Scientific Research Publishing

Journal of Data Analysis and Information Processing, 2024, 12, 462-478
https://www.scirp.org/journal/jdaip
ISSN Online: 2327-7203
ISSN Print: 2327-7211

A Machine Learning Classification Model for Detecting Prediabetes

A. K. M. Raquibul Bashar[1], Mahdi Goudarzi[2], Chris P. Tsokos[3]

[1]Department of Mathematics & Computer Science; Augustana College, Rock Island, Illinois, USA
[2]Independent Researcher, San Francisco, California, USA
[3]Department of Mathematics & Statistics, University of South Florida, Tampa, Florida, USA
Email: akmraquibulbashar@augustana.edu, m6goodarzi@gmail.com, ctsokos@usf.edu

How to cite this paper: Bashar, A.K.M.R., Goudarzi, M. and Tsokos, C.P. (2024) A Machine Learning Classification Model for Detecting Prediabetes. *Journal of Data Analysis and Information Processing*, **12**, 462-478.
https://doi.org/10.4236/jdaip.2024.123024

Received: July 12, 2024
Accepted: August 24, 2024
Published: August 27, 2024

Abstract

The incidence of prediabetes is in a dangerous condition in the USA. The likelihood of increasing chronic and complex health issues is very high if this stage of prediabetes is ignored. So, early detection of prediabetes conditions is critical to decrease or avoid type 2 diabetes and other health issues that come as a result of untreated and undiagnosed prediabetes condition. This study is done in order to detect the prediabetes condition with an artificial intelligence method. Data used for this study is collected from the Centers for Disease Control and Prevention's (CDC) survey conducted by the Division of Health and Nutrition Examination Surveys (DHANES). In this study, several machine learning algorithms are exploited and compared to determine the best algorithm based on Average Squared Error (ASE), Kolmogorov-Smirnov (Youden) scores, areas under the ROC and some other measures of the machine learning algorithm. Based on these scores, the champion model is selected, and Random Forest is the champion model with approximately 89% accuracy.

Keywords

Prediabetes, Machine Learning, SVM, Forest, Cumulative Lift

1. Introduction

Insulin is one of the many important hormones produced by our pancreas that works as an accelerator to break the blood sugar and process those sugars (glucose) in such a way so that micro-cells in our body can absorb those to produce energy and heat in the human body. If the cells in the body don't respond normally to insulin, then this state of sugar insulation is termed as the **Prediabetes** [1]. Approximately 84 million American Adults, more than 1 out of 3, have pre-

diabetes. Among those with prediabetes, about 90% don't know they have this hormonal condition. If this stage goes untreated, then there is an increased risk of developing type-2 diabetes, heart disease and stroke as per CDC [2]. In terms of preventing cardiovascular risk and preventing the strokes, it is very significant and important to detect the prevalence of prediabetes [3]. Also, the risk of cardiovascular disease and mortality is almost two times as high in individuals with a condition of prediabetes [4] [5]. Early detection, diagnosis, and intervention for prediabetes is highly desired preventive measure that can be taken by anyone to avoid all the complications, prevent the transition of state for individuals from prediabetes to other type of diabetes (type-2) and the model can be deployed to detect this condition with a very cost-effective way [6] [7].

In recent years, artificial intelligence research is used to quantify almost all areas of human intervention with disease diagnosis and treatment selection. Machine learning is one of the broad areas of artificial intelligence that uses statistical methods for data classification and clustering. There are a handful of machine learning techniques that have been utilized and applied in the clinical domain to predict any sort of disease condition and have implied higher accuracy for diagnosis rather than classical methods [8].

2. Methodology and Materials

2.1. Data Source

The National Center for Health Statistics (NCHS), Division of Health and Nutrition Examination Surveys (DHANES), part of the Centers for Disease Control and Prevention (CDC), has conducted a series of health and nutrition surveys since the early 1960's. The National Health and Nutrition Examination Surveys (NHANES) were conducted on a periodic basis from 1971 to 1994. In 1999, NHANES became continuous. Every year, approximately 5000 individuals of all ages are interviewed in their homes and complete the health examination component of the survey [9]. The sample size we have used in this study is 22,867 (*i.e.*: the number of individuals included in this study).

