2020 2nd International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI) | 978-1-7281-9638-1/20/$31.00 ©2020 IEEE | DOI: 10.1109/MLBDBI51377.2020.00037

Machine Learning in Predicting Diabetes in the Early Stage

Juncheng Ma
University of California Irvine
Irvine, CA, 92697, United States
junchenm@uci.edu

***Abstract*— Diabetes is a common disease and its early symptoms are not very noticeable, so an efficient method of prediction will help patients make a self-diagnosis. However, the conventional method to identify diabetes is to make a blood glucose test by doctors and the medical resource is limited. Therefore, most patients cannot get the diagnosis immediately. Since the early symptoms of diabetes are not obvious and the relationship between symptoms and diabetes is complex, the self-diagnosis results based on patients' own experience are not accurate. The process of Machine Learning is to train a computational algorithm for prediction based on a big dataset. It is popular for its efficiency and accuracy. Also, it has the advantage of dealing with tons of data, so we can make diagnoses for plenty of patients in a short time and the result will be more accurate. In this study, we used six classical machine learning models, including logistic regression, support vector machine, decision tree, random forest, boosting and neural network, to make a prediction model for diabetes diagnosis. Our data was from UCI Machine Learning Repository, which was collected by direct questionnaires from the patients of the Sylhet Diabetes Hospital in Sylhet, Bangladesh and approved by a doctor. We conduct parameter tuning on each model to tradeoff between the accuracy and complexity. The testing error shows that random forest, boosting and neural network had better performances than logistic regression, support vector machine and decision tree. The accuracy of neural network of the test dataset achieves 96 percent, which is the best model among these models for predicting diabetes.**

Keywords-component; Machine learning; predication;

I. INTRODUCTION

Diabetes is a very common disease. According to the National Diabetes Statics Report [1], as of 2015, 30.3 million people in the United States had diabetes, which means one of ten people in the United States is suffering from diabetes. In addition, one in ten of them do not even know they have the disease. It is also a chronic disease affecting the quality of life negatively, since most patients must deal with diabetes every day and it can lead to problems affecting almost every body systems. Hence, it is necessary to prevent and diagnose the disease.

An accurate and timely diagnosis will help patients prevent the diabetes, and it helps the patients find out whether they get diabetes in the early stage. However, the medical resource is limited, and doctors can only make diagnoses for certain number of patients in the limited time. Therefore, most people make an assessment based on their experience and symptoms. However, most patients lack professional medical knowledge, and they are just based on what they know and what they hear so it is inaccurate for patients to make diagnoses for themselves. Hence, it is necessary to make an efficient prediction model, which can save medical resources and help patients make a self-test accurately.

There are also some existed models for diabetes prediction. El Jerjawi et al [2] established a neural network model for diabetes prediction. They used some attributes such as PG Concentration (Plasma glucose at 2 hours in an oral glucose tolerance test), Diastolic BP (Diastolic Blood Pressure (mm Hg)). Most of them need some professional medical test, so it is not accessible for every person. Also, the final accuracy of the prediction is around 87 percent. Peter W.F et al [3] also uses the regression models to make predictions for diabetes mellitus. However, in its samples, 99 percent of them are white and non-Hispanic which does not include other races. Hence, we believe the model is not representative.

The objective of our study is to make machine learning [4] models to predict diabetes. The process of machine learning is like the computer algorithms are learning through experience instead of human, which means that they are much more efficient than humans. Also, machine learning is more and more popular nowadays. It can easily handle problems in high dimensional space. Logistic regression [5, 6] is to model the linear relationship between the binary dependent variable and independent variables based on logistic transformation. Support vector machine [7] (SVM) maps all the examples into high dimensional space and split the samples by a clear gap which is as wide as possible, and each side presents one category. Decision tree [8,9] is a tree-like structure. Each brunch represents different outcomes. Both SVM and decision tree can handle the non-linear relationship. Random forest [10] and Boosting [11] are two representative ensemble learning [12] methods. Random forest is a way to average multiple decision trees and reduce its variance. Boosting is also a way to reduce the variance by re-weighting each sample during the training process. Neural network is made up of neurons and layers. The data will travel from the first layer to the last. In this article, we employ the above-mentioned six machine learning models to construct diabetes prediction model.

