Integrated Intelligent Research (IIR)
International Journal of Computing Algorithm
Volume: 08, Issue: 01, June 2019, Page No.27-32
ISSN: 2278-2397

Analysis of Various Machine Learning Algorithms to Predict the Type II Diabetes Disease

S. Banumathi[1], A. Aloysius[2]
[1]Assistant Professor, Department of Computer Science, Holy Cross College, Trichy, Tamil Nadu, India
[2] Assistant Professor, Department of Computer Science, St. Joseph's College, Trichy, Tamil Nadu, India
Email: banu15.sm@gmail.com, aloysius1972@gmail.com

Abstract - The enormous growth of data in biomedical and healthcare communities need accurate analysis of medical data benefits in early disease detection, patient care and society services. However, the accuracy of analysis data will be condensed when the eminence of health care data is imperfect. In heath care community different regions exist with regional disease which would not be easily predicted with maximum of accuracy level. In this paper, we predict the diabetes disease and compare the algorithm which algorithm provide high performance, finally select the best algorithm to predict the diabetes disease at early stage. Machine learning algorithm can be applied for diabetes disease automating classification. This paper compares several Machine learning algorithms for classifying diabetes disease. Algorithms that involve Decision Tree, Naive Bayes, and KNN, SVM are proposed and assessed for this classification. These approaches have been tested with PIMA Indian Diabetes Dataset downloaded from UCI machine learning data repository. The performances of the algorithms have been compared in terms of Accuracy, Sensitivity, and Specificity with help of Sciklit-learn. Sciklit-learn are a free software machine learning library for the Python programming language. Finally comes with best suitable model for predict diabetes diseases.

Keywords: Machine learning algorithm, Decision Tree, Naive Bayes, KNN, SVM algorithm

I. INTRODUCTION

Data driven health care aims at effective promising utilization of big medical data which represents the collective learning in treating millions of patients to provide best and most personalized care is believed most potentially capable direction for the health care transformation [3]. The key aspect of extracting features from patient records is usually referred as "electronic prototyping", in medical informative. Lot of challenges is existing in health care data a few are high dimensionality, temporality, sacristy, and irregularity, bias. The analysis accuracy is less if there is incomplete and quality less health care data [4].

Diabetes is an extremely common chronic disease form which nearly 8.5% of the world suffer, 422 million people have struggle with diabetes [5]. Machine learning has driven advances in many domains including computer vision, natural language processing and automatic speech recognition to deliver powerful systems. Machine learning ability to extract information from data, paired with the centrality of data in healthcare, makes research in machine learning for health care crucial. The effective use of machine learning in health care presents many challenges and opportunities for researchers and the potential impact is vast. Machine learning methods offer a new approach to diabetes which is well suited to today's big data requirements.

The ability to deal with large volumes of both structured and unstructured data from different sources, big data analytical tool hold the promise to study outcomes of large scale population based longitudinal studies as well as to capture trends and propose predictive models for data generated from electronic medical and health records. EHR contains patient treatments and outcomes are rich but underused information. Traditional health care data centers used very enormous amount of data concern with disease diagnostics, lab test, medication and clinical data. Several ways of defining big data exist as a broad term to encapsulate the challenges related to the processing of massive amount of structured and unstructured data [7].The six v's of big data which can also be applied to health data shown in table.1.

II. RELATED WORK

The authors proposed predictive analysis system by analyzing the large diabetic data. The hadoop environment used to distribute the data among various server and the data replicated to several map reduce nodes. In mapping phase the master nodes splits large data with smaller tasks to all work nodes. The predictive analysis system act

as a pattern matching system. The pattern matching is the process of comparing the analyzed threshold value to the actual obtained value [15]. The author suggests location aware health care management and prediction system for rural area needed to provide treatment at low cost [1]. The author introduced new decision support system to predict kidney chronic disease. The data set from the patients collected and the proposed system used two machine learning algorithms SVM, KNN. The classifier used to predict the disease and the performance of the classifier evaluated. The proposed system used classification algorithms accuracy and that has been shown in confusion matrix. The proposed system used MAT Lab to analyze data and finally result accuracy proven KNN is the best predictor compared to SVM [2].

