www.revistageintec.net
ISSN: 2237-0722
revista GEINTEC
Gestão, Inovação e Tecnologias

Early Prognosis of Diabetes Using Supervised Learning Techniques: A Comparison of Performance

H.S. Niranjana Murthy[1]

[1]Associate Professor, Dept. of Electronics & Instrumentation Engineering, Ramaiah Institute of Technology, Bangalore.

[1]hasnimurthy@msrit.edu

Abstract

Supervised machine learning algorithms have been a predominant technique in information mining field. Disease forecast utilizing health information has shown a potential application region for these techniques. This investigation centers to distinguish the vital patterns among various kinds of supervised learning and their performance and utilization for diabetes prognosis. The point of this work is to analyze the performance of various machine learning (ML) classifiers. These ML classifiers are k-Nearest Neighbors, Support Vector Machine, ANN, Logistic Regression, Decision Tree and Ensemble classifiers which are applied on diabetes datasets for evaluating performance. The ML models are trained and tested with Pima Indian Diabetes dataset. The exploratory outcomes uncover that the SVM classifier beat the different classifiers with a most noteworthy accuracy of 83% in detecting diabetes.

Key-words: Machine Learning, SVM, KNN, ANN, Ensemble Classifier.

1. Introduction

Diabetes is a regular steady disorder and addresses an uncommon peril to human wellbeing. The property of diabetes is that the blood glucose is higher than the typical level, which is accomplished by lacking insulin surge or its agitated basic impacts, or both [1]. Diabetes can incite persevering harm and brokenness of various tissues, especially heart, veins, eyes, kidneys, and nerves. Diabetes can be separated into two arrangements, type 1 and type 2 diabetes. The regular clinical side effects are extended thirst and successive pee [2].

Received: 08.05.2021 – Accepted: 30.05.2021

With the headway of daily comforts, diabetes is logically customary in people's step by step life. Thus, fast and exact identification of diabetes is a point excellent examining . In prescription, the investigation of diabetes is as shown by fasting blood sugar, glucose opposition, and subjective blood sugar levels. The earlier assurance is gotten, significantly easier we can deal with it. Artificial intelligence can help to make a groundwork decision about diabetes according to their step-by-step real appraisal data, and it can fill in as a wellspring of point of view for experts [3]. For AI strategy, the fundamental issues are to pick the significant features and the correct classifier.

The extent of this exploration is principally on the presentation investigation of disease forecast approaches utilizing various variations of AI algorithms. Disease forecast and in a more extensive setting, clinical informatics, have acquired critical consideration from the information science research recently [4]. This is principally because of PC based innovation into the health care area in various structures and resulting accessibility of enormous health data sets for specialists. These information are used in a wide scope of medical care regions, for example, the investigation of medical services usage, estimating execution of an emergency clinic care network, investigating examples and creating illness risk forecast model, constant sickness reconnaissance, and contrasting disease predominance and risk expectation models implying AI calculations (e.g., logistic regression, ANN and SVM), explicitly - supervised learning. Models dependent on these calculations utilize marked information for training. Patients are characterized into a few gatherings like generally safe and high danger in test set [5].

In particular, we discovered little exploration that makes a thorough audit of articles utilizing diverse learning calculations for disease forecast. Thusly, this examination plans to recognize key patterns among various sorts of supervised AI calculations, their presentation accuracies for the disease being examined. Likewise, the benefits and impediments of various supervised AI calculations are summed up. The consequences of this investigation will assist the researchers with bettering comprehend latest things and focal points of illness forecast models utilizing AI calculations and figure their exploration objectives appropriately.

Customarily, standard factual techniques and specialist's instinct, information and experience had been utilized for anticipation and illness risk expectation. This training frequently prompts undesirable predispositions, mistakes and high costs, and contrarily influences the nature of administration given to patients. Expanding accessibility of e-health information, more robust and progressed computational methodologies, for example, AI have gotten more viable to apply and investigate in disease forecast zone. The majority of the related work used at least one AI calculations for a specific disease expectation. Thus, the correlation of various supervised AI calculations for disease detection is the essential focal point of this examination.

This article is figured out as follows: Segment 2 portrays technique; Segment 3 depicts results and discussions and Segment 4 demonstrate Conclusions.