In our study, we choose the attributes such as sudden weight loss, obesity to construct our prediction model, which are more understandable and accessible. The patients do not need to do some medical tests, which makes our model more understandable and applicable. We construct diabetes prediction models based on the above-mentioned six machine learning models. We also compare their performance in terms

978-1-7281-9638-1/20/$31.00 ©2020 IEEE
DOI 10.1109/MLBDBI51377.2020.00037

of the testing error. As a result, we find out boosting and neural network has the best accuracy which is 95.5% and 96.1% respectively.

The reset of the article is organized as follows. In section 3, we describe the detailed information about data and conduct descriptive statistic to investigate the relationship of each attributes with the dependent variable. In section 4, we discuss each model separately and tune the parameters with test data to trade-off between the complexity and accuracy. Finally, we summarize the test accuracy for each machine learning models.

II. DATA DESCRIPTION

The data set we used is from UCI machine learning repository, which is public available at http://archive.ics.uci.edu/ml/datasets/Early+stage+diabetes+risk+prediction+dataset. It is collected from the patients of Sylhet Diabetes Hospital in Sylhet. There are 17 attributes and 520 instances totally. Detailed information about the variables are listed in Table 1. Only age is a numerical variable, and others are categorical variables. 320 samples are positive cases and 200 samples are negative cases. We partition the whole data set into 2 subsets, the train set containing 364 cases and the test set containing 156 cases. We normalize the age to 0~1 by min-max normalization.

TABLE I. THE VARIABLES IN THE DATA SET

Age	Polyphagia	partial paresis
Sex	Genital thrush	muscle stiffness
Polyuria	visual blurring	Alopecia
Polydipsia	Itching	Obesity
sudden weight loss	Irritability	Class
weakness	delayed healing	

We conduct the descriptive statistics for each attribute to investigate the relation between each attribute and diabetes. First, we plot the average age for each class in Figure 1, and we can see that the average age of people with diabetes is higher. For the categorical variables, we calculated morbidity for each class of every attribute and the results are summarized in Table 2. As a result, people have symptoms like delayed healing, Obesity, muscle stiffness, Itching, Polyphagia, Polyuria, sudden weight loss, Genital thrush, Irritability, Polydipsia, partial paresis, weakness, visual blurring are more likely to get diabetes, while symptoms like Alopecia , Itching imply lower probability.

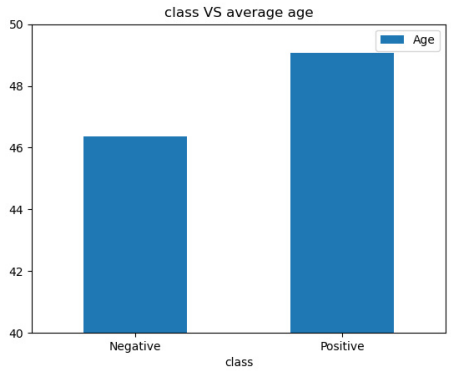

FIG. 1 THE AVERAGE AGE FOR POSITIVE AND NEGATIVE SAMPLES

TABLE II. MORBIDITY FOR EACH CLASS OF EVERY ATTRIBUTE

	Alopecia	**delayed healing**	**Gender**	**Itching**	**muscle stiffness**
No	71.0%	59.4%	90.1%	62.2%	56.9%
Yes	43.6%	64.0%	44.8%	60.9%	69.2%
	sudden weight loss	**Genital thrush**	**Irritability**	**partial paresis**	**Polydipsia**
No	43.6%	58.7%	53.3%	43.2%	33.1%
Yes	86.6%	71.6%	87.3%	85.7%	96.6%
	Obesity	**Polyphagia**	**Polyuria**	**visual blurring**	**weakness**
No	60.0%	46.3%	29.4%	50.5%	47.4%
Yes	69.3%	79.7%	94.2%	75.1%	71.5%

III. MODELS

In this section, we use logistic regression, support vector machine, decision tree, random forest, boosting and neural network to establish diabetes prediction models. We evaluate the model by calculation the train error and test error.

