VOLUME-8, ISSUE-6, JUNE-2019 • PRINT ISSN No. 2277 - 8160

Original Research Paper | Information Technology

DIABETES PREDICTION USING STACKING AND COST-SENSITIVE LEARNING

Ferhat Avdić International Burch University, Sarajevo, Bosnia and Herzegovina

Nejdet Doğru* International Burch University, Sarajevo, Bosnia and Herzegovina. *Corresponding Author

ABSTRACT According to the World Health Organization diabetes mellitus is the cause of millions of deaths around the world. Diabetes-related issues may be avoided when the patient is diagnosed and treated on time. Machine learning is used to create decision support systems in healthcare which aid in diagnosis. The Pima Indian dataset was used for patient classification. Algorithms such as Logistic Regression, Support Vector Machine, k-Nearest Neighbor, Naive Bayes, C4.5 Rules were used to create the prediction models. This study aimed to achieve a higher prediction accuracy for the onset of diabetes using a cost-sensitive stacking model. It performed better than the models in the literature, reaching approximately 85% accuracy, 86% specificity, and 85% sensitivity.

KEYWORDS : diabetes, decision support systems, data mining, stacking

INTRODUCTION

Diabetes is a chronic disease in which the levels of blood glucose in the body are too high due to the inability of the pancreas to produce enough insulin or to use it effectively. It is Statistics by the World Health Organization report that 8.5% of adults of age 18 and above had been diagnosed with diabetes by 2014. 1,6 million people have died due to diabetes in 2016, and another 2.2 million have died because of high blood glucose in 2012. [2]

Specialists use predictive analysis in health care mainly as help in the diagnosis of certain illnesses like diabetes, asthma, heart disease, and other lifelong diseases. Predictive analytics makes use of statistical techniques such as data mining, predictive modeling, and machine learning, that utilize gathered information on current or past events and conditions to predict the outcome of similar occasions in the future. [1][7]

This study is focused around improving the predictive model performance in diabetes diagnosis. It proposes a cost-sensitive stacking method that was more accurate compared to other singular algorithms and the ones from literature. It provides a platform for future advancements in decision support systems for diabetes diagnosis.

LITERATURE REVIEW

Diabetes prediction is a popular topic in machine learning related studies. Most of the literature is focused on building better performing prediction models. The studies used the Pima Indian dataset, performing different preprocessing methods and modifications on it such as replacing illogical values with the mean of the respective column[8][12][13], dropping the column entirely because of missing values [10][12], removing outliers [13], and feature selection [4][8]. In the literature, the performance of algorithms such as Decision Tree [8], the Ripper algorithm [10], C4.5 Rules [9], Artificial Neural Networks [4], Decision Stump [12] was analyzed and compared with other algorithms. Some studies used ensemble methods like majority voting [5], bagging [5] and boosting [5][12] to improve model performance. B. R. Prasad and S. Agarwal have performed feature engineering and discretization of column values to make algorithm decisions easier [8]. Kalaiselvi, C., and Nasira, G. M. used algorithms such as gradient descent and backpropagation to train the proposed model and an adaptive group based KNN for efficiency improvement. [4] Different tools were used to perform machine learning tasks for diabetes prediction such as Weka [10][12], Tanagra[10], Matlab [3][10], RStudio [13], and Keel [9]. The studies show that algorithm parameters also play a key role in training the model. All of them featured a certain train/test split for evaluation of their systems. The table below shows a summary of the methods used and accuracies achieved in previous studies:

TABLE 1 ACHIEVED ACCURACIES IN PREVIOUS STUDIES

Authors	Method	Accuracy (%)
B. R. Prasad and S. Agarwal	Decision Tree	83.95
H. Kaur and S. Batra	Simple Majority Voting	83.34
F. G. Woldemichael and S. Menaria	Back Propagation	83.11
Rahman, R. M., and Afroz, F.	Ripper (Jrip)	82.38
Purushottam, K. Saxena, and R. Sharma	C4.5 Rules	81.27
Jasim, I. S., Deniz Duru, A., Shaker, K., Abed, B. M., and Saleh, H. M.	Artificial Neural Network	80.86
V. V. Vijayan and C. Anjali	Decision Stump with AdaBoost	80.72
Kalaiselvi, C., and Nasira, G. M.	ANFIS with adaptive KNN	80.00

METHODOLOGY

The Pima Indian dataset was used to create machine learning models for diabetes prediction. It consists of 768 instances, out of which 268 were tested positive and 500 were tested negative for diabetes. One of the constraints of this dataset is that it consists of only females who were at least the age of 21.[11] The features of this dataset are described in the table below:

