ISSN: 0974-5823 Vol. 7 (Special Issue, Jan.-Mar. 2022)

International Journal of Mechanical Engineering

Classification of Diabetes Mellitus and Healthy Subjects using PIMA Dataset by Logistic Model Tree and Naive Bayes Machine Learning Algorithms

Vajeefa K[1], Usharani Thirunavukkarasu[2]

[1]Research Scholar, Department of Biomedical Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, Tamil Nadu. India. Pincode: 602105

[2]Project Guide, Corresponding Author, Department of Biomedical Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, Tamil Nadu. India. Pincode: 602105.

ABSTRACT

Aim: The aim of our study is to classify the healthy and diabetes mellitus subjects using a logistic model tree (LMT) and Naive Bayes (NB) classifiers. **Materials and Methods:** The proposed study used the LMT and NB machine learning algorithms to classify diabetes using Indian Diabetic Database (PIMA) dataset with healthy (n=44) and diabetic (n=44) subjects which are collected from kaggle, a machine learning repository. The normal and diabetic subjects were classified using Waikato Environment for Knowledge Analysis (WEKA) version 3.8.5, a data mining tool. The statistical analysis was performed using IBM SPSS software version 21. **Results:** The statistical significant difference ($p<0.01$) was observed between the groups. The NB classifier has achieved the classification accuracy rate as 64.77% and LMT classifier has obtained 62.5%. **Conclusion:** The classifiers have been trained, tested and validated using 10-fold cross-validation in WEKA tool, the NB classifier has achieved a higher classification accuracy rate than LMT classifier.

Keywords: Diabetes Mellitus, Logistic Model Tree, Innovative Naive Bayes, Data Mining Tool, Artificial Intelligence.

INTRODUCTION

Diabetes mellitus (DM) is a heterogeneous disorder which is characterized by hyperglycemia and glucose intolerance. Insulin allows the cells in the muscles, fat and liver to absorb glucose that is present in the blood. So it plays a major role in controlling the glucose levels in the blood (Burges 1998). The lack of insulin secretion, and defect in insulin action leads to type 1 and type 2 diabetes (American Diabetes Association 2014). An increase in the blood glucose levels leads to complications such as failure of various organs, including the kidneys, eyes, nerves, blood vessels (Kumari and Chitra 2013). Till today the diabetes was diagnosed by the invasive techniques. The importance of the study is to diagnose diabetes and pre-diabetes at the earlier stage to prevent the onset of disease and its complications (Rosenbauer et al. 1999). This study can also be applied to diagnose cardiovascular disease (CVD), chronic kidney disease (CKD) at an earlier stage (Yu et al. 2010).

Related to our work, we have obtained around 288 articles from google scholar and 5 articles from pubmed in the year of 2016-2021. The authors have obtained the classification accuracy of 0.77% and 0.78% for artificial neural network and logistic regression classifiers, with F1 measures of 0.83 and 0.84 for predicting diabetes (King, Aubert, and Herman 1998). The authors experimentally tested the NB, K-Nearest Neighbour (k-NN),

and Logistic Regression (LR) algorithms to classify the normal and diabetes from UCI data sets. Among the three classifiers, NB has a better classification accuracy rate than other classifiers (Jiang et al. 2005). Some authors have used the KNN, Adaboost, and LR to classify the diabetes and observed that the LR classifier obtained a better classification accuracy rate (Webb, Boughton, and Wang 2005). Some experimental results show that the adaboost algorithm with decision stump has obtained an accuracy rate as 80.72% which is greater compared to that of various artificial intelligence algorithms such as Support Vector Machine (SVM), NB and Decision Tree to classify the diabetes (Ning Wang and Guixia Kang 2012). The SVM classifier has achieved only 70% of the classification accuracy rate for categorizing the normal and diabetes which is found to be greater than KNN, LMT machine learning algorithms (Hassanein et al. 2017).Previously our team has a rich experience in working on various research projects across multiple disciplines(Ezhilarasan et al. 2021; Balachandar et al. 2020; Muthukrishnan et al. 2020; Kavarthapu and Gurumoorthy 2021; Sarode et al. 2021; Hannah R et al. 2021; Sekar, Nallaswamy, and Lakshmanan 2020; Appavu et al. 2021; Menon et al. 2020; Gopalakrishnan et al. 2020; Arun Prakash et al. 2020)

Based on the review of literature and according to our knowledge, the classification of diabetes using the Indian Diabetic Database (PIMA) dataset using a LMT classifier was found to be limited. So we have used the LMT machine learning algorithm for classifying the healthy and diabetes subjects using the PIMA dataset as the innovative approach in our proposed study. We have learnt about WEKA, a data mining tool and SPSS, statistical software. So, the aim of our study is to classify diabetes and healthy subjects using LMT and NB machine learning algorithms from PIMA dataset.