A. Logistic Regression

Logistic regression uses logistic functions to model the probabilities of possible outcomes. In a binary logistic regression model, the dependent variable has two categories. It converts the probability which is between 0 and 1 to any real number by function logit and finds out the linear relationship.

First, we need to find out the Logistic coefficients, and they are shown in Table 3. As we can see in the table, polyuria, polydipsia, sudden weight loss, weakness, polyphagia, genital thrush, visual blurring, Irritability and partial paresis are positively related with diabetes, while Age, itching, delayed healing, muscle stiffness, Alopecia and obesity are negatively related with diabetes, which is the same as the results in Table 2. The confusion matrix is summarized in Table 4. The train accuracy 87.6% and test accuracy 87.2% are close, which implies there is no overfitting [13].

TABLE III. COEFFICIENTS OF LOGISTIC REGRESSION.

Attribute	**Coefficient**	**Attribute**	**Coefficient**
intercept	-1.34	**visual**	0.76

		blurring	
Age	-0.35	**Itching**	-1.07
Gender	0.00	**Irritability**	1.02
Polyuria	2.45	**delayed healing**	-0.61
Polydipsia	2.02	**partial paresis**	1.25
sudden weight loss	0.79	**muscle stiffness**	-0.56
weakness	0.48	**Alopecia**	-0.54
Polyphagia	0.39	**Obesity**	-0.38
Genital thrush	0.81		

TABLE IV. THE CONFUSION MATRIX FOR LOGISTIC REGRESSION

	Train	
	Negative Predication	***Positive Predication***
Negative True	86.80%	13.20%
Positive True	11.80%	88.20%
	Test	
	Negative Predication	***Positive Predication***
Negative True	87.50%	12.50%
Positive True	13.00%	87.00%

B. Support Vector Machine

A support vector machine builds a hyperplane or set of hyperplanes in high dimensional space and it maps all the examples in a map and split the samples by a clear gap which is as wide as possible, and each side presents one category.

In this method, we must adjust the regularization parameter C, which controls the complexity of the model. Bigger C means stronger penalty on the misclassification, which makes the model is more likely overfitting. Hence, we use different C from 0.1 to 4.6, and we calculate the training and testing error for each model with different C, and the results are plotted in Figure 2. As the model becomes more complex, the train error continues to fall, and the test error first goes down and then goes up. Therefore, we choose the best parameter for our model, which is 2.1, and we get the corresponding confusion matrix which is summarized in Table 5. As a result, support vector machine has a better testing accuracy (90.4%) than logistic regression (87.2%).

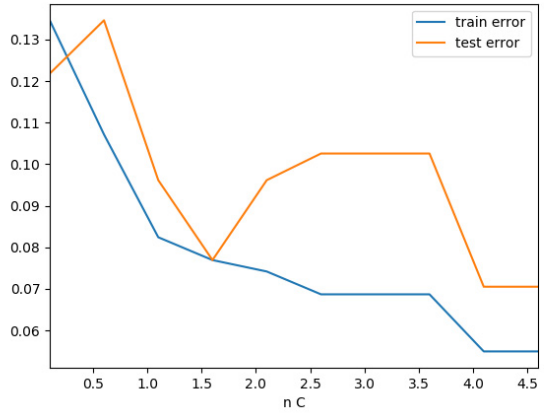

FIG. 2 THE TRAINING AND TESTING ERROR FOR EACH SVM MODEL WITH DIFFERENT TUNING PARAMETERS.

