ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 5, May 2019

Analysis and Classification of Diabetes using Machine Learning

Sakshi Bhosale[1], Revati Akole[1] , Kalpesh Patil[1], Rahul Patil[1]

U.G. Student, Department of Computer Engineering, SSBT's COET Bambhori Jalgaon, India[1]

ABSTRACT: The hyper triglyceridemic waist (HW) phenotype is strongly associated with diabetes however to date, no study has accessed the predictive power of phenotypes based on individual anthropometric measurements. In this paper, an approach for the classification of dynamic models of diabetes mellitus is presented. Principal component analysis and a support vector machine are used to reduce the feature space and to find a suitable classifier. The data covers type 1, type 2, and non-diabetic virtual subjects. A data informed framework is proposed for identifying the subjects with and without DM from EHR via feature engineering and machine learning. For evaluating and contrasting the identication performance of widely used machine learning models within the framework, k-Nearest-Neighbours, Naive Bayes, Decision tree, Random Forest, Support Vector Machine and Logistic Regression are used.

KEYWORDS: Naive Bayes, Type 1 and Type 2 Classification, phenotypes, EHR.

I. INTRODUCTION

Diabetes mellitus is chronic disorder of the glucose-insulin metabolism, which affects a growing number of people – worldwide 425 million adults are concerned. Diabetes may require life-style changes, insulin administration, and may cause further severe diseases. In the field of personalized medicine, model-based diagnosis can be an important step towards a better understanding of diseases such as diabetes.A diagnostic decision is a classification procedure for which medical guidelines provide definitions and diagnostic rules. Machine learning and data mining methods have become widely used in the field of biomedical engineering in the past years to automatically support physicians in their diagnostic decision making. The risk factors of diabetes were investigated extensively in the past studies, but it remains unknown that which risk factors were associated with diabetes than others. However no study has accessed the predictive power of using various phenotypes consisting of individual anthropometric measurements that uses the actual TG and WC values as components of HW phenotype. A widely adopted approach for identifying subjects with and without Diabetes Milletus is to have human experts manually design algorithms based on examination of EHR data. However such strategies prove to be limited and not scalable. Expert algorithms are often designed with conservative identification which fail to identify complex subjects and miss a potential number of significant cases and controls. So, an automated framework should be developed which reduces these complexities. Many researchers are conducting experiments for diagnosing the diseases using various classification algorithms of machine learning approaches like J48 ,SVM ,Naive Bayes, Decision Tree etc as researchers have proved that machine learning algorithms works better in diagnosing different diseases.

The main aim of this work is the detection of Diabetes Milletus using an hybrid model classification comprised of Bayesian Classification, Logistics Regression(LR), SVM, ID3, Neural Network.

II. LITERATURE SURVEY

In recent years, predictive classification is one of the most essential and important tasks in data mining and machine learning. Its application to the medical diagnosis has received a strong boost due to earnest research activities in the medical big data field. Many researchers have highlighted the potential of predictive classification to provide decision support for doctors and medical professionals. Over the last few years, a great deal of research has been conducted on different datasets to predictive diabetes. Many of them showed good classification accuracy. A journalism assessment acknowledgement several consequences taking place diabetes passed out by special methods and resources of diabetes difficulty in India. A lot of people urbanized as a result of the different calculation models with machine learning systematic technique be able to be used to guess the diabetes. The traditional neural network model is used to see coming the pre-processed dataset and also they replaced the absent importance with area of the matching feature.The obtained consequences pertaining to the level of danger which lying on your front to both heart bother or knock. The work of fiction pre-processing point with absent worth declaration for mutually statistical and unconditional information. A fusion combination of arrangement and failure grass (CART) and Genetic Algorithms to impute missing continuous values and Self Organizing Feature Maps (SOFM) to impute categorical values was improved. J. Pradeep Kandhasamy and S. Balamurali using data sample from UCI machine learning data repository to compare the performance of four common classifiers (J48 Decision Tree, K-Nearest Neighbors, Random Forest, and Support Vector Machines) to classify diabetes mellitus patients[9]. The result shows that the J48 decision tree classifier achieves higher accuracy of 73.82 % than other three classifiers before data preprocessing and both KNN (k=1) and Random Forest show better performance than the other three classifiers after data preprocessing [16]. Rashedur M. Rahman and Farhana Afroz present a comparative study of different classification techniques by using three different data mining tools named WEKA, TANAGRA and MATLAB to analyze the performance of different classification algorithms for a large dataset. The study shows that the best algorithm in WEKA is J48graft with an accuracy of 81.33%, Naive Bayes classifier provides an accuracy of 100% in TANAGRA and ANFIS has 78.79% accuracy in MATLAB [1]. Xue-Hui Meng developed three predictive models (logistic regression, artificial neural networks and decision tree) then compare the performance by using 12 risk factors.