MATERIALS & METHODS

The study has been performed in the Digital Signal Processing laboratory at the Department of Biomedical Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences. Our research does not require ethical approval. For the classification task, we used WEKA software version 3.8.5. The IBM SPSS (statistical package for the social sciences) software version 21 was used to conduct the statistical analysis for our research.

The PIMA dataset was obtained from kaggle, a machine learning repository which was used in our research. The parameters such as number of pregnancies, glucose levels, blood pressure, skin thickness, insulin, BMI (Body Mass Index), diabetes pedigree function, and age are available in the PIMA dataset. We have calculated the sample size using clincalc.com with alpha value as 0.05, a 95% confidence interval of 80% and enrollment ratio as 1 with mean accuracy of group 1 as 62.5% and group 2 as 64.77% (Rahman, Afroz, and Others 2013).

Our proposed study has a total sample size of 88 individuals, with group 1 (healthy,n=44) and group 2 (diabetes, n=44). The PIMA dataset was compiled in.xlsx format and converted to.arff format to perform the classification task in the WEKA software.

Statistical Analysis

The unpaired t test was performed to find the mean, standard deviation, and standard error mean. The independent sample test was performed to find the statistical significance between the groups. The classification accuracy rate (%) obtained from the machine learning algorithms are dependent variables in our study. There are no independent variables in our study.

RESULTS

The confusion matrix of the LMT classifier was tabulated in Table 1. It represents, true positive (TP) is n=29, the healthy people are correctly identified as healthy. The true negative (TN) is n=15, the diabetic patients are correctly identified as diabetic. The false positive (FP) is n=18, healthy people are incorrectly identified as diabetic. The false negative (FN) is n=26, diabetic are incorrectly identified as healthy.

The confusion matrix of the NB classifier was tabulated in Table 2. It represents, the true positive (TP) is n=27, the healthy people correctly identified as healthy. The true negative (TN) is n=17, the diabetic correctly identified as diabetic. The false positive (FP) is n=14, the healthy people incorrectly identified as diabetic. The false negative (FN) is n=30, the diabetic incorrectly identified as healthy.

From Table 3, it was observed that NB classifiers have obtained a better classification accuracy rate as 64.77%. The sample size of group 1 is 44, the sample size of group 2 is 44. From Table 4, it was observed that the mean accuracy of group 2 (NB classifier) is higher than the mean accuracy of group 1 (LMT classifier). An independent sample T test results of LMT and NB classifiers was represented in Table 5.

The comparison of mean accuracy of LMT and NB classifiers was shown in Fig. 1. The mean accuracy of the NB machine learning algorithm was observed to be better than the LMT machine learning algorithms.

DISCUSSION

From Table 5, an independent sample T test, it was observed that there was a statistical significant difference ($p<0.01$) between LMT and NB classifiers. The innovative naive bayes classifier has obtained 64.77% as the classification accuracy rate which appears to be higher than LMT classifier (62.5%).

(Thaiyalnayaki 2021)et.al has used Multi Layer Perceptron (MLP) and Support Vector Machine (SVM) classifiers with 18 parameters to classify diabetes and normal from PIMA dataset. They have correctly classified 595 instances with a classification accuracy of 77.47% and incorrectly classified 173 instances with 22.52% as classification accuracy rate using MLP. Similarly the SVM classifier correctly classified 500 instances with a classification accuracy rate as 65.10% and incorrectly classified 268 instances with 34.89% as a classification accuracy rate. The MLP classifier has outdone the SVM classifier by its classification accuracy rate using PIMA dataset (Thaiyalnayaki 2021).

(Thulasi, Ninu, and Shiva 2017) et.al used three artificial intelligence algorithms such as random forest (RF), SVM, and NB to perform the classification of diabetes and non-diabetes from PIMA dataset. The performance of these algorithms is evaluated on different measures like precision, recall, F-measure, and accuracy rate. Their experimental results show that the NB classifier outperforms with the highest classification accuracy rate compared with RF and SVM algorithms (Thulasi, Ninu, and Shiva 2017). (Patra and Khuntia 2021) et. al has used the Pima Indian Diabetes Dataset (PIDD) in their study. They have splitted the dataset into 90% for training and 10% for testing and found that KNN classifier has obtained maximum classification accuracy rate as 83.2% for predicting the diabetes (Patra and Khuntia 2021).