2. Methodology

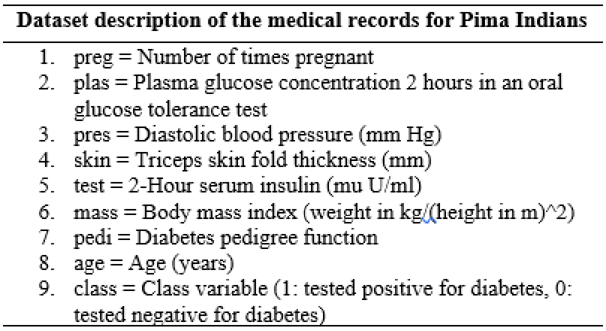
Dataset

For assessment of viability of the proposed model, we utilize the openly accessible "Pima Indian Diabetes" dataset. In this work, we lean toward this dataset in light of the fact that it gives experiences about indications that can anticipate the event just as the class of diabetes. This dataset is most broadly utilized for assessment of various ML models and it offers a chance to contrast the proposed model with existing techniques.

The dataset comprises the clinical history of 768 ladies patients. All patients lie in the age scope of 21 to 81years. The dataset contains the accompanying eight features as autonomous factors (clinical indicator) and one ward (target) variable as shown in Table 1.

70% of the information were used as preparing set and staying 30% of information were used for testing.

Table I - Description of Pima Indians Diabetes Dataset



Dataset description of the medical records for Pima Indians
1. preg = Number of times pregnant
2. plas = Plasma glucose concentration 2 hours in an oral glucose tolerance test
3. pres = Diastolic blood pressure (mm Hg)
4. skin = Triceps skin fold thickness (mm)
5. test = 2-Hour serum insulin (mu U/ml)
6. mass = Body mass index (weight in kg/(height in m)^2)
7. pedi = Diabetes pedigree function
8. age = Age (years)
9. class = Class variable (1: tested positive for diabetes, 0: tested negative for diabetes)

Machine Learning Algorithms

In this work, six machine learning techniques are used for identifying the diabetes. These models are utilized to improve forecast. These classifiers are contrasted to discover the best predictor of diabetes in patients.

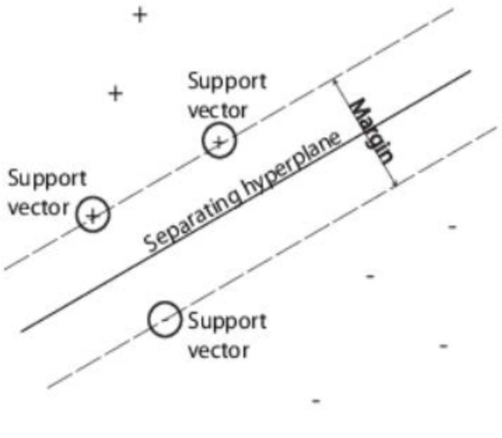
Logistic Regression (LR)

LR is a well-established and incredible technique for supervised grouping [6]. It is an expansion of normal regression furthermore, can display just a dichotomous which typically addresses the event or nonoccurrence of an occasion. LR discovers the likelihood that new occurrence has a place with a specific class. Since it is a likelihood, the result lies somewhere in the range of 0 and 1. Along these lines, to utilize the LR as a double classifier, an edge should be relegated to separate two classes. For instance, a likelihood esteem more than 0.50 for an info example will characterize it as 'class 1'; something else, 'class 2'.

Support Vector Machine (SVM)

A classifier that has gotten impressive consideration is SVM. This procedure has its underlying foundations in statistical learning hypothesis. As an undertaking of characterization, it looks for ideal hyperplane isolating the tuples of one class from another. SVM functions admirably with higher dimensional information and in this manner dodges dimensionality issue. Albeit the SVM based grouping is incredibly sluggish, the outcome, is anyway profoundly exact. Further, testing an obscure information is extremely quick. SVM is less inclined to over fitting than different strategies. It likewise encourages conservative model for classification.

Fig. 1 - Support Vector Machine for Classification



Decision Tree (DT)

DT learning is a regulated AI strategy for actuating a DT from training information. A DT is a prescient model which is a planning from perceptions about a thing to decisions about its target. In the tree structures, leaves address classifications, nonleafy hubs are highlights, and branches address conjunctions of highlights that lead to the groupings. Building a DT that is reliable with a given informational collection is simple. The test lies in building great DT, which normally implies the littlest DTs [7].

K-Nearest Neighbor

KNN is the most straightforward and soonest classification calculations. It tends to be thought a less difficult form of a Naive Bayes (NB) classifier. In contrast to the NB method, the KNN calculation doesn't need to consider likelihood esteems. KNN is a non-parametric, slow learning algorithm. Its inspiration is to use a data base in which the data centers are separated into a couple of classes to predict the portrayal of another example point. KNN Algorithm relies upon comparability. A thing is assembled by a larger part vote of its neighbors, with the article being allotted to the class commonly standard among its k nearest neighbors.