2.2. Data Description

In this dataset, Risk for prediabetes is the response variable and all other covariates are subdivided into different variable clusters based on the attributes of those variables such as, Demographic, Diet Behavior, Weight, Hight, Physical Activities, and Symptoms as shown in **Figure 1**. The NHANES sample represents the total non-institutionalized civilian U.S. population residing in the 50 states and District of Columbia. As with previous NHANES samples, a four-stage sample design was used in NHANES 2011-2014. The first stage consisted of selecting PSUs from a frame of all U.S. counties.

2.3. Risk Factors

At the beginning of variable inclusion in the machine learning algorithm, all the

attributes under the sub-cluster of covariates are taken into model feed and sequentially variables are ranked in the final machine learning model according to their importance determined by the relative importance calculated using the actual model. In this modeling, Age, Weight, Height, Poverty Ratio, and Blood pressures, are the continuous variables and rest of the attributes are categorical in variable measurements.

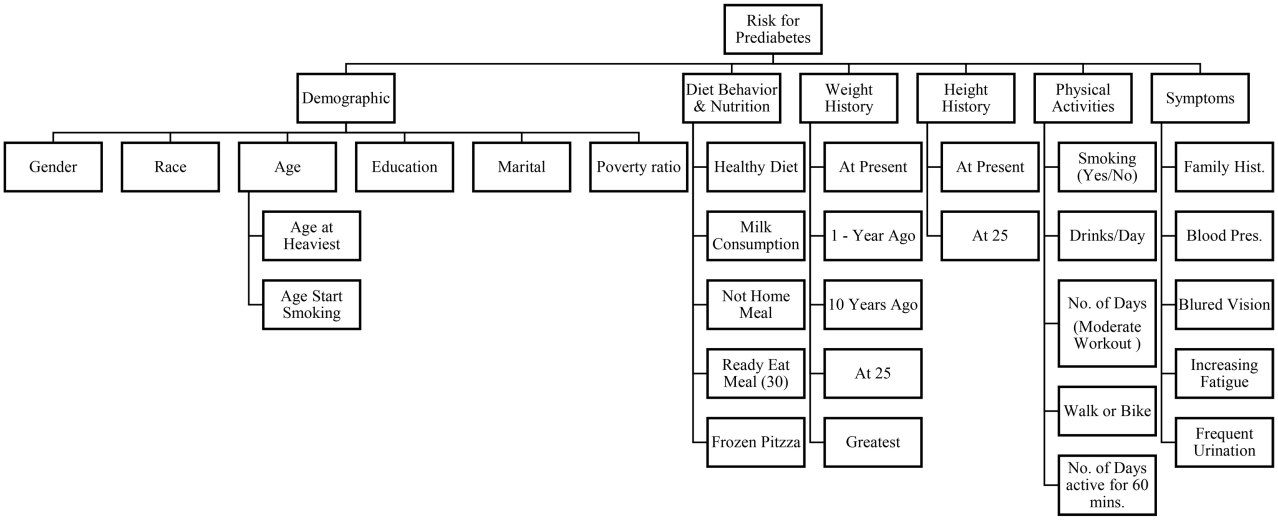

Figure 1. Schematic diagram of prediabetes data.

2.4. Machine Learning Modeling

In this study, we have used a supervised learning algorithm such as Decision Tree, Support Vector Machine (SVM), Gradient Boosting, Random Forest, Logistic Regression, and Neural Network. Also, for all the algorithms, all the observations were subdivided in 65% for **Training Set**, 25% for **Validation** and 10% for **Testing**. Then compared all the models according to their *Average Squared Error* (***ASE***), *Captured Response Percentage* (***CRP***) and *Areas Under* (***ROC***). The champion model is selected with the lowest value of ASE and the highest value of CRP and areas under ROC. In the following sections, some machine learning algorithms are discussed very briefly.