Table.1 Characteristics of Health Care Data

Value	Critically relevant data longitudinal studies
Volume	High throughput technologies continuous monitoring of vital signs
Velocity	High speed processing for fast clinical decision support
Variety	Heterogeneous and unstructured data sources
Veracity	Data quality is unreliable Data coming from uncontrolled environments
Variability	Seasonal health effect and disease evolution Non-deterministic models of illness and health

The proposed system uses decision tree and KNN as classification model for diabetes disease which reduces the time and cost of diagnoses. The model proved KNN has to highest accuracy of 95% than decision tree [6]. The author used a pair of different UCI machine learning repository data set for predicting the disease. The experimental results shown that SVM has higher accuracy of 95.556% as average for all data sets in the disease prediction [8]. Breiman's categorize two approaches for statistical modeling namely data model approach and machine learning approach. The data model assumes that data is generated by stochastic data model where the output is predicted by estimating the parameters of the input data. The machine learning approach views data output as arising from unknown input-output mapping process and overarching the goal of statistical modeling to learn function or an algorithm that best approximates mapping process [9].

The Machine learning algorithms precisely improves the algorithms automatically through experiences such as decision tree, random forest, neural network, and support vector machines. A recent survey shows that 15% of the hospitals using machine learning predictive models for clinical purpose [11]. Farahmandian et al. proposed and used various data mining methods for the diagnosis of diabetes using Pima Indians Dataset. The author proposed algorithms used 80% of data for training the learning algorithms and 20 % for testing. The author compare the algorithms used and the experimental result shows SVM algorithm is more accurate than other algorithms and the accuracy rate of 81.77% [12].

Vijayan et al. carried out an experiment on Pima Indians Dataset. The data set consist of seven hundred and sixty eight samples. The experiment consists of various data mining techniques such as KNN, Amalgam KNN, Kmeans, EM, and ANFIS. Comparing these algorithms, amalgam and the authors proposed prediction model identifies KNN was more accurate than others [13].

The authors established prediction out the diagnosis of diabetes on using Pima Indians Dataset. Data mining techniques are significant techniques for diagnosis of diseases. The authors used data mining algorithms such as Naive Bayes, RBF Network, and J48 were used to diagnose type II diabetes [14]. There were 768 samples for diagnoses of the disease among them 230 samples were selected for testing. Naive Bayes algorithm with accuracy rate of 76.95% had been proved as highest accuracy compared to J48 and RBF Network for diagnosis.

III. PROPOSED METHODOLOGY

Classification is a process discovery model (functions) that describe and distinguish classes of data or concept that aims to be used to predict the class of the object which label class is unknown. Classification is part of data mining, where data mining is a term used to describe the knowledge discovery in databases [16].

The classification process is based on four components they are class Categorical dependent variable in the form that represents the 'label' contained in the object. Predictor is the independent variables are represented by characteristic (attribute) data. Training dataset: One data set that has the assessment of both components above is

International Journal of Computing Algorithm
Volume: 08, Issue: 01, June 2019, Page No.27-32
ISSN: 2278-2397

used to determine a suitable class based on predictor. Testing dataset: Containing new data which will be classified by the model. The figure.1 shows the machine learning algorithms and its types. This research works focuses on only supervised learning algorithms with constraint of minimum data set to predict the type-II diabetes disease.