Artificial Neural Network (ANN)

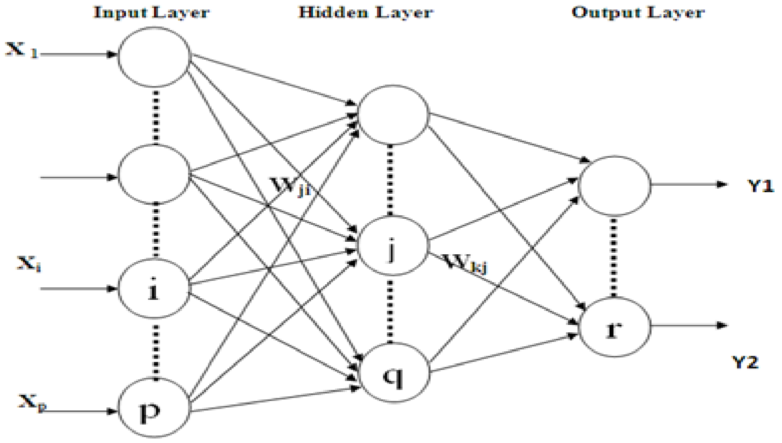
ANNs are normally inspired PC programs planned to imitate the way the human frontal cortex process information. ANNs amass their knowledge by distinguishing the models and associations in data and learn through experience, not from programming [8]. Another critical contrast between ANNs programming and other PC programs is that the calculations utilized for information investigation are adaptable. They can be changed whenever during the advancement of examination. The unmistakable component of ANNs is their capacity to manage multidimensional issues, including a few huge numbers of highlights. An ANN is framed from many single units, for example artificial neurons, associated with coefficients, which comprise the neural design and are coordinated in layers [9]. The capacity of neural calculations comes from associating neurons in an organization. The better the neurons are associated in networks, the better is the expectation as yield.

Ensemble Classifier

Ensemble learning improves AI results by joining a few models. This methodology permits the creation of better classifier contrasted with a solitary model.

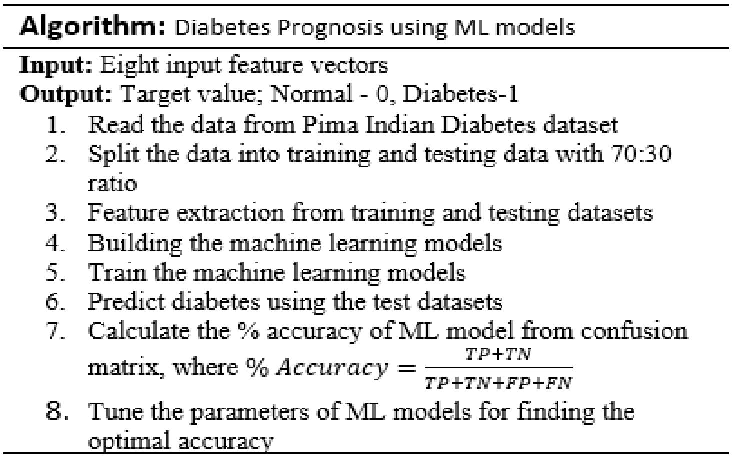
Received: 08.05.2021 – Accepted: 30.05.2021

Fig 2. Structure of Artificial Neural Network



Ensemble strategies are meta-calculations that consolidate a few AI procedures into one prescient model to diminish variance (bagging), biasing (boosting), or improve forecasts (stacking)[12]. Ensemble techniques can be partitioned into two gatherings specifically sequential and parallel strategies. In sequential strategies where the base learners are created consecutively (for example AdaBoost). The essential inspiration of sequential strategies is to utilize the reliance between the base learners. Parallel strategies where the base learners are created in parallel (for example random Forest). The essential inspiration of parallel strategies is to utilize freedom between the base learners since the error can be decreased significantly by averaging. All together for ensemble strategies to be more precise than any of its individual elements, the base learner must be pretty much as exact as could be expected.

In this investigation, ML models utilize the Pima Indian Diabetes datasets for predicting the diabetes. The overall prediction algorithm is presented in below.