2.4.1. Decision Tree

The decision tree algorithm falls under the category of supervised learning [10]. They can be used to solve both regression and classification problems. The decision tree uses the tree representation to solve the problem in which each leaf node corresponds to a class label and attributes are represented on the internal node of the tree. We can represent any Boolean function on discrete attributes using the decision tree. While using the decision tree, there are some basic assumptions are made as follows:

- at the beginning, the whole training dateset is considered as the root
- featured values are preferred to be categorical
- on the basis of attributable values records are distributed recursively

• statistical methods are used for ordering attributes as root or internal node.

2.4.2. Support Vector Machine (SVM)

More formally, a support-vector machine [11] constructs a hyperplane or set of hyper-planes in a high- or infinite-dimensional space, which can be used for classification, regression, or other tasks like outliers detection. Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the nearest training-data point of any class (so-called functional margin), since in general the larger the margin, the lower the generalization error of the classifier.

2.4.3. Gradient Boosting

Gradient boosting [12] is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. It builds the model in a stage-wise fashion as other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function.

2.4.4. Forest Model

A forest is an ensemble of decision trees [13], each one able to predict its own response to a set of input variables. The results from the individual trees are combined to provide the final prediction. For a categorical target, the forest model's prediction is either the most popular class (as determined by a vote) or the average of the posterior probabilities of the individual trees. For an interval target, the forest model's prediction is the average of the estimates from the individual decision trees. In Model Studio, the forest algorithm uses the following process to build each tree:

1) The algorithm selects a sample of cases, with replacement, from the original training data.

2) Then, for each node, the algorithm selects a sample of input variables from all available inputs.

3) From this sample, the input that has the strongest association with the target is used in the splitting rule for that node.

Therefore, the method of selecting the input variable for a splitting rule is different for a forest than it is for the split-search process used to build an individual tree. Each tree is created on a different sample of the cases, and each splitting rule is based on a different sample of the inputs. This process ensures that the individual models in the ensemble are more varied. The process that the forest algorithm uses to build the individual trees and then combine the results of the predictions in a more stable model than a single tree. Training each tree with different data reduces the correlation of the predictions of the trees. This, in turn, is likely to improve the predictions of the forest as compared to the naïve method of using the same data to build all the trees in a forest. The forest algorithm also takes random samples of the inputs. Therefore, the trees in the forest use different combinations of cases and inputs to determine the splits. This additional perturbation (beyond bagging) leads to greater diversity in the trees and

often better predictive accuracy than with bagging. Each sample of the original training data that is selected to train a specific decision tree is called bagged data. For each tree in the forest, the data that are withheld from training are called an out-of-bag sample. Model assessment measures (such as miss-classification rates and average squared error) and iteration plots are constructed on both the entire training data set and the out-of-bag sample.

2.4.5. Artificial Neural Network (ANN)

An artificial neural network (ANN) [14] is a network of simple elements called artificial neurons, which receive input, change their internal state (activation) according to that input, and produce output depending on the input and activation.

An artificial neuron mimics the working of a biophysical neuron with inputs and outputs but is not a biological neuron model.

The network forms by connecting the output of certain neurons to the input of other neurons forming a directed, weighted graph. The weights, as well as the functions that compute the activation, can be modified by a process called learning which is governed by a learning rule.

2.5. Statistical Analyses

2.5.1. Ranking Variable Importance

In this study, we have taken 22 covariates at the initial consideration. But considering the 16 variables in total to build the machine learning model determined by TREE SPLIT procedure [15] [16]. It measures variable importance based on the following metrics:

- Count-based variable importance simply counts the number of times in the tree that a particular variable is used in a split.
- Surrogate-count-based variable importance tallies the number of times that a variable is used in a surrogate splitting rule.
- RSS-based variable importance measures variable importance based on the change of RSS when a split is found at a node.

Figure 2 shows the relative importance determined by this process. From that figure, it turns out that Age is the most important risk factor for predicting prediabetes.