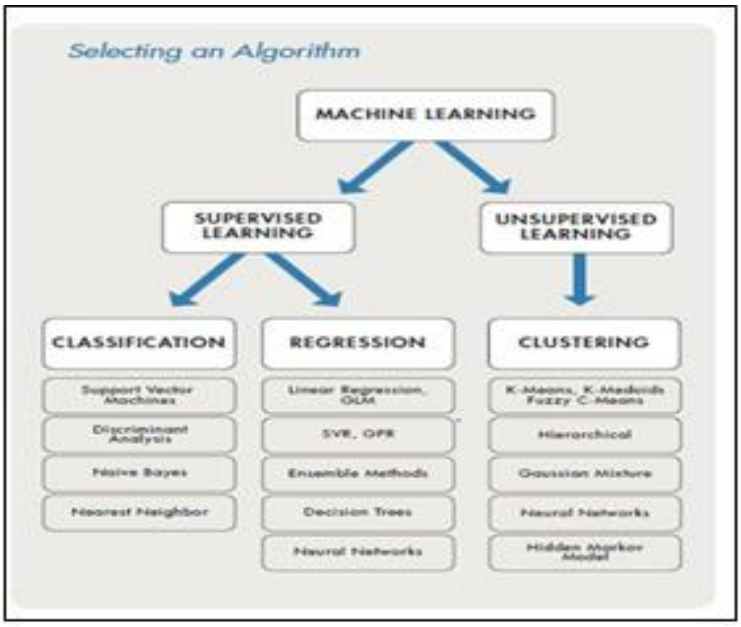

Fig.1 Machine Learning Algorithms

IV. RESULT AND DISCUSSION

Sciklit-learn is a liberated software for machine learning library with the Python programming language. Sciklit-learn are designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy. Those algorithms have been tested with PIMA Indian Diabetes Dataset downloaded from UCI machine learning data repository. The dataset consists of 9 essential attributes with 768 instances necessarily relates with the health care prediction. Attributes are exacting, all patients now are females at least 21 years old of Pima Indian heritage. Attributes are listed below in the below Table 2.

Table 2: Attributes of Data Set

S.No.	Features	Description for each feature
1.	Pregnancy	Number of times pregnant
2.	Plasma	Plasma glucose concentration
3.	Pres	Diastolic blood pressure (mm Hg)
4.	Skin	Triceps skin fold thickness (mm)
5.	Insulin	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 (in years)
9.	Class	Class variable (0 or 1)

The performances of the algorithms have been compared in terms of Accuracy, Sensitivity, and Specificity. Based on PIMA Indian dataset, confusion matrix figure.2 can be taken as, **TP (True Positive):** The no. of people who actually suffer from ‘diabetes’ among those who were diagnosed ‘diabetic’. **TN (True Negative):** States the number of people ‘healthy’ among those who were diagnosed ‘diabetic’. **FP (False Positive):** Depicts the number of persons who are unhealthy that is, ‘diabetic’ but was diagnosed as ‘healthy’. **FN (False Negative):** The number of people found to be ‘healthy’ among those who were diagnosed as ‘diabetic’. The performance of classification can be measure based on the criteria of Sensitivity must have high percentage, Specificity must have low percentage and Accuracy must have high percentage. Classification accuracy is the

International Journal of Computing Algorithm
Volume: 08, Issue: 01, June 2019, Page No.27-32
ISSN: 2278-2397

percentage of instances that are correctly classified by the model. Consider N be the total number of samples. The classification accuracy for total number of samples N is calculated as the sum of correct classification divided by N. Sensitivity is the measure of the ability of a classification model to select instances of certain class from the dataset. It is the proportion of actual positive which are predicted positive. Specificity is a measure that is commonly used in two class problems where the focus is on a particular class. The true negative rate is the percentage of the negative class that was predicted negative. The mathematical formulas for calculating sensitivity, specificity and accuracy have shown below.