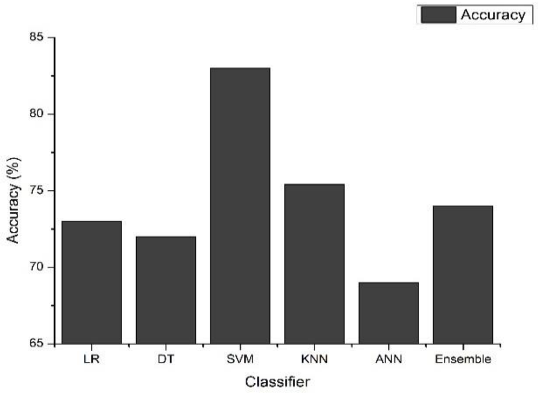
```
Algorithm: Diabetes Prognosis using ML models
Input: Eight input feature vectors
Output: Target value; Normal - 0, Diabetes-1
  1. Read the data from Pima Indian Diabetes dataset
  2. Split the data into training and testing data with 70:30
     ratio
  3. Feature extraction from training and testing datasets
  4. Building the machine learning models
  5. Train the machine learning models
  6. Predict diabetes using the test datasets
  7. Calculate the % accuracy of ML model from confusion
     matrix, where % $Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$
  8. Tune the parameters of ML models for finding the
     optimal accuracy
```

3. Results and Discussions

The investigations are directed for the forecast of Diabetes illness by applying different AI classifiers. From the test results, we distinguish that SVM classifier perform better when contrasted with other six ML classifiers in the forecast of diabetes.

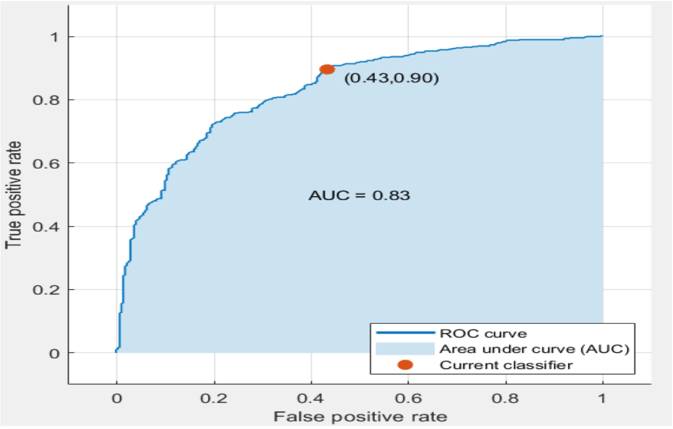
The undereath Fig 3 shows execution of different ML strategies in the forecast of diabetes. The exploratory outcomes show the examination of LR, DT, SVM, KNN, ANN and Ensemble classifiers and assess the exhibition based on exactness. Altogether classifiers, SVM perform better with a precision of 83%.

Fig. 3 - Comparison of accuracy of various classifiers



The fig.4 shows the ROC graph of the SVM classifier for detecting the diabetes. ROC curves are as a rule used to show in a graphical way the relationship among sensitivity and specificity for each test.

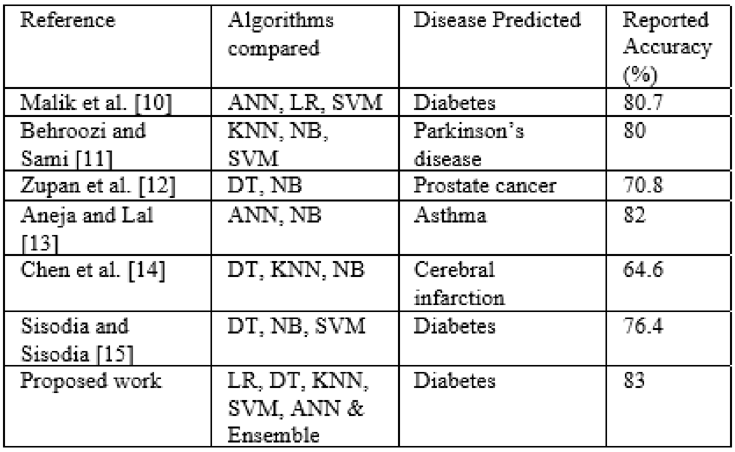
Fig. 4 - ROC curve of SVM classifier



ISSN: 2237-0722
Vol. 11 No. 4 (2021)
Received: 08.05.2021 – Accepted: 30.05.2021

Besides, the region under the ROC bend (AUC) gives an idea with respect to the upside of using the test(s) being alluded to. AUC estimates the whole two-dimensional territory under the whole ROC curve. AUC gives a total measure of performance across all classification limits. Table 2 shows the comparison of proposed work with earlier works carried out by using other datasets.