2.5.2. Missing Data Imputation

Because of the fact that the data was collected from the survey and missing information is inherent characteristics for this data-set, we have used multiple imputation [17] method to impute missing information for each of the attributable variables considered in the model. We have imputed mean for the scale variables or continuous variables and mode for qualitative measures of categorical variables for imputation to the missing cells of information. Since we have missing systematic missing information without any pattern, multiple imputations have enabled us to impute missing values and use the complete case information in the data-set of the subject matter.

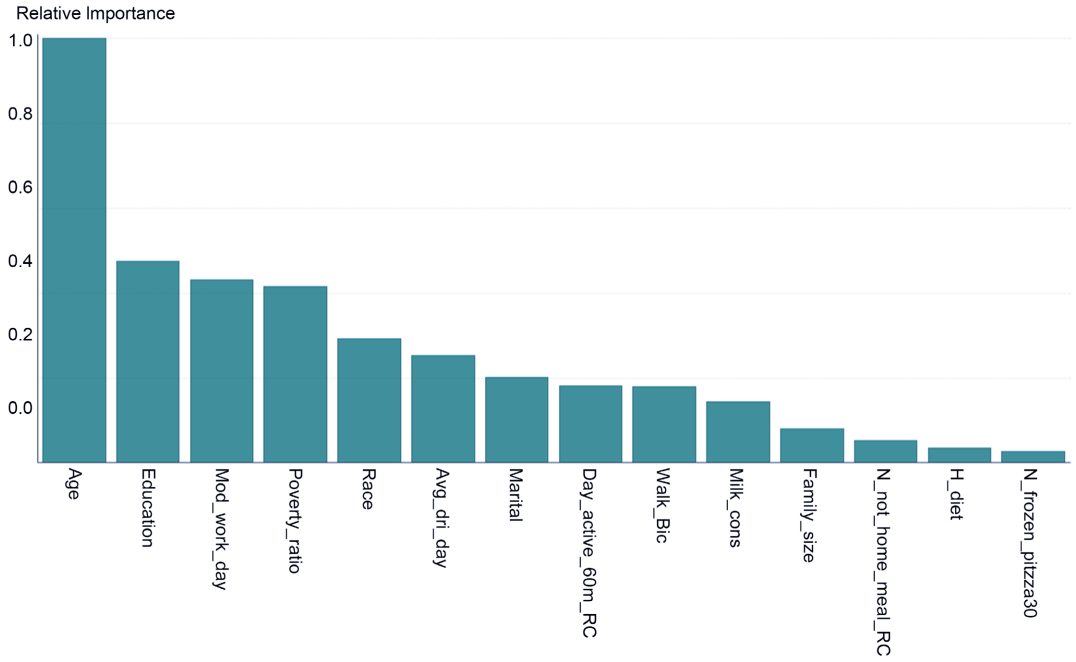

Figure 2. Ranking of important variables in the dataset.

2.5.3. Variable Selection

For each of the machine learning algorithms mentioned in this study, the variable selection procedure is the same as the initial process of variable selection at the beginning of the analysis. Relative importance procedure is used to identify and rank the variables for each algorithm and it turns out that for each algorithm, the ranking of the variable has changed from process to process. The following Table 1 shows the ranking of the most important variables for each of the algorithms. In this table, only those variables are considered as input variables accepted by all the machine learning algorithms.

Table 1. Variable selection by all algorithms.

Name	Variable Level	Role	Reason
RISK_DIAB	BINARY	TARGET	
AGE	INTERVAL	INPUT	
AVG_DRI_DAY	INTERVAL	INPUT	
DAY_ACTIVE_60M_RC	NOMINAL	INPUT	
EDUCATION	NOMINAL	INPUT	
FAMILY_SIZE	NOMINAL	INPUT	
GENDER	BINARY	INPUT	
GREATEST_WEIGHT	INTERVAL	INPUT	
HEIGHT	INTERVAL	INPUT	
H_DIET	NOMINAL	INPUT	
MARITAL	NOMINAL	INPUT	
MILK_CONS	NOMINAL	INPUT	
MOD_WORK_DAY	NOMINAL	INPUT	