(Bilous and Donnelly 2010) et. al has obtained the classification accuracy rate as 91% for NB and KNN machine learning algorithms to classify the diabetes and normal with 9 % of error rate using PIMA dataset (Bilous and Donnelly 2010). (Nurjahan et al. 2021) constructed a predictive model through examining several machine learning techniques such as decision tree, KNN, NB, SVM, LR, extreme Gradient Boosting, MLP and RF on two different datasets of diabetes patients namely Pima Indian diabetes datasets and Sylhet Diabetes Hospital datasets. Their experimental results revealed that RF classifier has obtained the highest classification accuracy rate as 97.5%, F-measure as 97.5%, Area under Receiver Operating Characteristic Curve as 99.80% for Sylhet hospital datasets. For the PIMA dataset, LR classifier has obtained the highest classification accuracy rate as 77.7%, F-measure as 77% and Area under Receiver Operating Curve as 83% (Nurjahan et al. 2021). (Sarwar, Kamal, and Hamid 2018) et.al have obtained 77% as accuracy rate for SVM and 71% accuracy rate for RF classifier for the classification of diabetes mellitus using PIDD dataset (Sarwar, Kamal, and Hamid 2018). The opposing findings were not observed in our study according to our knowledge from the review of literature. The factors such as hereditary, unhealthy diet, sedentary lifestyle and aging of the subjects might affect the results of our study.

As the limitation of this study, the modifications can be made in the machine learning algorithms to obtain a better classification accuracy rate (%) and sampling size can also be increased in our study. Furthermore, machine learning algorithms can be incorporated for the dataset (PIMA) to obtain a better classification accuracy rate to classify diabetes mellitus in future. The computer aided diagnostic (CAD) system can be developed in future to help the clinicians to predict the diabetes mellitus at the earliest.

CONCLUSION

In our research, the innovative naive bayes and logistic model tree classifiers were used to classify the healthy and diabetic subjects using a PIMA dataset. Overall, the innovative naive bayes classifier has achieved a higher accuracy rate as 64.77% than the logistic model tree classifier (62.5%).

DECLARATIONS

Conflict of Interests

The authors declare no potential conflict of interest.

Authors Contributions

Author VK was involved in data collection, data analysis, manuscript writing. Author UT was involved in conceptualization, data validation, and critical review of manuscript.

Acknowledgement

The authors would like to thank the Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences (formerly known as Saveetha University) management for providing us all the necessary facilities to complete this project successfully.

Funding

We thank the following organizations for providing financial support that enabled us to complete the study.

1. Healthminds Pvt, Ltd, Karnataka.
2. Saveetha University.
3. Saveetha Institute of Medical and Technical Sciences.
4. Saveetha School of Engineering.

REFERENCES

1. American Diabetes Association. 2014. “Diagnosis and Classification of Diabetes Mellitus.” *Diabetes Care* 37 Suppl 1 (January): S81–90.
2. Appavu, Prabhu, Venkata Ramanan M, Jayaprabakar Jayaraman, and Harish Venu. 2021. “NOx Emission Reduction Techniques in Biodiesel-Fuelled CI Engine: A Review.” *Australian Journal of Mechanical Engineering* 19 (2): 210–20.
3. Arun Prakash, V. R., J. Francis Xavier, G. Ramesh, T. Maridurai, K. Siva Kumar, and R. Blessing Sam Raj. 2020. “Mechanical, Thermal and Fatigue Behaviour of Surface-Treated Novel Caryota Urens Fibre–reinforced Epoxy Composite.” *Biomass Conversion and Biorefinery*, August. https://doi.org/10.1007/s13399-020-00938-0.
4. Balachandar, Ramalingam, Logalakshmanan Baskaran, Ananthanarayanan Yuvaraj, Ramasundaram Thangaraj, Ramasamy Subbaiya, Balasubramani Ravindran, Soon Woong Chang, and Natchimuthu Karmegam. 2020. “Enriched Pressmud Vermicompost Production with Green Manure Plants Using Eudrilus Eugeniae.” *Bioresource Technology* 299 (March): 122578.
5. Bilous, R., and R. Donnelly. 2010. “Handbook of Diabetes. 4th Edn West Sussex.” Blackwell Publishing.
6. Burges, Christopher J. C. 1998. “A Tutorial on Support Vector Machines for Pattern Recognition.” *Data Mining and Knowledge Discovery* 2 (2): 121–67.
7. Ezhilarasan, Devaraj, Thangavelu Lakshmi, Manoharan Subha, Veeraiyan Deepak Nallasamy, and Subramanian Raghunandhakumar. 2021. “The Ambiguous Role of Sirtuins in Head and Neck Squamous Cell Carcinoma.” *Oral Diseases*, February. https://doi.org/10.1111/odi.13798.
8. Gopalakrishnan, R., V. M. Sounthararajan, A. Mohan, and M. Tholkapiyan. 2020. “The Strength and Durability of Fly Ash and Quarry Dust Light Weight Foam Concrete.” *Materials Today: Proceedings* 22 (January): 1117–24.
9. Hannah R, Pratibha Ramani, WM Tilakaratne, Gheena Sukumaran, Abilasha Ramasubramanian, and Reshma Poothakulath Krishnan. 2021. “Author Response for ‘Critical Appraisal of Different Triggering Pathways for the Pathobiology of Pemphigus vulgaris—A Review.’” Wiley. https://doi.org/10.1111/odi.13937/v2/response1.
10. Hassanein, Mohamed, Monira Al-Arouj, Osama Hamdy, Wan Mohamad Wan Bebakar, Abdul Jabbar, Abdulrazzaq Al-Madani, Wasim Hanif, et al. 2017. “Diabetes and Ramadan: Practical Guidelines.” *Diabetes Research and Clinical Practice* 126 (April): 303–16.
11. Jiang, L., H. Zhang, Z. Cai, and J. Su. 2005. “Evolutional Naive Bayes.” In *Proceedings of the 2005*