TABLE V. THE CONFUSION MATRIX FOR SVM

	Train	
	Negative Predication	***Positive Predication***
Negative True	94.10%	5.90%
Positive True	8.30%	91.70%
	Test	
	Negative Predication	***Positive Predication***
Negative True	92.20%	7.80%
Positive True	10.90%	89.10%

C. Decision Tree

Decision tree is a tree-like structure to create a model that make predictions based on input variables, and it is built by splitting the source set. The model classifies examples by a set of splitting rules based on the classification features. Decision tree is one of the most popular machine learning algorithms which has a feature of intelligibility and simplicity.

In this model, we use min_samples_split which means the minimum number of samples required to split an internal node to control the complexity of the model. As the minimum number becomes smaller, the model becomes more complex. We use different parameter from 2 to 29 to build the model and calculate train error and test error which is presented in Figure 3. The trend is like the support vector machine. Hence, we choose 9 as min_samples_split, and the corresponding confusion matrix is shown in Table 6. The testing accuracy is 93.6, which is much better than SVM and logistic regression, which is reasonable since SVM and logistic regression are designed for linear relationship modeling, while the decision tree can deal with nonlinear relationship.

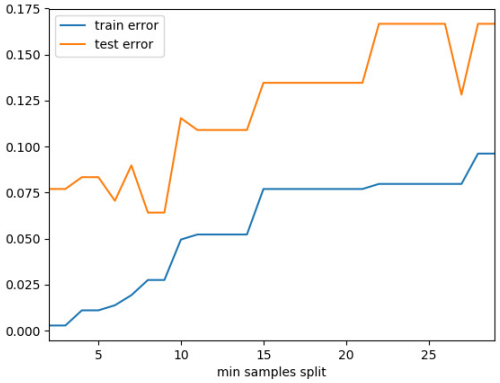

FIG. 3 THE TRAINING AND TESTING ERROR FOR EACH DECISION TREE WITH DIFFERENT TUNING PARAMETERS.

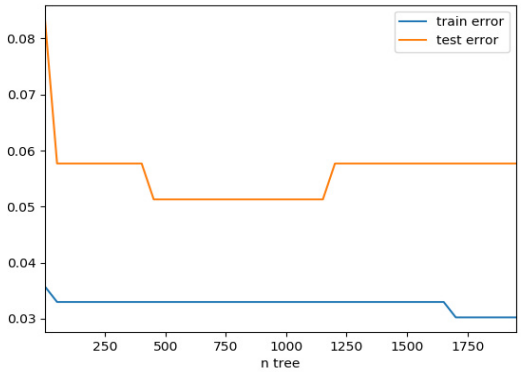

FIG. 4 THE TRAINING AND TESTING ERROR FOR RANDOM FOREST WITH DIFFERENT NUMBER OF TREES

TABLE VI. THE CONFUSION MATRIX FOR DECISION TREE

	Train	
	Negative Predication	***Positive Predication***
Negative True	97.80%	2.20%
Positive True	3.10%	96.90%
	Test	
	Negative Predication	***Positive Predication***
Negative True	90.60%	9.40%
Positive True	4.30%	95.70%

TABLE VII. THE CONFUSION MATRIX FOR RANDOM FOREST.

	Train	
	Negative Predication	***Positive Predication***
Negative True	100.00%	0.00%
Positive True	5.30%	94.70%
	Test	
	Negative Predication	***Positive Predication***
Negative True	93.80%	6.30%
Positive True	4.30%	95.70%

D. Random Forest

Due to the instability of single decision tree, it comes out the ensemble learning which combines multiple models to improve the prediction accuracy overall and reduce the variance. Random forest is a typical ensemble learning method, and the number of decision trees controls the complexity of the model, so we try different number of trees and calculate their errors which is plotted in Figure 4. Different from SVM and decision tree, testing error of random forest is not likely going up with larger number of trees, which implies random forest is less likely to over fit. Even though the training error continues to go down, the testing error keeps stable after falling, which means it is not likely to be an overfitting model relatively. The confusion matrix is summarized in Table 7. Compared with single decision tree, the testing accuracy of random forest, which is 94.9%, increased near 1.7 percent.