Asma A. AlJarullah conducts a diabetes prediction model by using the decision tree algorithm. In this study, Weka's J48 decision tree classifier was applied to the dataset to construct the decision tree model. The accuracy of the resulting model was 78.1768% [4]. Wei Yu presents a potentially useful alternative approach based on support vector machine (SVM) techniques that can be used to classify persons with and without diabetes. The study used the data from the U.S. National Health and Nutrition Examination Survey to develop SVM models two classification schemes, one is diagnosed or undiagnosed diabetes vs. pre diabetes or no diabetes, the other is undiagnosed diabetes or pre-diabetes vs. no diabetes. The results show the area under the receiver operating characteristic (ROC) curve were respectively 83.5% and 73.2%. The result indicates SVM modeling is a promising classification approach for detecting common diseases like diabetes [7]. Mira Kania Sabariah, Aini Hanifa and Siti Sa'adah combine Classification and Regression Tree method (CART) and Random Forest (RF) to build the classification model that can be used in the early detection of Type 2 diabetes. The study shows that the average accuracy of the proposed model is 83.8%, which is higher than the single classifier CART [8].

III. PROPOSED SYSTEM AND DISCUSSION

The following figure illustrates the flow of the proposed model which consists of training the data and then applying the property machine learning algorithms so as to remove the in-consistencies and get the accurate data for the electronic health records from PIMA dataset.