International Symposium on Intelligent Computation and Its Application, ISICA, 344–50.
12. Kavarthapu, Avinash, and Kaarthikeyan Gurumoorthy. 2021. "Linking Chronic Periodontitis and Oral Cancer: A Review." *Oral Oncology*, June, 105375.
13. King, Hilary, Ronald E. Aubert, and William H. Herman. 1998. "Global Burden of Diabetes, 1995–2025: Prevalence, Numerical Estimates, and Projections." *Diabetes Care* 21 (9): 1414–31.
14. Kumari, V. Anuja, and R. Chitra. 2013. "Classification of Diabetes Disease Using Support Vector Machine." *International Journal of Engineering Research and Applications* 3 (2): 1797–1801.
15. Menon, Soumya, Happy Agarwal, S. Rajeshkumar, P. Jacquline Rosy, and Venkat Kumar Shanmugam. 2020. "Investigating the Antimicrobial Activities of the Biosynthesized Selenium Nanoparticles and Its Statistical Analysis." *BioNanoScience* 10 (1): 122–35.
16. Muthukrishnan, Sivaprakash, Haribabu Krishnaswamy, Sathish Thanikodi, Dinesh Sundaresan, and Vijayan Venkatraman. 2020. "Support Vector Machine for Modelling and Simulation of Heat Exchangers." *Thermal Science* 24 (1 Part B): 499–503.
17. Ning Wang, and Guixia Kang. 2012. "A Monitoring System for Type 2 Diabetes Mellitus." In *2012 IEEE 14th International Conference on E-Health Networking, Applications and Services (Healthcom)*, 62–67. ieeexplore.ieee.org.
18. Nurjahan, Mohammad Abu Tareq Rony, Md Shahriare Satu, and Md Whaiduzzaman. 2021. "Mining Significant Features of Diabetes through Employing Various Classification Methods." In *2021 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD)*, 240–44.
19. Patra, Radhanath, and Bonomali Khuntia. 2021. "Analysis and Prediction Of Pima Indian Diabetes Dataset Using SDKNN Classifier Technique." *IOP Conference Series: Materials Science and Engineering* 1070 (1): 012059.
20. Rahman, Rashedur M., Farhana Afroz, and Others. 2013. "Comparison of Various Classification Techniques Using Different Data Mining Tools for Diabetes Diagnosis." *Journal of Software Engineering and Applications* 6 (03): 85.
21. Rosenbauer, J., P. Herzig, R. von Kries, A. Neu, and G. Giani. 1999. "Temporal, Seasonal, and Geographical Incidence Patterns of Type I Diabetes Mellitus in Children under 5 Years of Age in Germany." *Diabetologia* 42 (9): 1055–59.
22. Sarode, Sachin C., Shailesh Gondivkar, Gargi S. Sarode, Amol Gadbail, and Monal Yuwanati. 2021. "Hybrid Oral Potentially Malignant Disorder: A Neglected Fact in Oral Submucous Fibrosis." *Oral Oncology*, June, 105390.
23. Sarwar, M. A., N. Kamal, and W. Hamid. 2018. "Prediction of Diabetes Using Machine Learning Algorithms in Healthcare." *On Automation and …*. https://ieeexplore.ieee.org/abstract/document/8748992/.
24. Sekar, Durairaj, Deepak Nallaswamy, and Ganesh Lakshmanan. 2020. "Decoding the Functional Role of Long Noncoding RNAs (lncRNAs) in Hypertension Progression." *Hypertension Research: Official Journal of the Japanese Society of Hypertension*.
25. Thaiyalnayaki, K. 2021. "Classification of Diabetes Using Deep Learning and SVM Techniques." *International Journal of Current Research and Review* 13 (01): 146.
26. Thulasi, K. S., E. S. Ninu, and K. K. M. Shiva. 2017. "Classification of Diabetic Patients Records Using Naïve Bayes Classifier." In *2017 2nd IEEE International Conference on Recent Trends in Electronics, Information Communication Technology (RTEICT)*, 1194–98.
27. Webb, Geoffrey I., Janice R. Boughton, and Zhihai Wang. 2005. "Not So Naive Bayes: Aggregating One-Dependence Estimators." *Machine Learning* 58 (1): 5–24.
28. Yu, Wei, Tiebin Liu, Rodolfo Valdez, Marta Gwinn, and Muin J. Khoury. 2010. "Application of Support Vector Machine Modeling for Prediction of Common Diseases: The Case of Diabetes and Pre-Diabetes." *BMC Medical Informatics and Decision Making* 10 (March): 16.