TABLE 2 PIMA INDIAN DATASET COLUMN DESCRIPTIONS

Column Name	Description	Value
Pregnancies	Number of times pregnant	Integer
Glucose	2 hour-load plasma glucose concentration	mg/dl
Blood Pressure	Diastolic blood pressure	mm/Hg
Skin Thickness	Triceps skinfold thickness	mm
Insulin	2-hour serum insulin	μU/ml.
BMI	Body mass index	kg/m2
Diabetes Pedigree Function	Calculated according to diabetes mellitus family history related to the subject	Real number
Age	Age in years	Integer

The non-logical values were replaced, discretized values were added, and normalization or standardization were left for preprocessing by default settings in singular Weka algorithms. The table below shows discretized values:

TABLE 3 VALUE DISCRETIZATION METHOD

Column	Condition	Value
Pregnancies	Was pregnant	True/False
Glucose	<=110	Normal
	>110 & <126	Impaired Fasting Glucose
	>=126 & <180	Impaired Glucose Tolerance
	>180	Diabetes
Blood Pressure	<=80	Normal
	>80 & <89	High
	>89	Risky
BMI	<=18.5	Underweight
	>18.5 & <25	Normal
	>=25 & <30	Overweight
	>=30	Obese
Insulin	>=16 & <66	Normal
	>=66 & <116	Medium
	>=116 & <166	High
	>166	Very High
Skin Thickness	>4 & <=10	Excellent
	>10 & <=14	Good
	>14 & <=20	Average
	>20 & <=25	Fair
	>25	Poor
Age	Ranges: 25-35; 35-45; 45-55; 55-65; >65	Particular ranges of values

The tool used in this study for performing machine learning is the Waikato Environment for Knowledge Analysis (Weka). Algorithms such as Logistic Regression, Support Vector Machine, Naïve Bayes, k-Nearest Neighbor, C4.5 Rules were used to train the models. They were compared to a proposed stacking method that included C4.5 Rules and Support Vector Machine.

Cost-sensitive learning was introduced to the training process to improve the prediction accuracy of the models. This method considers misclassification cost and handles different misclassification differently.[6] A cost-sensitive meta-classifier in Weka called Threshold-Selector was used for this purpose. It selects a mid-point threshold on the probability output of a learning algorithm to optimize a certain performance measure. This method is also known as thresholding.

The models were evaluated by their predictive accuracy, sensitivity, and specificity. As studies in the literature have used a train/test split for model validation, an 80:20 split ratio has been chosen in this study. This ratio provided 614 instances of training data for the model, and 154 test instances for validation.

RESULTS AND DISCUSSION

Table 4 shows the results for accuracy using five different single algorithm models and the proposed stacking method. Interestingly, Logistic Regression performed the same on the initial and modified dataset. All other algorithms show significant improvement in accuracy when used on the modified dataset. By using the cost-sensitive meta-classifier called Threshold-Selector in Weka, the results were improved even more across all models. The highest achieved accuracy is 84.42% by using the stacking method which featured C4.5 rules and SVM as base learners using Logistic Regression as the meta-learner. By minimizing the cost function this model achieved over 85% accuracy.

TABLE 4 ACHIEVED ACCURACIES BEFORE AND AFTER PREPROCESSING, AND AFTER SUBSEQUENTLY USING THRESHOLDING

Algorithm	Initial	Preprocessed	Thresholding
Logistic Regression	79.22	79.22	82.47
Support Vector Machine	77.92	79.87	82.47
K Nearest Neighbor	74.03	78.57	81.82
Naive Bayes	75.32	78.57	83.76
C4.5 Rules	76.62	79.22	79.87
Stacking	77.27	84.42	85.06

Table 5 shows the confusion matrix numbers obtained via models trained on the modified dataset and whose score threshold values on the probability outputs were modified by the Threshold-Selector, thus yielding slightly better results than in the experimental phase:

TABLE 5 CONFUSION MATRIX RESULTS PER ALGORITHM AFTER PREPROCESSING AND THRESHOLDING

Algorithm	Score Threshold	TP	FP	FN	TN
Stacking	0.5322	31	18	5	100
Logistic	0.5836	27	22	5	100
SVM	1.000	32	17	10	95
KNN	0.48	33	16	12	93
Naive Bayes	0.9536	27	22	3	102
C4.5 Rules	0.75	29	20	11	94

Accuracy, sensitivity, and specificity percentages were obtained through calculating ratios from the confusion matrices. Table 6 shows how stacking compares to singular algorithms in terms of performance. It is the most stable method among them, showing a balance between sensitivity and specificity, as well as the highest prediction accuracy. Naive Bayes has promising results as well, reaching as high as 90% in specificity while having slightly less accurate results than the stacking method. However, a model is evaluated by the sensitivity and specificity together as a balance between these two must be obtained to have a well-performing prediction model. Similarly, KNN has the highest score in specificity, precisely 85.32%. C4.5 rules gave the least accurate results, reaching less than 80%. Logistic Regression and SVM share the same accuracy results while they differ in sensitivity and specificity.