Continued

N_FROZEN_PITZZA30	INTERVAL	INPUT	
RACE	NOMINAL	INPUT	
SMOKING	NOMINAL	INPUT	
WEIGHT	INTERVAL	INPUT	
N_NOT_HOME_MEAL_RC	NOMINAL	REJECTED	Combination Criterion
N_READY_EAT30_RC	NOMINAL	REJECTED	Combination Criterion
WALK_BIC	NOMINAL	REJECTED	Combination Criterion

It turns out that, no. of not a home meal, no. of ready to eat meal in the last 30 days, and walking-biking variables are rejected by all the algorithms. In our analysis, we have considered those variables accepted by all the algorithms because of the fact that each and every algorithm has its own selection criterion to be considered in the final analysis [18].

After selecting appropriate covariates for the model building, we have tried most of the commonly known machine learning algorithms and at the end of the analysis, we have compared all the models to determine the champion model based on ASE (Average Squared Error) [19], CRP (Captured Response Percentage), and ROC [20]. **Figure 3** shows the complete flow chart of the analysis.

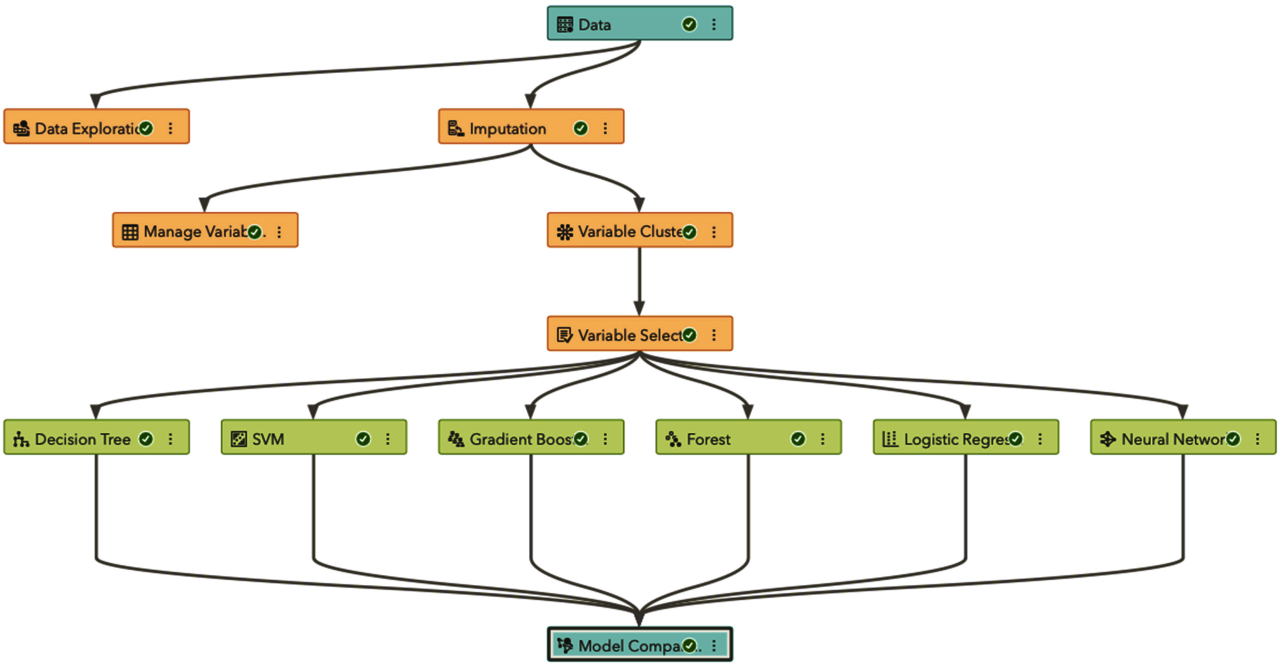

Figure 3. Flow chart of the Analysis.