E. Boosting

Boosting is another classical ensemble learning method, and it reduces the variance by re-weighting the samples during the training process for each decision tree. Like the random forest, we also use the number of decision tree to determine the complexity of the model and the result is presented in Figure 5. Compared to the random forest, boosting is relatively more likely overfitting. The train error goes down first and then becomes a straight line. The test error goes down and then goes up like other models. According to the figure, we choose 700 trees to build our boosting model, and the confusion matrix is shown in Table 8. As we can see the train accuracy is almost 100 percent, which proves that the boosting is more likely to be overfitting. Compared to the random forest, its accuracy of boosting is further improved by around 0.7 percent.

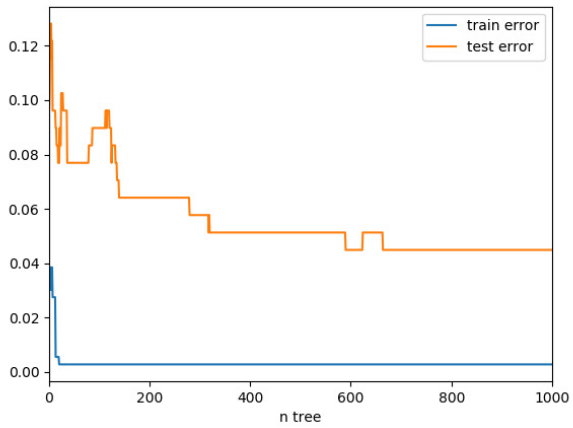

FIG. 5 THE TRAINING AND TESTING ERROR FOR BOOSTING WITH DIFFERENT NUMBER OF TREES

TABLE VIII. THE CONFUSION MATRIX FOR BOOSTING

	Train	
	Negative Predication	***Positive Predication***
Negative True	99.30%	0.70%
Positive True	0.00%	100.00%
	Test	
	Negative Predication	***Positive Predication***
Negative True	92.20%	7.80%
Positive True	2.20%	97.80%

F. Neural Network

Neural network is made up of layers and neurons, and it learns by processing samples from first layer to the last layer. Once enough samples have been processed, the algorithm can construct the right model. The more samples being processed, the more accurate the model is.

We designed three layers, and each one has 64,16,8 neurons respectively. We also choose the activation function which is logistic and learning rate which is 0.01. Then we get the confusion matrix in Table 9. As we can see, the neural network has the best performance with accuracy 96.2, which is better than all other models, because it has more parameters than others which can fit the data better.

Due to the instability of single decision tree, it comes out the ensemble learning which combines multiple models to improve the prediction accuracy overall and reduce the variance. Random forest is a typical ensemble learning method, and the number of decision trees controls the complexity of the model, so we try different number of trees and calculate their errors which is plotted in Figure 4. Different from SVM and decision tree, testing error of random forest is not likely goes up with larger number of trees, which implies random forest is less likely to over fit. Even though the training error continues to go down, the testing error keeps stable after falling, which means it is not likely to be an overfitting model relatively. The confusion matrix is summarized in Table 7. Compared with single decision tree, the testing accuracy of random forest, which is 94.9%, increased near 1.7 percent.