TABLE 6 MODEL PERFORMANCES PER ALGORITHM

Algorithm	Accuracy (%)	Sensitivity (%)	Specificity (%)
Stacking	85.06	86.11	84.75
Naive Bayes	83.77	90.00	82.26
Logistic	82.47	84.23	81.97
SVM	82.47	76.19	84.82
KNN	81.82	73.33	85.32
C4.5 Rules	79.87	72.50	82.46

CONCLUSIONS

Preprocessing the dataset and cost-sensitive learning improved results across all models. Combining Support Vector Machine and C4.5 Rules algorithms using stacking has proven to improve the predictive accuracy as well as sensitivity and specificity of the trained model. This method resulted in better model performance than any single algorithm scoring above 85% accuracy which is higher than the results in the literature. This sets a platform for further research in this area.

REFERENCES

[1] Eckerson, W. (n.d.). Predictive Analytics: Extending the Value of Your Data Warehousing Investment. Retrieved January 29, 2019, from http://www.bi-bestpractices.com/view-articles/5642

[2] Fact Sheets - Detail - Diabetes. (n.d.). Retrieved January 29, 2019, from World Health Organization website: https://www.who.int/news-room/fact-sheets/detail/diabetes

[3] Jasim, I. S., Deniz Duru, A., Shaker, K., Abed, B. M., & Saleh, H. M. (2017). Evaluation and measuring classifiers of diabetes diseases. 2017

International Conference on Engineering and Technology (ICET), 1–4. https://doi.org/10.1109/ICEngTechnol.2017.8308165

[4] Kalaiselvi, C., & Nasira, G. M. (2014). A New Approach for Diagnosis of Diabetes and Prediction of Cancer Using ANFIS. 2014 World Congress on Computing and Communication Technologies, 188–190. https://doi.org/10.1109/WCCCT.2014.66

[5] Kaur, H., & Batra, S. (2017). HPCC: An ensembled framework for the prediction of the onset of diabetes. 2017 4th International Conference on Signal Processing, Computing and Control (ISPCC), 216–222. https://doi.org/10.1109/ISPCC.2017.8269678

[6] Ling, C. X., & Sheng, V. S. (2008). Cost-Sensitive Learning and the Class Imbalance Problem. Springer, 8.

[7] Nyce, C. (n.d.). Predictive Analytics White Paper. 24.

[8] Prasad, B. R., & Agarwal, S. (2014). Modeling risk prediction of diabetes — A preventive measure. 2014 9th International Conference on Industrial and Information Systems (ICIIS), 1–6. https://doi.org/ 10.1109/ICIINFS. 2014.7036646

[9] Purushottam, Saxena, K., & Sharma, R. (2015). Diabetes mellitus prediction system evaluation using C4.5 rules and partial tree. 2015 4th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions), 1–6. https://doi.org/10.1109/ ICRITO. 2015. 7359272

[10] Rahman, R. M., & Afroz, F. (2013). Comparison of Various Classification Techniques Using Different Data Mining Tools for Diabetes Diagnosis. Journal of Software Engineering and Applications, 06, 85. https://doi.org/10.4236/jsea.2013.63013

[11] Smith, J. W., Everhart, J. E., Dickson, W. C., Knowler, W. C., & Johannes, R. S. (1988). Using the ADAP Learning Algorithm to Forecast the Onset of Diabetes Mellitus. Proceedings of the Annual Symposium on Computer Application in Medical Care, 261–265.

[12] Vijayan, V. V., & Anjali, C. (2015). Prediction and diagnosis of diabetes mellitus — A machine learning approach. 2015 IEEE Recent Advances in Intelligent Computational Systems (RAICS), 122–127. https://doi.org/ 10.1109/ RAICS. 2015.7488400

[13] Woldemichael, F. G., & Menaria, S. (2018). Prediction of Diabetes Using Data Mining Techniques. 2018 2nd International Conference on Trends in Electronics and Informatics (ICOEI), 414–418. https://doi.org/10.1109/ ICOEI. 2018. 8553959