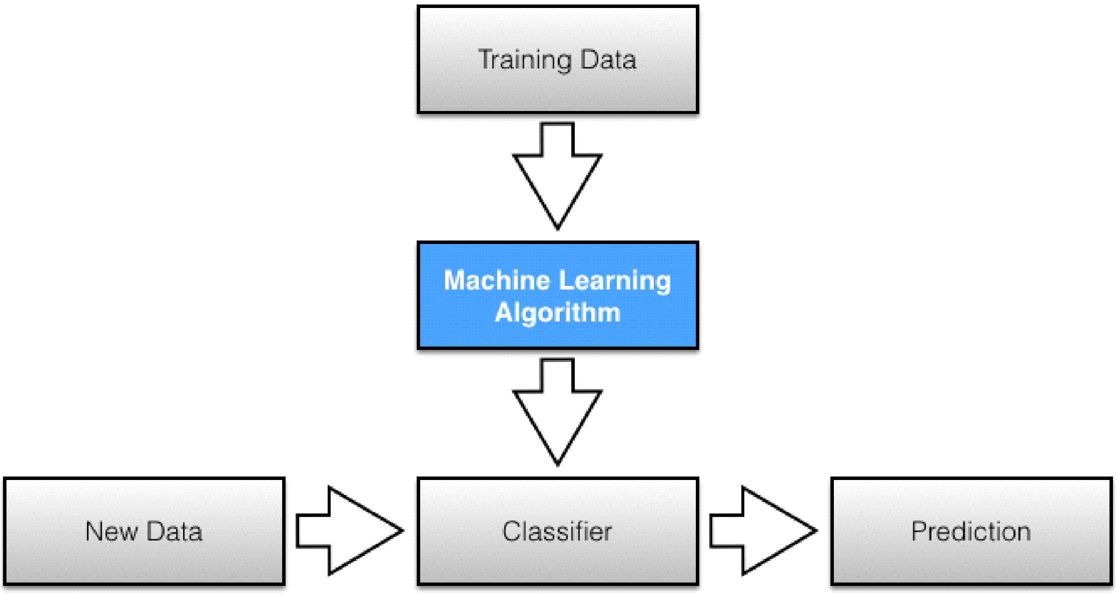


Fig 1: Proposed Model

In our proposed work, we collect the dataset of diabetic patients which comprised of 403 observations. The 10th attribute represents class diabetic and non diabetic with values 0 and 1 respectively. In the first step we applied the data pre processing for the row dataset. Then applied principal component analysis to know the most important features.

Feature selection

PCA is used to structure, simplify and illustrate extensive data sets by approximating a large number of statistical variables by a smaller subset of linear combinations that are as meaningful as possible. Consequently, potentially redundant information in the data is removed and only a few variables are required to describe most of the variances in the original data.

Naive Bayes:-

After selecting the features we applied the Naive Bayesian Classifier for finding probability

$$P(a|b)=P(b|a)P(a)/P(b)$$

where P(a|b)= probability of class with attribute
P(a)=prior probability of class
P(b|a)=likelihood which is the probability of predictor given class
P(b)=prior probability of predictor

Steps:-

1. Load diabetes data set which is in .csv format and split into training and testing dataset.
2. Summarize the properties of training dataset so that we can calculate the probabilities and make prediction.
3. Calculate the class probabilities of all the attribute values for a data instance.
4. After calculating the class probability we predict the laegest probability and return the associated values.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

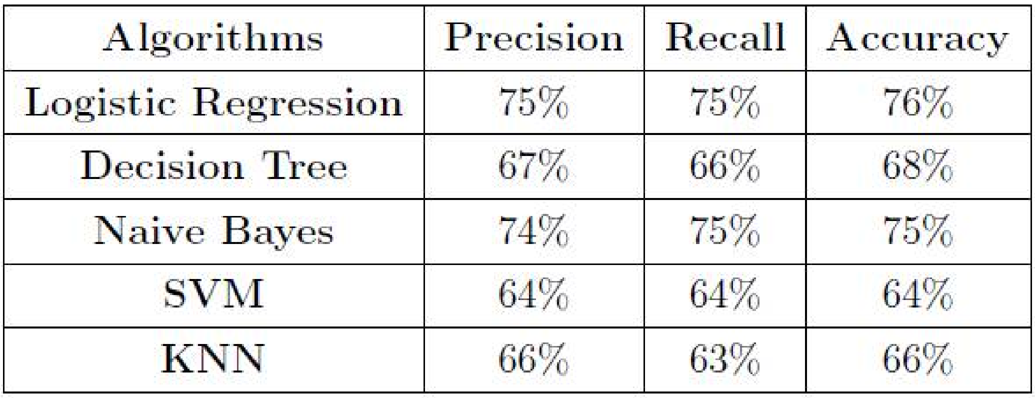
(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 5, May 2019

IV. EXPERIMENTAL RESULTS

The table represents different performance values of all classification algorithms calculated on various measures. And hence it is analyzesd that our model has achieved the highest accuracy of 76% with 6 features with the help of Logistics Regression.



Algorithms	Precision	Recall	Accuracy
Logistic Regression	75%	75%	76%
Decision Tree	67%	66%	68%
Naive Bayes	74%	75%	75%
SVM	64%	64%	64%
KNN	66%	63%	66%

Table 1: Estimated measures

V. CONCLUSION AND FUTURE WORK

Thus we have studied that as hypertriglyceridemic waist (HW) phenotype is strongly associated with type 1 and 2 diabetes however to date, no study has assessed the predictive power of phenotypes based on individual anthropometric measurements and triglyceride (TG) levels. We measured fasting plasma glucose and TG levels and performed anthropometric measurements. We employed binary logistic regression (LR) to examine statistically significant differences between normal subjects and those with type 1 and 2 diabetes using HW and individual anthropometric measurements. For more reliable prediction results, two machine learning algorithms, naive Bayes (NB), SVM, ID3 and Neural Network were used to evaluate the predictive power of various phenotypes. Experiments are performed on Pima Indians Diabetes Database. In future work, the designed system with the used machine learning classication algorithms can be used to predict or diagnose other diseases. The work can be extended and improved for the automation analysis including some other.

