system.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | Impact Factor: 7.488 |

|| Volume 8, Issue 8, August 2020 ||

A. Machine learning classifiers used in diagnosis of diabetes

The variation in glucose levels is cause of diabetes. Insulin balances the blood glucose level in the body, deficiency of which cause diabetes. For the prediction of diabetes machine learning is used, these have many steps like image pre-processing/data preprocessing followed by a feature extraction and then classification. We can use any of the mentioned machine learning classifiers to predict this disease. In the above section we have learning about many classification algorithms, we can either use any one of these to predict the disease or we can explore the techniques to use the hybrid methodology to improve the accuracy over using a single one. Currently, the researchers have used the a single classification algorithm and have come up to accuracy of 70 to 80% for detection of the diabetes disease. Depending on the application and nature of the dataset used we can use any classification algorithms mentioned below. As there are different applications, we can not differentiate which of the algorithms are superior or not. Each of classifiers have its own way of working and classification.

IV.PROPOSED SYSTEM

The proposed system predicts the disease of diabetes in patients with maximum accuracy. We shall talk about various machine learning, the algorithm which can help in decision making and prediction. We shall use more than one algorithm to get better accuracy of prediction.

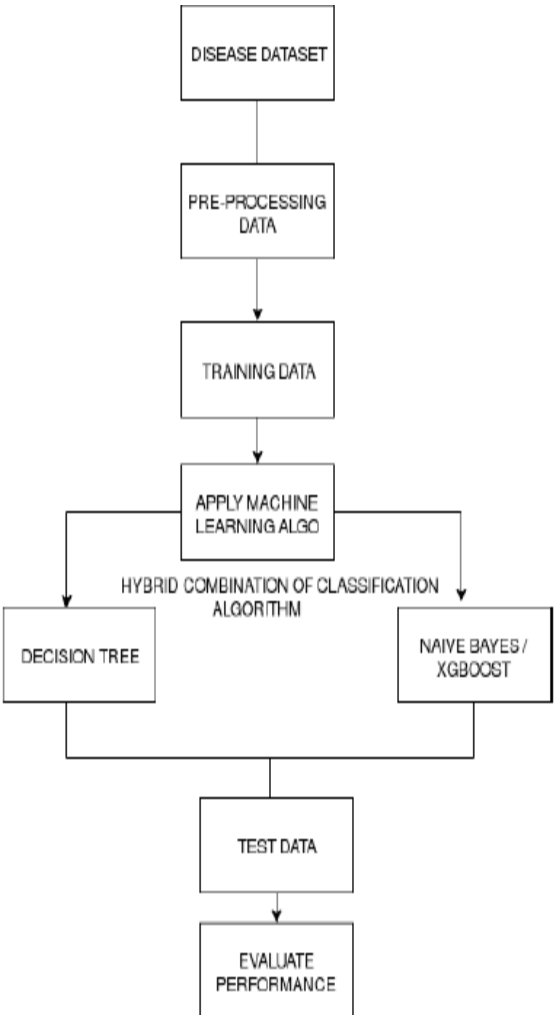


Fig 1. Proposed System Block diagram

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | Impact Factor: 7.488 |

|| Volume 8, Issue 8, August 2020 ||

V.CONCLUSIONS

This paper discusses the commonly used supervised algorithms. The primary goal was to prepare a comprehensive review of the key ideas and present different techniques for every supervised learning method. The paper makes it clear that every algorithm differs according to area of application and no algorithm is more powerful than the other in different scenarios. The choice of an algorithm should be made depending on the type of problem given to us and the data available. The accuracy can be increased by using two or more algorithm together in suitable conditions. The machine learning methods can support the doctors to identify and cure diabetic diseases. We shall conclude that the improvement in classification accuracy helps to make the machine learning models get better results. The performance analysis is in terms of accuracy rate among all the classification techniques such as decision tree, logistic regression, k-nearest neighbors, naive bayes, and SVM , random forest , adaboost , xgboost. We have also seen that the accuracy of the existing system is less than 70% hence we proposed to use a combination of classifiers known as Hybrid Approach. Hybrid approach takes advantage by aggregating the merits of two or more techniques.

REFERENCES

[1] Domingos, P. “A few useful things to know about machine learning”, Communications of the ACM, 55(10),2012 pp.1.
[2] Mohri, M., Rostamizadeh, A. and Talwalker, A. “Foundations of machine learning”, Cambridge, MA: MIT Press,2012.
[3] Nguyen, T. and Shirai, K. “Text Classification of Technical Papers Based on Text Segmentation”, Natural Language Processing and Information Systems, 2013,pp.278-284.
[4] Deng, L. and Li, X. “Machine Learning Paradigms for Speech Recognition: An Overview”, IEEE Transactions on Audio, Speech, and Language Processing, 21(5), 2013, pp.1060-1089.
[5] Siswanto, A., Nugroho, A. and Galinium, M. “Implementation of face recognition algorithm for biometrics based time attendance system”, 2014 International Conference on ICT For Smart Society (ICISS).
[6] Chen, Z. and Huang, X. “End-to-end learning for lane keeping of self-driving cars”, 2017 IEEE Intelligent Vehicles Symposium (IV).
[7] Yong, S., Hagenbuchner, M. and Tsoi, A. “Ranking Web Pages Using Machine Learning Approaches”, 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology.
[8] Wei, Z., Qu, L., Jia, D., Zhou, W. and Kang, M. “Research on the collaborative filtering recommendation algorithm in ubiquitous computing”, 2010 8th World Congress on Intelligent Control and Automation.
[9] Kononenko, I. “Machine learning for medical diagnosis: history, state of the art and perspective”, Artificial Intelligence in Medicine, 23(1), 2011, pp.89-109.
[10] Jordan, M. “Statistical Machine Learning and Computational Biology”,IEEE International Conference on Bioinformatics and Biomedicine (BIBM 2007).
[11] Thangavel, S., Bkaratki, P. and Sankar, A. “Student placement analyzer: A recommendation system using machine learning”, 4th International Conference on Advanced Computing and Communication Systems (ICACCS-2017).

INNO SPACE

SJIF Scientific Journal Impact Factor

Impact Factor:
7.488

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details