Table 2 - Comparison of performance of ML techniques with other datasets



Reference	Algorithms compared	Disease Predicted	Reported Accuracy (%)
Malik et al. [10]	ANN, LR, SVM	Diabetes	80.7
Behroozi and Sami [11]	KNN, NB, SVM	Parkinson's disease	80
Zupan et al. [12]	DT, NB	Prostate cancer	70.8
Aneja and Lal [13]	ANN, NB	Asthma	82
Chen et al. [14]	DT, KNN, NB	Cerebral infarction	64.6
Sisodia and Sisodia [15]	DT, NB, SVM	Diabetes	76.4
Proposed work	LR, DT, KNN, SVM, ANN & Ensemble	Diabetes	83

4. Conclusion

This exploration endeavored to examine relative performance of various ML algorithms in disease prognosis. The proposed approach is compared with six distinct methodologies, to be explicit LR, DT, SVM, KNN, ANN and Ensemble classifiers to show the improvement of Diabetes detection. The outcomes uncovered that, SVM calculation can recognize the Diabetes with improved exactness precision of 83%. Thus, the SVM model for predicting Diabetes can help the clinical specialists with taking key steps in early identification, which is a critical worry of clinical consideration in India.

References

A. Lonappan, J.Jacob, C. Rajasekaran, G.Bindu, , Thomas, V., , and Mathew, K. T., “Diagnosis of diabetes mellitus using microwaves”, *J. Electro. Wave.* 21, pp. 1393–1401, 2007.

A. Krasteva, Kisselova A., Panov, V. and Krastev, Z., “Oral cavity and systemic diseases—*Diabetes Mellitus”, Biotec. Equip.* 25, pp. 2183–2186, 2011.

I. Iancu, Iancu, E. and Mota M., and, “Method for the analysing of blood glucose dynamics in diabetes mellitus patients,” in *Proc. of the IEEE Int. Conf. on Automation, Qual. and Test., Robo.,* 2008.

Received: 08.05.2021 – Accepted: 30.05.2021

American Diabetes Association, “Diagnosis and classification of diabetes mellitus. *Diabetes Care”,* 35(Suppl.1), S64–S71, 2012.

I. Kavakiotis,, A. Salifoglou, , N. Maglaveras, Vlahavas, I., Tsave, O., and Chouvarda, I., “Machine learning and data mining methods in diabetes research”, *Comput. Struct. Biotech. J.* 15, 104–116, 2017.

D W Hosmer, S. Lemeshow, Sturdivant RX, “*Applied logistic regression”,* Wiley, 2013.

Cruz JA, Wishart DS. Applications of machine learning in cancer prediction and prognosis. *Cancer Informat.* Vol.2: 59–77, 2006.

H.S.N. Murthy, M. Meenakshi, “Novel and efficient algorithms for early detection of myocardial ischemia. *Int. J. Medical Eng. Informatics 9*(4): 351-372, 2017.

H.S.N Murthy, "Comparison of Ensemble Classifiers for Identifying the Wart Treatment Therapy", *Journal of Advanced Research in Dynamical & Control Systems,* vol. 12, issue no. 3, pp. 539-544, August 2020

Malik S, Khadgawat R, Anand S, Gupta S. “Non-invasive detection of fasting blood glucose level via electrochemical measurement of saliva”, *Springer Plus.,* 5(1):701, 2016.

Behroozi M, Sami A., “A multiple-classifier framework for Parkinson’s disease detection based on various vocal tests”, *Int J Telemed Appl.* Pp.1–9, 2016.

Zupan B, DemšAr J, Kattan MW, Beck JR, Bratko I., “Machine learning for survival analysis: a case study on recurrence of prostate cancer”, *Artif Intell Med.,* 20(1):59–75, 2000

Aneja S, Lal S., “Effective asthma disease prediction using naïve Bayes—Neural network fusion technique”, *IEEE International Conference on Parallel, Distributed and Grid Computing (PDGC),* pp. 137–40, 2014.

Chen M, Hao Y, Hwang K, Wang L, Wang L., “Disease prediction by machine learning over big data from healthcare communities”, *IEEE Access,* no.5, 8869–79, 2017

D. Sisodia, DS Sisodia, “Prediction of diabetes using classification algorithms”, *Proc. Comp. Sci.,* vol.132, pp.1578–85, 2018.