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

$$\text{Specificity} = \frac{TN}{FP + TN}$$

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

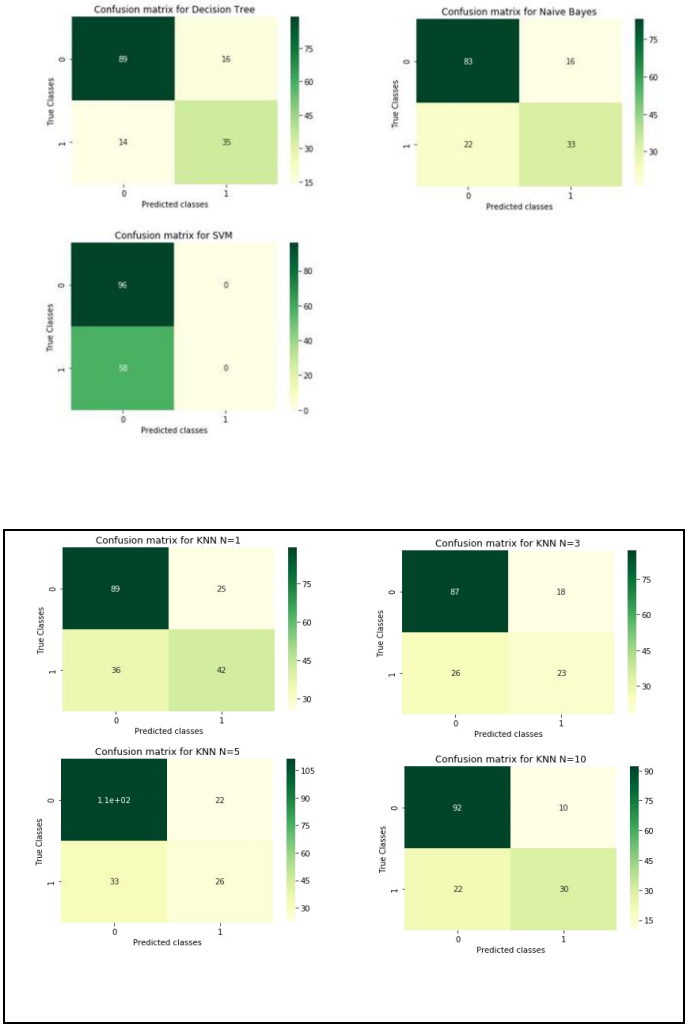

Fig.2 Confusion Matrixes

International Journal of Computing Algorithm
Volume: 08, Issue: 01, June 2019, Page No.27-32
ISSN: 2278-2397

The figure.2 represents confusion matrix for four machine learning algorithms such as Decision tree, Support vector machine, Naive bayes and KNN with neighbourhood of N= 1,3,5,10 values. The prediction model compared by the iteration calculations. More than five iterations have considered. In each iterations the values of accuracy, sensitivity, specificity compared and found that the values are differs in floats only. There is no much deviations identified in the measured values. The table3 shows the values for all algorithms. The following graph shows the accuracy, sensitivity, specificity value distribution among four algorithms with training and testing dataset for the prediction of disease.

Table.3 Comparison of Algorithms

S.no	Algorithms	Accuracy	Sensitivity	specificity
1	Decision tree	0.734	0.774	0.715
2	Naïve bayes	0.776	0.693	0.815
3	KNN	0.635	0.532	0.684
4	SVM	0.677	0.531	0.695

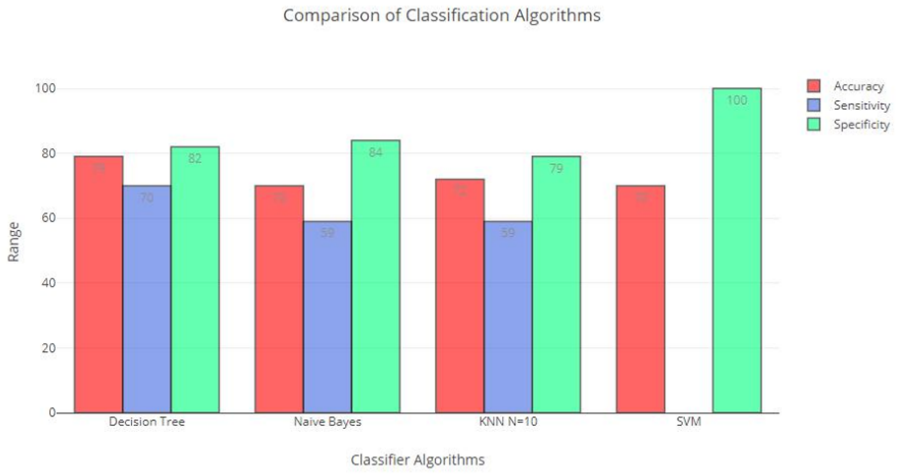

Figure.3 Comparison of Algorithms

V. CONCLUSION

In summary, we have compared four prediction models for predicting diabetes disease using eight important attributes after preprocessing. Here these projects conclude that the decision tree classifier achieves higher accuracy of 79.82 % than other three classifiers, highest sensitivity and also second lowest specificity. Decision tree classifier can be used to predict diabetes disease as best than other models. The future work can be implementing and comparing the algorithms with large data set.