TABLE IX. THE CONFUSION MATRIX FOR NEURAL NETWORK

	Train	
	Negative Predication	***Positive Predication***
Negative True	97.80%	2.20%
Positive True	1.80%	98.20%
	Test	
	Negative Predication	***Positive Predication***
Negative True	93.80%	6.30%
Positive True	2.20%	97.80%

IV. CONCLUSION

An efficient diabetes prediction model will help doctors make accurate diagnoses and help patients get timely treatment. We conduct descriptive statistics for diabetes risk prediction dataset to investigate the variables which influence the diabetes. We build diabetes prediction models based on six machine learning models including logistic regression, support vector machine, decision tree, random forest, boosting and neural network, and their train and test error are listed in Table 10. For the first three models, logistic regression, SVM and decision tree are simple and intuitional, and it has lower accuracy than the last three models which are more complex. As we can see, the more complex the model is, the test accuracy is higher. In the Future, we can try models which have better learning and adaptive ability and use more variety of datasets to improve the prediction accuracy.

TABLE X. TRAIN AND TEST ERROR OF ALL MODELS

	Train	**Test**
Logistic	87.6%	87.2%
SVM	92.6%	90.4%
Decision Tree	97.3%	93.6%
Random Forest	96.7%	94.9%
Boosting	99.7%	95.5%
Neural Network	98.1%	96.2%

REFERENCES

[1] Centers for Disease Control and Prevention. National diabetes statistics report, 2017. Centers for Disease Control and Prevention website. www.cdc.gov/diabetes/pdfs/data/statistics/national-diabetes-statistics-report.pdf External link (PDF, 1.3 MB) . Updated July, 18 2017. Accessed August 1, 2017.

[2] El_Jerjawi, Nesreen Samer & Abu-Naser, Samy S. (2018). Diabetes Prediction Using Artificial Neural Network. International Journal of Advanced Science and Technology 121:54-64.

[3] Wilson PWF, Meigs JB, Sullivan L, Fox CS, Nathan DM, D'Agostino RB. Prediction of Incident Diabetes Mellitus in Middle-aged Adults: The Framingham Offspring Study. Arch Intern Med. 2007;167(10):1068–1074. doi:10.1001/archinte.167.10.1068

[4] Alpayd, Ethem. Introduction to Machine Learning, 2nd ed.[J]. Methods Mol Biol, 2014, 1107(1107):105-128.

[5] Tolles J, Meurer W J. Logistic Regression: Relating Patient Characteristics to Outcomes[J]. JAMA The Journal of the American Medical Association, 2016, 316(5):533.

[6] Cramer, J. S. (2003), Logit Models from Economics and Other Fields, Cambridge University Press, p. 13, ISBN 9781139438193..

[7] "1.4. Support Vector Machines — scikit-learn 0.20.2 documentation". Archived from the original on 2017-11-08. Retrieved 2017-11-08.

[8] Shalev-Shwartz, Shai; Ben-David, Shai (2014). "18. Decision Trees". Understanding Machine Learning. Cambridge University Press.

[9] Wu, Xindong; Kumar, Vipin; Ross Quinlan, J.; Ghosh, Joydeep; Yang, Qiang; Motoda, Hiroshi; McLachlan, Geoffrey J.; Ng, Angus; Liu, Bing; Yu, Philip S.; Zhou, Zhi-Hua (2008-01-01). "Top 10 algorithms in data mining". Knowledge and Information Systems. 14 (1): 1–37. doi:10.1007/s10115-007-0114-2. ISSN 0219-3116. S2CID 2367747.

[10] Liaw, A, Wiener, M, Liaw, A. Classification and Regression with Random Forest[J]. R News, 2002, 23(23).

[11] Kégl, Balázs. The return of AdaBoost.MH: multi-class Hamming trees[J]. Computer Science, 2013.

[12] Krogh A, Sollich P. Statistical mechanics of ensemble learning[J]. Phys.rev.e, 1997, 55(1).

[13] Piryonesi S. Madeh; El-Diraby Tamer E. (2020-03-01). "Data Analytics in Asset Management: Cost-Effective Prediction of the Pavement Condition Index". Journal of Infrastructure Systems. 26 (1): 04019036. doi:10.1061/(ASCE)IS.1943-555X.0000512.