TABLES AND FIGURES

Table 1: Represents the confusion matrix of the LMT classifier.

	Healthy	**Diabetic**
Healthy	29	15
Diabetic	18	26

Table 2: Represents the confusion matrix of the NB classifier.

	Healthy	**Diabetic**
Healthy	27	17
Diabetic	14	30

Table 3: Comparison of the LMT and NB classifiers in terms of accuracy (%).

Classifiers	**Accuracy (%)**
LMT	62.5
NB	64.77

Table 4: Represents the group statistics description of the LMT and NB machine learning algorithms.

	Group	N	Mean	Std. Deviation	Std. Error Mean
Accuracy	LMT	44	61.4150	3.25725	.49105
	NB	44	64.6564	.70820	.10677

Table 5: Independent sample test for significance and standard error determination. P value is less than 0.05 considered to be statistically significant and 95% confidence intervals were calculated.

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Diff.	Std. Error Diff.	Lower	Upper
Accuracy	**Equal variances assumed**	18.14	.000	-6.45	86	.000	-3.24	.50	-4.24	-2.24
	Equal variances not assumed			-6.45	47	.000	-3.24	.50	-4.25	-2.23

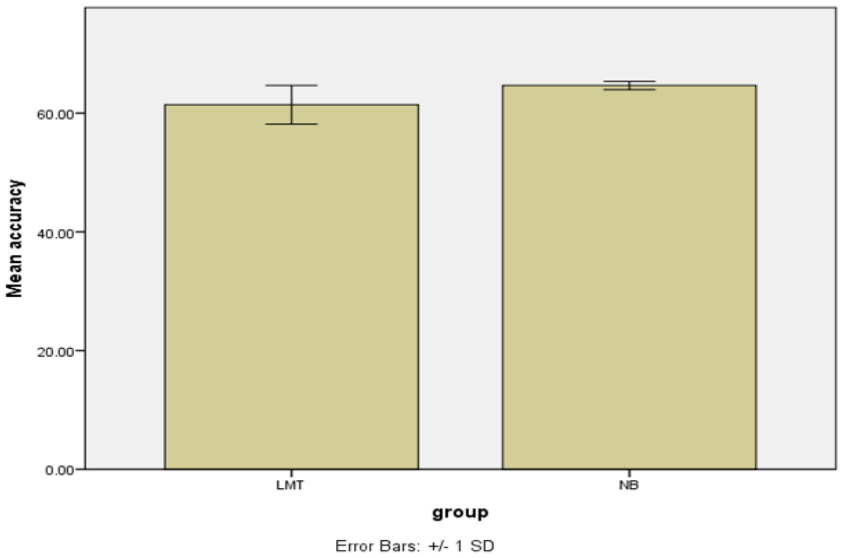

Fig. 1 Comparison of LMT and NB classifiers in terms of mean accuracy. The mean accuracy of NB is better than LMT and the standard deviation of LMT is slightly better than NB. X Axis: LMT classifier vs NB classifier; Y Axis: Mean accuracy of detection ± 1 SD.