References

[1] Dr. Saravanakumar N M, Eswari T, Sampath P, Lavanya S, "Predictive methodology for Diabetic Data Analysis in Big Data" , Elsevier, 2015.
[2] Parul Sinha, Poonam Sinha, :comparative study of chronic kidney disease predition using KNN and SVM", International journal of Engineering Research & Technology (IJERT), Dec, 2015, vol.4, issue.12.
[3] Yu cheng, Fui Wang, Ping Zhang, Jinaying Hu, "Risk predictin with Electronic Health recods, A Deep learning Approach", Proceedings of the 2016 SIAM International Conference on Data Mining ,10.1137/1.9781611974348.49.
[4] Minchen, YixueHao, KaiHwang, Luwang, Linwang, "Disease prediction by machine learning over big data health care communities", IEEE, june 2017.
[5] MinChen, JunYang, Jiehan Zhou, Yixue Hao, Jing Zhang, Chan Hyun Youn, " 5G smart diabteres: toward personalized diabetes diagnosis with health care big data clouds", IEEE, 2017.
[6] K.Lakshmi, D. Iyajaz Ahmed, G. Sivakumar, " A smart clinical decision support system to predict diabetes disease using classification techniques", IJSRSET, Vol.4, Issue.1.
[7] Javier Andrue_perez, Carmen c.Y.Poon, Robert D.Merrifield, Stephen T.C.Wong, and Guang-Zhong Yang,

"Big data for Health", IEEE, July 2015, Vol.19, No.4
[8] R.Thayammal, S.Kamalakannan, P.Kavith," Big data analytics in diabetic disease data prognis using naivebayes classifier algorithm", Internation journal of pure and applied mathematics, vol.119, no.10, 2018, pg.21-27.
[9] YakubSebastuan , Xun Ting Tiong, Valliappan Raman, Alan Yean Yip Iong, Patrick Hung Huithea, " Advances in Diabetic analytics from clinical and machine learning perspectives", International journal of design, analysis and tools for integrated circuits and systems, vol.6, no.1, oct, 2017.
[10] Gang Luo, "predit-ML a tool for automating machine learning model building with big clinical data ", Life health information science system, 2016, DOI. 10.1186/2/13755-016-0018-1.
[11] Marzyeh Ghassemi, Tristan Naumann, Peter Schulam, Andrew L.Beam, Rajesh Ranganath,"opportunities in machine learning for health care", arXiv:1806.00388.
[12] Farahmandian M, Lotfi Y, Maleki , "Data Mining Algorithms Application In Diabetes Diseases Diagnosis: A Case Study", MAGNT Research Report, 2015.
[13] V. Vijayan V , A. Ravikumar, "Study on Data Mining Algorithms For Prediction Diagnosis Of Diabetes Mellitus", International Journal of Computer Applications , Vol.95, No.17, pp.12-16, June 2014.
[14] Amanj Maleki , Zahra Panbechi , Sadri , "Comparison Of Data Mining Algorithms In The Diagnosis Of Type Ii Diabetes" ,International Journal on Computational Science & Applications (IJCSA) Vol.5, No.5,October 2015 DOI:10.5121.
[15] S. Banumathi, A. Aloysius," Big Data Prediction Using Evolutionary Techniques: A Survey", Journal of Emerging Technologies and Innovative Research**,** September 2016, Volume 3, Issue 9, Pg. 89-91.
[16] S. Banumathi, A. Aloysius," Prediction Model with Selection of Best Prediction Algorithm for Big Data", Saudi Journal of Engineering and Technology (SJEAT), March, 2018, ISSN 2415-6264.

Authors Profile

S. Banumathi doing her Doctoral Degree in Computer Science at Bharathidasan University, Tiruchirappalli, Tamilnadu, India. She has more than Seven years of teaching and research experience. She is currently working as Assistant Professor, Department of Computer Science, Holy Cross College, Tiruchirappalli, Tamilnadu, India. She had published several papers in international journal. She had presented various papers in national and international conferences. Presently she is doing her research on Big data analytics, Machine learning.

Dr. A. Aloysius is working as an Assistant Professor in Department of Computer Science, St. Joseph's College, Trichy, Tamil Nadu, India. He has 19 years of experience in teaching and research. He has published many research articles in the National/ International conferences and journals. He has also presented research articles in the International Conferences on Computational Intelligence and Cognitive Informatics in Indonesia. He has acted as a chair person for many national and international conferences. His current area of research is Cognitive Aspects in Software Design, Big Data, and Cloud Computing.