REFERENCES

1. P. T. Katzmarzyk, C. L. Craig, and L. Gauvin, “Adiposity, physical fitness and incident diabetes: The physical activity longitudinal study,” *Diabetologia*, vol. 50, no. 3, pp. 538–544, Mar 2007.
2. Rajendra A U, Tan P H, Subramaniam T, et al.“ Automated Identification of Diabetic Type 2 Subjects with and without Neuropathy Using Wavelet Transform on Pedobarograph,” Journal of Medical Systems, 2008.
3. Yu W, Liu T, Valdez R, et al. “ Application of support vector machine modeling for prediction of common diseases, the case of diabetes and pre-diabetes,” BMC Medical Informatics and Decision Making, 2010
4.Jarullah A A A.“ Decision tree discovery for the diagnosis of type II diabetes,” International Conference on Innovations in Information Technology. IEEE, 2011.
5. S. Fazeli Farsani, M. P. Van Der Aa, M. M. J. Van Der Vorst, C. A. J. Knibbe, and A. De Boer, “Global trends in the incidence and prevalence of type 2 diabetes in children and adolescents: A systematic review and evaluation of methodological approaches,” Diabetologia, vol. 56, no. 7, pp. 1471–1488, 2013.
6. Kerner W, Brückel J. Definition, “classification and diagnosis of diabetes mellitus Experimental and clinical endocrinology & diabetes official journal,” German Society of Endocrinology [and] German Diabetes Association, 2014.

7. Hasim N, Haris N A. “A study of open-source data mining tools for forecasting,” International Conference on Ubiquitous Information Management and Communication. ACM, 2015.

8. Sabariah M T M K, Hanifa S T A, Sa'Adah M T S. “Early detection of type II Diabetes Mellitus with random forest and classification and regression tree (CART),” Advanced Informatics: Concept, Theory and Application IEEE, 2015.

9. Kandhasamy J P, Balamurali S. “Performance Analysis of Classifier Models to Predict Diabetes Mellitus,” Procedia Computer Science, 2015

10. C. A. C. Montanez, P. Fergus, A. Hussain, D. Al-Jumeily, B. Abdulaimma, J. Hind, and N. Radi, “Machine learning approaches for the prediction of obesity using publicly available genetic profiles,” Proc. Int. Jt. Conf. Neural Networks, vol. 2017–May.

11. Katon, W. J., Rutter, C., Simon, G., Lin, E. H., Ludman, E.,Ciechanowski, P., ...& Von Korff, M.“ The association of comorbid depression with mortality in patients with type 2 diabetes.Diabetes care,” 28(11), (2005).

12. Dai, W., & Ji, W.“ A maeduce implementation of C4. 5 decision tree algorithm,” . International Journal of Database Theory and Application,. (2014)

13. Nanri, A., Mizoue, T., Noda, M., Takahashi, Y., Matsushita, Y., Poudel- Tandukar, K., ...& Japan Public Health Center–based.“ Prospective Study GroupFish int ake and type 2 diabetes,” in Japanese men and women: the Japan Public Health Center–based Prospective Study. The American journal of clinical nutrition, . (2011). .

14. Shirkhorshidi, A. S., Aghabozorgi, S., Wah, T. Y., & Herawan, .“ Big data clustering: a review.,” In International Conference on Computational Science and Its Applications (pp. 707-720). Springer International Publishing.. (2014, June).

15. UCI Machine Learning Repository.
http://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes