2019 HAWAII UNIVERSITY INTERNATIONAL CONFERENCES
SCIENCE, TECHNOLOGY & ENGINEERING, ARTS, MATHEMATICS & EDUCATION JUNE 5 - 7, 2019
HAWAII PRINCE HOTEL WAIKIKI, HONOLULU, HAWAII

STATISTICAL DATA MINING ALGORITHMS FOR THE PROGNOSIS OF DIABETES AND AUTISM

MARIANI, MARIA C. ET AL
UNIVERSITY OF TEXAS AT EL PASO
EL PASO, TEXAS

Dr. Maria C. Mariani
Department of Mathematical Sciences and Computational Science Program
University of Texas at El Paso
El Paso, Texas

Mr. Md Al Masum Bhuiyan
Mr. Osei K. Tweneboah
Computational Science Program
University of Texas at El Paso
El Paso, Texas

Statistical Data Mining Algorithms for the Prognosis of Diabetes and Autism

Synopsis:

The early detection of these diseases could help the prognosis and chance of survival significantly. This manuscript is devoted to the application of Machine Learning (ML) technique to Diabetes and Autism disease data. Several important variables that cause diabetes and autism disease are studied in this work. We propose three supervised machine learning (ML) techniques, which can predict with great accuracy the likelihood of diabetes and autism in patients. These techniques allow computers to learn and to order the important variables that causes the diseases. We predict the test data based on the key variables and compute the prediction accuracy using the Receiver Operating Characteristic (ROC) curve to train a good classifier. The results suggest that the ML techniques are effective in classifying the patients regarding diabetes and autism disorder. Similar methodology can also be applied to other diseases such as Cancer and Heart Disease data.

Statistical Data Mining Algorithms for the Prognosis of Diabetes and Autism

Maria C. Mariani[*], Md Al Masum Bhuiyan[†]and Osei K. Tweneboah[‡]

Abstract

This paper studies three statistical data mining algorithms used in the diagnosis and prognosis of diabetes and autism spectrum disorder (ASD). The persistent hyperglycemia of diabetes results in the damage, malfunction, and failure of different vital organs such as heart, kidneys, eyes, and several others. So, the early detection of diabetes is very important for timely treatment. Autism spectrum is a neurological disorder characterized by limited social interaction and restricted repetitive behaviors. The early intervention significantly improves the quality of life of people with ASD. In this study, we propose three supervised machine learning (ML) techniques, which can prognosticate with great accuracy the likelihood of diabetes and autism in patients. These techniques allow computers to learn and to order the important variables that causes the diseases. The techniques have been used on the basis of the correlation of variables in the data. We predict the test data based on the key variables and compute the prediction accuracy using the Receiver Operating Characteristic (ROC) curve to train a good classifier. The results suggest that the ML techniques are effective in classifying the patients regarding diabetes and autism disorder.

Keywords: Diabetes, Autism, Statistical Data Mining Algorithms, ROC Curve.

[*]Department of Mathematical Sciences and Computational Science Program, University of Texas at El Paso.

[†]Computational Science Program, University of Texas at El Paso.

[‡]Computational Science Program, University of Texas at El Paso.

1 Introduction

Machine learning (ML) techniques help to identify and discriminate by certain chronological disorders and diseases such as cancer, heart disease, tumor, diabetes, autism and so on. The methodologies are useful to many statistical and optimization processes that allow computers to learn past observations and to identify the pattern of the disease. The reason is that the ML techniques help us to identify the sources of the disease and also orders the variables in terms of importance in causing chronological disease.

Diabetes is a chronological disease which occurs when the blood glucose level rises in the circulatory system and results in stroke, blindness, kidney failure, and premature death. According to the World Health Organization report in 2018, 400 million adults live with diabetes across the planet and it caused the death of 1.6 million people in 2015 [1]. The factors responsible for diabetes are the consumption of extra carbohydrate, highly processed food, overweight, obesity, and bad consumption habits. In 2017, the National Diabetes Statistic Report [2] for Center Disease Control and Prevention (CDC) tell us that about 30.3 million people suffer from diabetes. According to this report, 23.1 million are analyzed and 7.2 million patients are undiscovered in the United States [3]. Different modeling techniques have been used to diagnose diabetes based on its key variables [4]. Boyle et al. worked on the duality analysis to predict the diabetes population in 2050 [5]. To obtain greater accuracy, Upadhyay and Patel proposed a fuzzy classifier model to classify the diabetes condition in patients [6]. Other researchers have also worked on different machine learning techniques like Artificial Intelligent, Linear Discriminant Analysis, and Neural Network techniques to increase diagnostic accuracy and to reduce costs and human resources [7], [8], and [9].

Autism Spectrum Disorder (ASD) is a brain development disorder linked to genetic and neurological factors that limits certain communication and social behaviors from natural growth. It is mainly diagnosed by in terms of behavioral indicators such as social interaction, imaginative ability, repetitive behaviors, and communication. The study conducted by Wiggin et al. found that 33% of children with problems other than ASD have some ASD symptoms while not meeting its full classification criteria [10]. The number of children afflicted with ASD has grown considerably over the last four decades. It is now estimated that this disease has affected $\sim 1.5\%$ of children [11]. However, it is a heterogeneous disorder that depends on behavioral features, age, gender, congenital jaundice, pervasive developmental disorder for any immediate family member, etiology, treatment response, and so on. Although ASD is currently diagnosed and treated with psychometric tools, the high prevalence rate of ASD have led some researchers to turn to machine learning intelligent methods in preference to traditional statistical methods [12] -[15]. For instance, Ecker et al. used support vector machine technique and had 81% classification accuracy of whole-brain structural volumetric changes in ASD based on cross-validation procedure [16].

A notable feature of classification problem is that it orders the important variables. In the present work, we propose three supervised ML techniques: Logistic regression, Random Forest, and Support vector machine for diabetes and Autism prognosis. We use the open source databases through openML benchmark for Diabetes [17] and UCI machine learning repository for Autism spectrum disorder [18]. We study the correlation of all variables in the data, which are very effective in capturing the characteristics and patterns of data correctly [19]. However, the problem of machine learning techniques is to fit it into the data using their tuning parameters. In this case, we use the cross-validation to estimate the tuning parameters of each model corresponding to the low mean squared error, which are discussed in the later sections. We determined the adequacy and predictive ability of the data by computing sensitivity, specificity, accuracy, and confidence interval using ROC curve.

The paper is organized as follows: section 2 offers a brief background of supervised machine learning techniques. The regularization techniques to obtain the best predictive model with test data are also discussed. Section 3 is devoted to a brief description of our data. We then analyze the correlation among the variables of data in section 4. Section 5 deals with the results when the machine learning techniques are applied to the autism and diabetes disease data. We explain the important variables in the data that play important roles to cause these diseases. Finally, section 6 contains the conclusion and discusses the suitability of our models to diagnose the diabetes and autism disorder.

2 Methodology

This section describes the three ML techniques that will be used to classify the diabetes and autism data. We will discuss the regularization techniques to find the best model and important features in data. The tuning parameters of the models are also estimated to obtain the lowest mean squared error, lowest miss-classification rate, and higher predictive accuracy.

2.1 Logistic Regression Model

Logistic Regression is a powerful classification algorithm that is used to predict the probability of a categorical variable. We assume that the predictors (x_k) are independent of each other, so the model has little or no multi-collinearity. We express the model as:

$$\text{logit}(p) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_k x_k, \tag{1}$$

where p is the probability of presence of the characteristic of interest; $\beta_0, \beta_1, \cdot\cdot\cdot\cdot \beta_k$ are the coefficient parameters. The logit transformation is defined as the logged odds as: $ln\frac{p}{1-p}$. So the logistic regression model is similar to a linear regression, but it is constructed using the natural logarithm of the "odds" of target variable. We say:

$$ln\frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 +_3 x_3 + \cdots + \beta_k x_k \tag{2}$$

Since logistic regression predicts probabilities, rather than just classes, we can fit it into the data using the likelihood technique. The likelihood helps to find the best model that explains the dataset well. The dataset used in this study contains a vector of features (x_i) and an observed class (y_i). We assume the probability of that class is either p, when $y_i = 1$, or $1 - p$, when $y_i = 0$. So the likelihood function of Eq.(2) is as follows:

$$L(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k) = \prod_{i=1}^{N} p(x_i)^{y_i}(1 - p(x_i)^{1-y_i}). \tag{3}$$

We now seek to estimate the parameters $\beta_0, \beta_1, \cdots \beta_k$ that maximize the likelihood function $L(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)$ in Eq. (3). At this point, it is mathematically convenient to maximize the logarithm of likelihood function [20] as follows:

$$l(\beta) = \sum_{i=1}^{N}(y_i\beta^T x_i - log(1 + e^{\beta^T x_i}). \tag{4}$$

We then use the regularization technique to obtain a parsimonious model with important features from the original model. The regularization technique penalizes the magnitude of coefficients of features to minimize the error between predicted and actual observations. In this study, we use the L_1 regularization is used.

2.1.1 L_1 regularization

We use the lasso-regularization technique by adding a L_1 penalty term in Eq. (4). It forces the sum of the absolute value of the regression coefficients to be less than a fixed value. This is due to the fact that the tuning parameter makes certain coefficients to be set to zero, effectively by choosing a simpler model that does not include those coefficients [21]. So we maximize the penalized versions as follows:

$$l_\lambda(\beta) = \sum_{i=1}^{N}(y_i\beta^T x_i - log(1 + e^{\beta^T x_i})) - \lambda\sum_{j=1}^{p}|\beta_j| \tag{5}$$

where λ is a tuning parameter that controls the strength of penalty term. So we select λ in a way that the resulting model minimizes the out of sample error.

2.2 Random Forest

The random forest technique [25] is a type of additive model that predicts the data by combining decisions from a sequence of base models. It reduces the variance by avoiding overfitting of the model. The class of base models can be expressed as follows:

$$g(x) = f_0(x) + f_1(x) + f_(x) + \cdots \tag{6}$$

where the final model g is the sum of simple base models f_i. We define each base classifier as a simple decision tree. So it is an ensemble technique that considers multiple learning algorithms to obtain best predictive model.

At this point, all the base models or trees are made independently using a different subsample of the data. Once we have a new generated training set, we divide it randomly into two parts. The two-third samples are used to build a tree and the one-third samples are used to obtain the predictions of trees. We take the majority vote of these one-third predictions as the predicted value for the data point and then we estimate the error. For a comprehensive study of Random Forest, please see [24].
We now present the algorithm of random forest that is used in this study:

1. We first take a random sample of size N with replacement from the data.

2. Take a random sample without replacement of the predictors.

3. Construct a split by using predictors selected in step 2.

4. Repeat steps 2 and 3 for each subsequent split until the tree is as large as desired.

5. Drop the out-of-bag data down the tree. We then store the class assigned to each observation along with each observation's predictor values.

6. Repeat steps 1-5 a large number of times.

7. For each observation in the dataset, we count the number of trees that it is classified in one category over the number of trees.

8. Assign each observation to a final category by a majority vote over the set of trees. Thus, if 51% of the time over a large number of trees a given observation is classified as a "1", that becomes its classification.

So the random forest include three main tuning parameters such as node size, number of trees (ntree), and number of predictors sampled (mtry) for splitting. To build a best predictive model, we estimate the best tuning parameters and important variables using mean decrease accuracy (MDA) and mean decrease Gini (MDG) indices. The MDA determines the importance of a variable by measuring the change in prediction accuracy, when the values of the variable are randomly permuted compared to the original observations. However, the MDG index is a measure of how each variable contributes to the homogeneity of the nodes and leaves in the resulting random forest. For the details of these methodologies, consult [25] and references therein.

2.3 Support Vector Machine (SVM)

SVM is a powerful supervised ML technique that is used for both classification and regression tasks. It identifies the optimal decision boundary that separates data points from different classes, and then it predicts the class of test observations using the separation boundary [26]. So we first normalize and scale the variables to obtain a decision boundary as follows:

$$y_i(\beta^T x_i + \beta_0) \geq M(1 - \xi_i) \text{ for, } i = 1, \cdots, n. \tag{7}$$

which is a non-convex programming problem. Here, $\xi_i \geq 0$ measures the degree of misclassification of the i-th individual. So it computes the distance from a wrongly classified observation x_i to its corresponding margin. We now consider the problem as:

$$\max_{\beta,\beta_0,M} M; \text{ subject to } ||\beta|| = 1, \text{ and } y_i(\beta^T x_i + \beta_0) \geq M(1 - \xi_i) \tag{8}$$

where $\sum_{i=1}^{n} \xi_i \leq C$ and the constant C determines the possible miss-classification rate. The higher the value of C, the less likely it is that the SVM algorithm will misclassify a point. So in this study our approach is to tune the parameter C.

Now we drop the constraint $||\beta|| = 1$ and set $M = 1/||\beta||$ in the problem and we obtain:

$$\begin{aligned} &\min_{\beta,\beta_0,} \frac{1}{2}||\beta||^2 + C.\sum_{i=1}^{n} \xi_i, \\ &\text{s.t. } \xi_i \geq 0 \text{ and } y_i(\beta^T x_i + \beta_0) \geq 1 - \xi_i \end{aligned}$$

It is now a convex programming problem and can be solved via Lagrangian and KKT techniques as follows:

$$\mathcal{L}(\beta, \beta_0, \xi, \alpha, \gamma) = \frac{1}{2}||\beta||^2 + C\sum_{i=1}^{n} \xi_i - \sum_{i=1}^{n} \alpha_i\{y_i(\beta^T x_i + \beta_0) - (1 - \xi_i)\} - \sum_{i=1}^{n} \gamma_i \xi_i. \tag{9}$$

Setting $\frac{\partial \mathcal{L}}{\partial \beta}, \frac{\partial \mathcal{L}}{\partial \beta_0}, \frac{\partial \mathcal{L}}{\partial \xi_i}$ to be 0 gives the following conditions:

$$\hat{\beta} = \sum_{i=1}^{n} \alpha_i y_i x_i, \ \sum_{i=1} \alpha_i y_i = 0, \text{ and } \alpha_i + \gamma_i = C, \forall i$$

where, $\alpha_i \geq 0, \gamma_i \geq 0$, and $\xi_i \geq 0$. We can solve the Lagrangian problem via coordinate descent method. Now the KKT complementary slackness condition yields the following case:

$$\begin{cases} \text{If } \alpha_i = 0, \; y_i(\beta^T x_i + \hat{\beta}_0) \geq 1 \\ \text{If } \alpha_i = C, \; y_i(\beta^T x_i + \hat{\beta}_0) \leq 1 \\ \text{If } 0 < \alpha_i < C, \; y_i(\beta^T x_i + \beta_0) = 1 \end{cases}$$

In this study, we compute the SVM for different values of C and choose the optimal C that maximizes the accuracy. We use the caret package in R software to compute the SVM kernel technique. For more details of the SVM methodology please see [27].

3 Data background

In this study, we used the diabetes and autism disorder datasets, that are available in openML repository and UCI machine learning repository, respectively. The diabetes dataset has 768 observations with 9 variables. Table 1 summarizes the short description of the diabetes dataset. The "*Class*" column is the response variable that includes the status of a testing diabetes as "*Yes*" or "*No*". Our object is to predict the "*Class*" variable and to conclude whether a patient has diabetes based on the important variables.

For the autism data, we used ten behavioral features (AQ-10-Child) with six individuals characteristics that detect the ASD cases from controls in behaviour science. Table 3 summarizes the short description of behavioral features and individual characteristics used in this study. The target variable is "*Autism*" that includes the status of autism disorder as "*Yes*" or "*No*". Our objective is to predict the "*autism*" variable and to conclude whether an individual has autism based on the behavioral features.

Variables	Features
preg	Number of times pregnant.
plas	Plasma glucose concentration (2 hours in an oral glucose tolerance test)
pres	Diastolic blood pressure (mm Hg)
skin	Triceps skin fold thickness (mm)
insu	2-Hour serum insulin (μ U/ml)
mass	Body mass index (weight in $kg/(\text{height in } m)^2$)
pedi	Diabetes pedigree function
age	Age (years)
Class	testing diabetes (0 or 1)

Table 1: Background of Diabetes data

Table 2: Background of Autism data

Variables	Description of Features
q1	I often notice small sounds when others do not.
q2	I usually concentrate more on the whole picture, rather than the small details
q3	I find it easy to do more than one thing at once
q4	If there is an interruption, I can switch back to what I was doing very quickly
q5	I find it easy to 'read between the lines' when someone is talking to me
q6	I know how to tell if someone listening to me is getting bored
q7	When I'm reading a story I find it difficult to work out the character's intentions
q8	I like to collect information about categories of things.
q9	I find it easy to work out what someone is thinking or feeling just by looking at their face
q10	I find it difficult to work out people's intentions
score	total scores of the behaviors mentioned above
age	Age in years
gender	Male or Female
jaundice	Whether the case was born with jaundice
pdd	Whether any immediate family member has a PDD.
screening	Whether the user has used a screening app.

4 Exploratory Data Analysis

This subsection deals with the correlation among the variables of data. We used “Hmisc” library in R program to determine the correlation among predictors. Figures 1 and 2 represent the correlation among the variables for diabetes and autism data, respectively. The size of circles in these figures indicates the strength of the relationship among the variables. In Figure 1, we observe that the variables *preg* (number of times pregnant) and *age* are highly positively correlated. On the other hand, the variables *insu* (2-hour serum insulin) and *age* are negatively correlated, which is not very strong. In Figure 2, we see that the variables *score* (total score of behavioral features) and *autism* are positively correlated. The correlation among other variables are also shown in these two figures.

5 Results & Discussion

This section presents the results and applications of machine learning (ML) techniques when applied to the diabetes and autism datasets. We trained three ML models to accurately predict whether an individual has been diagnosed with diabetes or Autism spectrum disorder. We randomly split the data into training set (67% for building a predictive model) and test set (33% for evaluating the model). We then compute the prediction mean squared error (PMSE), missclassification rate (MCR), and prediction accuracy using the ROC curve that validates the fitted model with data. An R statistical module was developed for the analyses.

5.1 Analysis of Fitted models

We first trained the logistic regression technique to build a predictive model. In this case, we used the lasso regularization with L_1 penalty and obtained the tuning parameter λ with cross-validation. The L_1 penalty is used for both variable selection and shrinkage, since it has the effect of forcing some of the coefficient estimates to be zero. Table 2 and 3 represent important predictors using the best predictive model with L_1 penalty. It is clear that *preg*, *plas*, *mass*, *pedi*, and *age* for diabetes (Table 2); and behavioral features *q6*, *q9*, and *score* for autism (see Table 3) are important predictors. We then predict the test data using this predictive model and compute the accuracy.

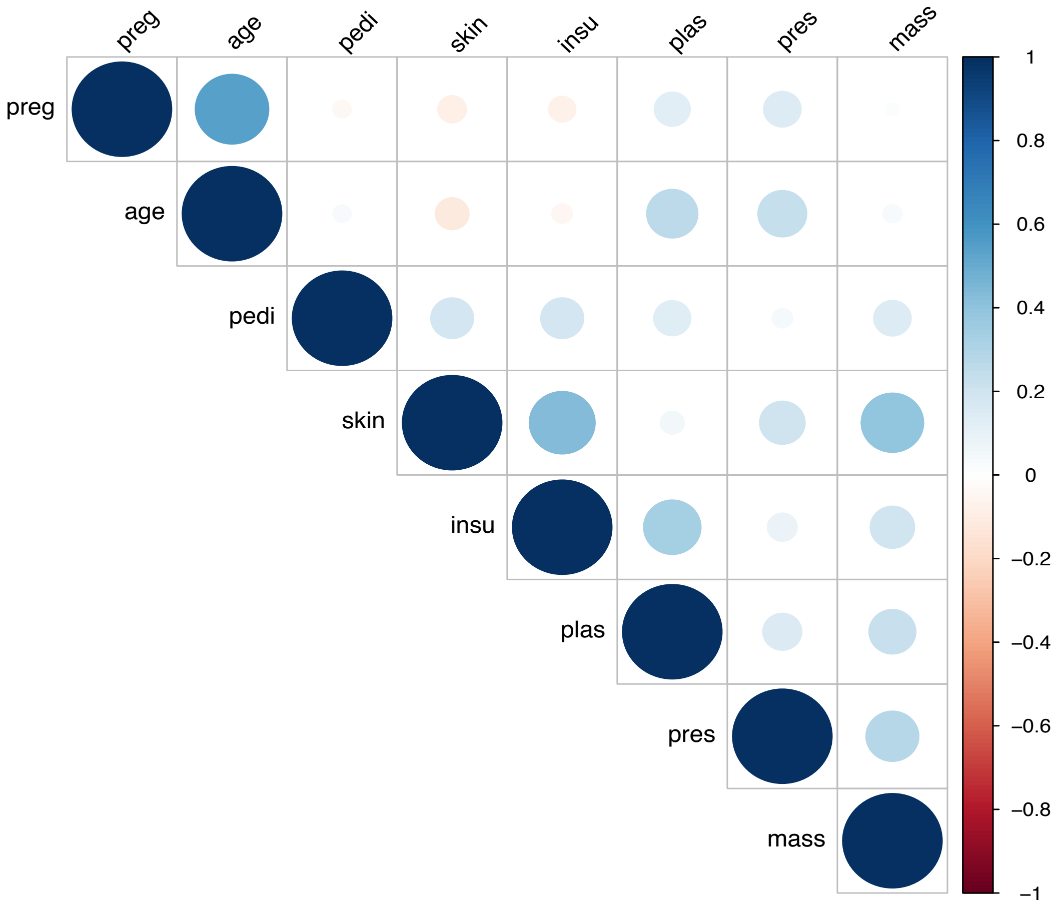


Figure 1: Correlation among the Diabetes variables.

Variables	Coefficients
preg	0.0323
plas	0.0250
pres	0
skin	0
insu	0
mass	0.0407
pedi	0.2950
age	0.0121

Table 2: Coefficients of important predictors using LGR (L_1) model for Diabetes data.

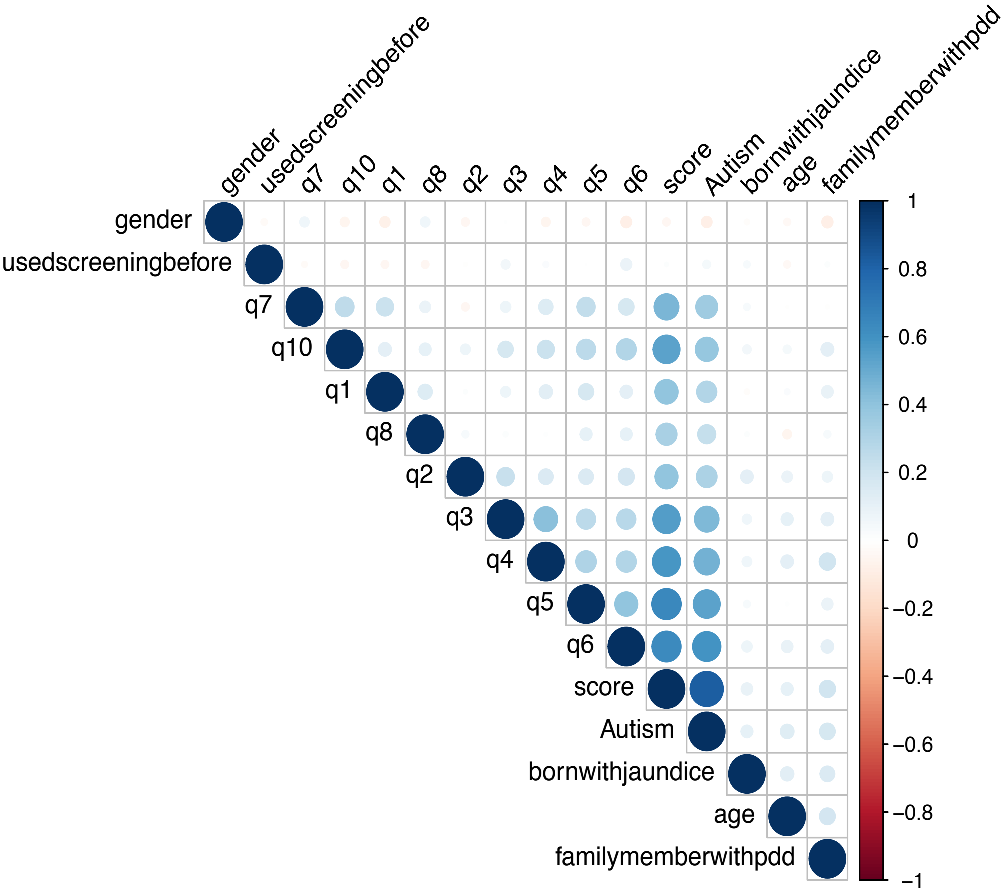


Figure 2: Correlation among the Autism variables.

Variables	Coefficients
q1	0
q2	0
q3	0
q4	0
q5	0
q6	0.02458
q7	0
q8	0
q9	0.18395
q10	0
score	2.66378
age	0
gender	0
bornwithjaundice	0
familywithpdd	0
usedscreeningbefore	0

Table 3: Coefficients of important predictors using LGR (L_1) model for Autism data.

We fitted the random forest model with train data to build a predictive model. In this case, we used 500 trees and sampled 3 variables at each split. We obtained a very good prediction accuracy on test data, which are 74.80% for diabetes disease, and 100.00% for autism disorder. An important feature of random forest is that we made an order of the predictors in terms of importance using Mean Decrease Accuracy and Mean Decrease Gini indices. From Fig. 3, it is clear that the plasma-glucose-concentration (*plas*) is the most important variable and triceps-skin-fold-thickness (*skin*) is the least important variable in causing diabetes. From Fig. 4, we see that the behavioral features *score* is the most important predictor in causing autism.

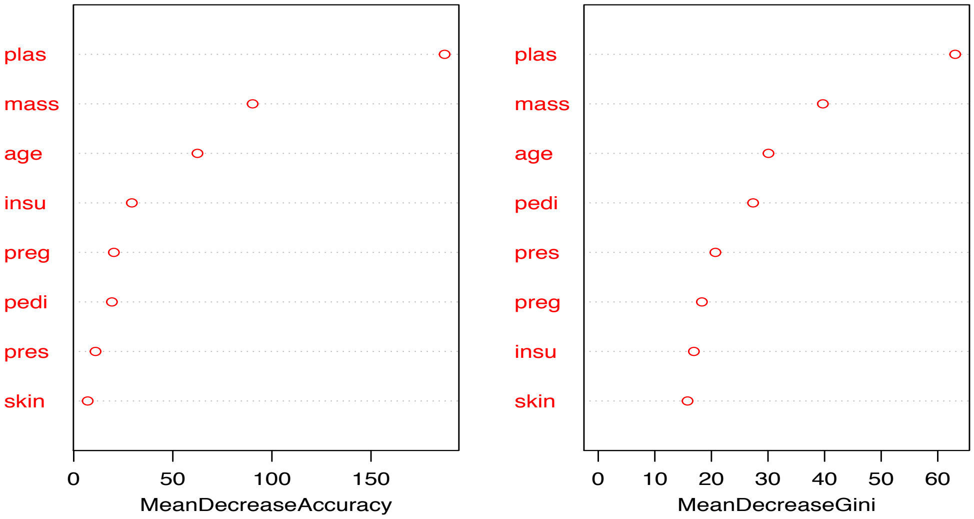


Figure 3: Variable importance plot using Random Forest model for Diabetes data.

Finally, we used the SVM kernal to classify the data used in this study. To fit the model, we first standardized the data and used 10-fold cross-validation to train the data. We evaluated different cost levels (C) to obtain the best predictive model with optimal cost. We achieved the highest accuracy when γ is 0.01 (diabetes data) and 0.1 (autism data). Table 4 summarizes the mean square error (pmse) for the predicted probabilities and missclassification rate (MCR) for both diabetes and autism data.

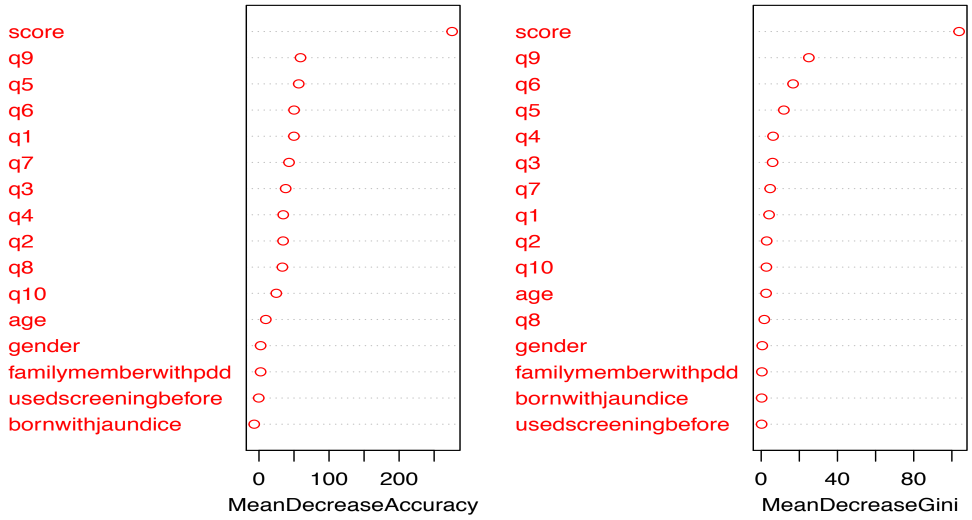


Figure 4: Variable importance plot using Random Forest model for Autism data.

Table 4: **Model Evaluation**

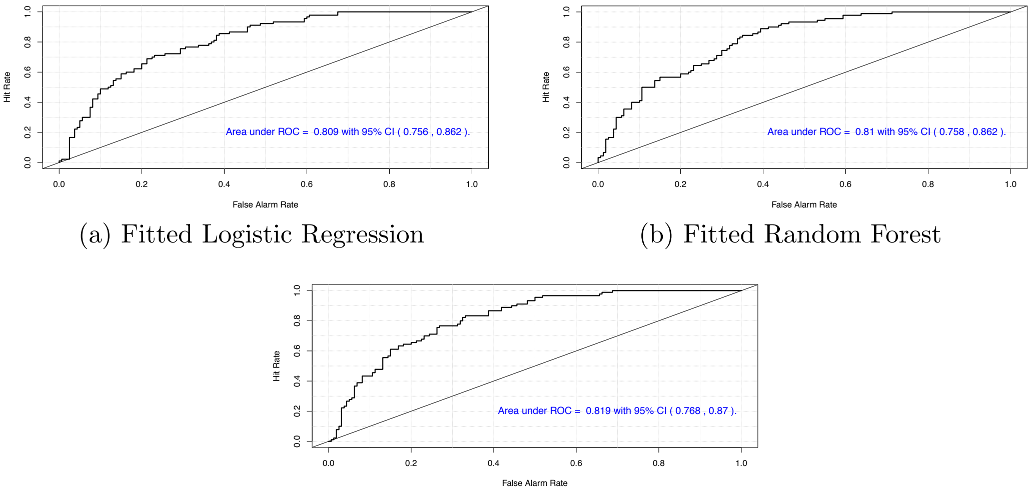
Models	Diabetes Dataset		Autism Dataset	
	PMSE	MCR	PMSE	MCR
LGR	0.173	0.256	0.007	0.000
RF	0.169	0.252	0.0018	0.000
SVM	0.169	0.24	0.0130	0.0172

5.2 Model Accuracy

This subsection presents the accuracy of our predictive models obtained in this study. Tables 5 & 6 show the sensitivity, specificity, accuracy, and confidence interval with 95% significance level for both diabetes and autism data. Here, the sensitivity of models measures the proportion of people with the disease (diabetes or autism) who will have a positive result. So the sensitive test correctly identifies patients with a disease. For example, the Logistic regression (L_1) test is 100% sensitive for autism data, that is, the model classifies all individuals who have the autism disorder (see Table 6). On the other hand, the specificity measures the proportion of people without the disease (diabetes or autism) who will have a negative result. For instance, the logistic regression (L_1) for diabetes test is 70% specific, meaning that the model identifies 70% of patients who do not have the diabetes (see Table 5). We plotted the ROC curve between the True Positive Rate (Y-axis) and the False Positive Rate (X-axis) of our predictive models (see Figs. 5 and 6). In these figures, the diagonal line represents the threshold (0.5) of the ROC curve. We see that the areas under the curve are almost 0.8 for all models in diabetes data. However for the autism data, the area under the ROC curve is 1 for Logistic regression (L_1) and Random Forest models. Thus we conclude that ML techniques have good predictive ability on diabetes and autism data.

Table 5: Model Evaluation using ROC Curve for Diabetes data.

Models	Sensitivity (%)	Specificity (%)	Accuracy (%)	Conf. Interval (%)
LGR	75.81	70.31	74.44	(68.52 - 79.69)
RF	86.25	54.44	74.80	(68.94 - 80.06)
SVM	79.41	68.75	76.00	(70.21 - 81.16)



(a) Fitted Logistic Regression

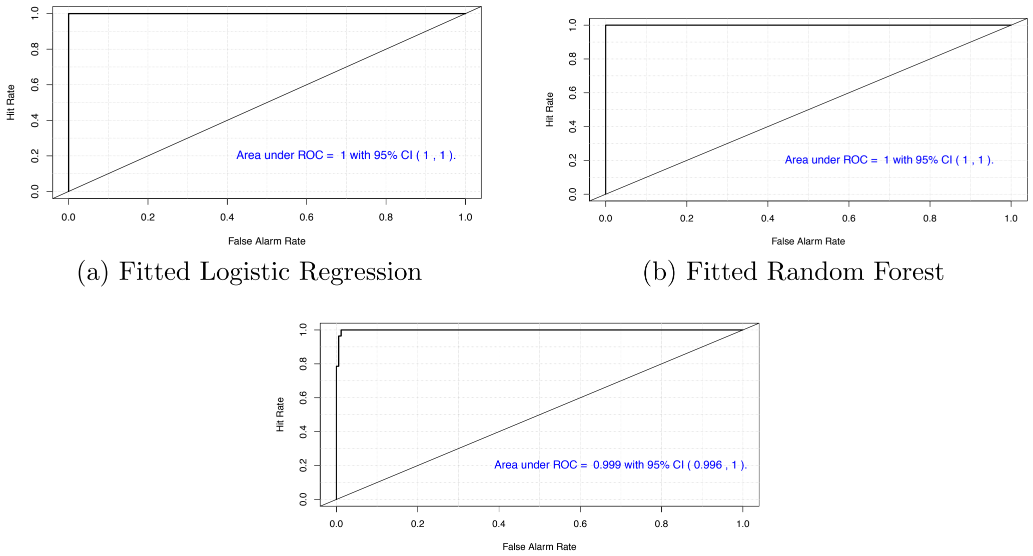
(b) Fitted Random Forest

(c) Fitted SVM model

Figure 5: Model Evaluation using ROC Curve for Diabetes data.

Table 6: Model Evaluation using ROC Curve for Autism data set.

Models	Sensitivity (%)	Specificity (%)	Accuracy (%)	Conf. Interval (%)
LGR	100.0	100.00	100.00	(98.42 - 100.0)
RF	100.00	100.00	100.00	(98.42 - 100.00)
SVM	98.86	96.43	98.28	(95.64 - 99.53)



(a) Fitted Logistic Regression

(b) Fitted Random Forest

(c) Fitted SVM model

Figure 6: Model Evaluation using ROC Curve for Autism data.

6 Conclusion

In this study, we have analyzed three different machine learning techniques namely; Logistic regression with L_1 regularization, Random Forest, and Support vector machine techniques. We used these algorithms to predict the presence of diabetes and autism spectrum disorder in people. We first did the exploratory data analysis to see the dynamics of the predictors and response variable. The exploratory data analysis provides correlation among the variables of disease data (see subsection 4). We then randomly split the data into training set (67% to build a predictive model) and test set (33% to evaluate the model).

We trained the logistic regression with L_1 penalty term and determined the important variables that are useful in predicting diabetes and autism disorder. We then analyzed the Random Forest model and thus obtained important predictors of these diseases. Regarding the SVM technique, we used the SVM kernel to obtain higher accuracy through several cost functions (see subsection 5.1).

Since these models had all the variables of data initially, we improved the models by removing the insignificant variables through cross-validation and different tuning parameters. In order to obtain the most efficient model possible, we compared the prediction mean squared error and miss-classification rate among these models (see Table 4). We also compared the prediction accuracy, sensitivity, specificity, and confidence interval to find the best classification accuracy possible (see Tables 5 and 6). We predicted 76.00% of the patients accurately regarding their diabetes with SVM kernel model. Similarly, 100.00% of individuals with autism spectrum disorder were predicted using the Logistic regression and the Random Forest model. Thus the ML techniques are effective in classifying patients regarding their diabetes and autism disorder. We also concluded that a patient suffering from diabetes are highly affected by the following variables; number of times pregnant (*preg*), plasma glucose concentration (*plas*), body mass index (*mass*), diabetes pedigree function (*pedi*), and age of patients (*age*). We also observed that the patient suffering from autism disorder is highly affected by the behavioral features (*q6*), (*q9*), and total scores (see Tables 2 - 3, and Figs. 3 - 4).

References

[1] Guidelines on second-and third-line medicines and type of insulin for the control of blood glucose levels in non-pregnant adults with diabetes mellitus, World Health Organization, 2018, ISBN: 9789241550284.

[2] Centres for Disease Control and Prevention, National Diabetes Statistics Report, Atlanta: Centers for Disease Control and Prevention, US Department of Health and Human Services, 2017.

[3] Introduction: Standards of Medical Care in Diabetes-2019 Diabetes Care, 2019; 42 (Supplement 1): S1-S2. https://doi.org/10.2337/dc19-Sint01.

[4] Meng, X. H., Huang, Y. X., Rao, D. P., Zhang, Q., & Liu, Q. (2013), Comparison of three data mining models for predicting diabetes or prediabetes by risk factors, *The Kaohsiung journal of medical sciences*, **29(2)**, 93-99.

[5] Boyle, J. P., Honeycutt, A. A., Narayan, K. V., Hoerger, T. J., Geiss, L. S., Chen, H., & Thompson, T. J. (2001). Projection of diabetes burden through 2050, Diabetes care, **24(11)**, 1936-1940.

[6] Upadhyay A., Patel V. R., (2016), Comparative Study - Prediction of Diabetes and Heart Disease using Data Mining Approaches, International Journal of Engineering Technology, Management and Applied Sciences, 4(1), 70-76

[7] Polat, K., ahan, S., & Gune, S., (2006), A new method to medical diagnosis: Artificial immune recognition system (AIRS) with fuzzy weighted pre-processing and application to ECG arrhythmia. Expert Systems with Applications, 31(2), 264-269.

[8] Dogantekin, E., Dogantekin, A., Avci, D., & Avci, L. (2010). An intelligent diagnosis system for diabetes on linear discriminant analysis and adaptive network based fuzzy inference system: LDA-ANFIS. Digital Signal Processing, 20(4), 1248-1255.

[9] Kala, R., Shukla, A., & Tiwari, R., (2009), Comparative analysis of intelligent hybrid systems for detection of PIMA indian diabetes. In Nature Biologically Inspired Computing, 2009, World Congress on (pp. 947-952), IEEE.

[10] Wiggins LD, Reynolds A, Rice CE, Moody EJ, Bernal P, Blaskey L, Rosenberg SA, Lee LC, Levy SE. Using 710 standardized diagnostic instruments to classify children with autism in the study to explore early development. J Autism Dev Disord. 2014a; **45(5)**,271-1280, doi:10.1007/s10803-014-2287-3.

[11] Howsmon DP, Kruger U, Melnyk S, James SJ, Hahn J (2017), Classification and adaptive behavior prediction of children with autism spectrum disorder based upon multivariate data analysis of markers of oxidative stress and DNA methylation, Plos Computational Biology, **13(3)**, e1005385, https://doi.org/10.1371/journal.pcbi.1005385

[12] Mythili M, Shanavas Mohamed R, A study on Autism spectrum disorders using classification techniques, Ijcsit, 2014, **5(6)**:7288-7291

[13] Pancers K, Derkacz A (2015), Consistency-based pre-processing for classification of data coming from evaluation sheets of subjects with ASDS, Federated conference on Computer Science and Information Systems, 63-67.

[14] Duda M, Ma R, Haber N, Wall DP, Use of machine learning for behavioral distinction of autism and ADHD, Transl Psychiatry, 2016, **9(6)**, 732, doi:10.1038/tp.2015.221.

[15] Bone D, Goodwin MS, Black MP, Lee -C-C, Audhkhasi K, Narayanan S, 2015, Applying machine learning to facilitate autism diagnostics: pitfalls and promises, *Journal of Autism and Developmental Disorders*,**45(5)**, 1-16.

[16] Ecker C., Rocha-Rego V., Johnston P., Mourao-Miranda J., Marquand A., Daly E.M., Brammer M.J., Murphy C., Murphy D.G., 2009, Investigating the predictive value of whole-brain structural MR scans in autism: a pattern classification approach NeuroImage, **49 (1)**, 44-56.

[17] https://www.openml.org/

[18] http://archive.ics.uci.edu/ml/datasets/Autism+Screening+Adult

[19] Mariani, M.C., Bhuiyan, M.A. Masum, Tweneboah, O.K., (2018), Estimation of stochastic volatility by using Ornstein-Uhlenbeck type models, *Physica A: Statistical Mechanics and its Applications*, **491**, 167-176.

[20] Mariani, M.C., Bhuiyan, M.A. Masum, Tweneboah, O.K., Huizar, H.G., Florescu I. (2018), Volatility models applied to geophysics and high frequency financial market data, *Physica A: Statistical Mechanics and its Applications*, **503**, 304-321.

[21] James, G., Witten, D., Hastie, T., TIbshirani., R. (2017), An Introduction to Statistical Learning with Applications in R, *Springer*, 337-349, ISBN:978-1-4614-7138-7.

[22] https://en.wikipedia.org/wiki/Tikhonovregularization.

[23] Jolliffe, I.T. (2002), Principal Component Analysis, Second edition, *Springer*, 167-195.

[24] Hastie T., TIbshirani, R., Friedman, J. (2008), The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd edition, *Springer*.

[25] Leo Breiman (2001), Random Forest, *Springer*, **45**, 5-32.

[26] James, G., Witten, D., Hastie, T., Tibshirani, R. (2013), An Introduction to Statistical Learning with Applications in R, *Springer*.

[27] https://docs.google.com/viewer?a=vpid=sitessrcid=ZGVmYXVsdGRvbWFpbnx1dGVwc3RhdDU0OTR8Z3g6MTI1ZTZhOWM2MDMwNzFhYQ.