In the flow chart figure above, the whole process of data analysis and machine learning model building is postulated with respect to our analysis. In this workflow, we have considered 6 most commonly known machine learning algorithms among those 5 are under supervised machine learning algorithms (Decision Tree, SVM, Gradient Boost, Forrest, Logistic Regression) and 1 is under supervised and unsupervised machine algorithm (Neural Network).

3. Proposed Champion Model

In the process of building a machine learning model, we have implemented six different types of algorithms and compared their results with each other to determine the best model. After considering the numerical values of ROC, ASE, Captured response percentage, and KS (Youden) [21] and it turns out the **Random Forest** model is the champion machine learning model for classifying the prediabetes patients. In the following Table 2, the comparative results are shown.

Table 2. Model comparison for prediabetes data.

Algorithm Name	ASE	KS (Youden)	ROC Area	CRP	Champion
Forest	0.115	0.1298	0.593	5.226	✠
Neural Network	0.249	0.0000	0.500	5.074	
SVM	0.192	0.0320	0.501	5.175	
Logistic Regression	0.117	0.0720	0.539	5.124	
Decision Tree	0.118	0.0730	0.552	5.256	
Gradient Boosting	0.117	0.0903	0.545	4.819	

From the above table, we see that the greatest KS (Youden) among all the models is for Forest about 0.1298 and areas under the ROC curve is 0.593 and this model has the smallest ASE (Average Squared Error) among all the model algorithm as per our analysis. So, to determine the champion model we have selected the "Forest" is the best model among machine learning algorithms.

According to the definition of the Random Forest for regression or classification has the implementation of the ideas known as bagging [22]. Because of the bagging this algorithm reduces the variance. This algorithm is setup as follows:

1) For $b = 1$ to B:

a) Draw a bootstrap sample $\mathbf{Z}^*$ of size N from the training data.

b) Grow a random-forest tree T_b to the bootstrapped data, by recursively repeating the following steps for each terminal node of the tree, until the minimum node size $n_{\min}$ is reached.

i) Select m variables at random from the p variables.

ii) Pick the best variable split-point among the m.

iii) Split the node into two daughter nodes.

2) Output the ensemble of trees $\{T_b\}_1^B$.

To make a prediction at a new point x:

• Regression: $\hat{f}_{rf}^B(x) = \frac{1}{B}\sum_{b=1}^B T_b(x)$

• Classification: Let $\hat{C}_b(x)$ be the class prediction of the b^{th} random-forest tree. Then $\hat{C}_{rf}^B(x) = \text{majority vote}\left\{\hat{C}_b(x)\right\}_1^B$

The random forests cannot over-fit the data because of the fact that B does not cause the random forest sequence to over-fit like bagging, the random forest estimate does the approximation on the expectation

$$\hat{f}_{rf}(x) = E_{\Theta} T(x;\Theta) = \lim_{B\to\infty} \hat{f}(x)_{rf}^{B} \quad (1)$$

For this specific study, the random forest (Forest) is selected because this algorithm handles non-linear relationships and interactions better compared to logistic regression which is simpler and assumes linear relationships between and among predictors and response variable. The forest algorithm is less sensitive to parameter tuning and capable of handling large datasets compared to support vector machine (SVM) which needs careful tuning and it could be computationally expensive for larger dataset. Since the random forest is easier to tune and less prone to over-fitting compared to gradient boosting machines (GBM), decision tree, neural network [23], etc.

4. Results & Discussion

In accordance with the study purpose, we have extended the research area that were initially presented in the work done by Bashar & Tsokos *et al.* [24]-[27]. Since the Forest algorithm is the champion one as a machine learning model, so we will discuss the results of this model in detail. Forest model is the ensemble of Decision Tree and the options we have used to build this model is as follows when we were splitting the tree:

- The number of trees used was 50.
- During the Tree splitting options, the class target criterion used is the Entropy.
- Maximum depth of the Tree was 12.
- Minimum leaf size used was 15.
- The number of bins for the continuous variable was 100.

After setting all the values in the algorithm, we have acquired the Forest model and as per our champion model, we have ranked the attributable variables according to the relative importance in **Figure 4** below.

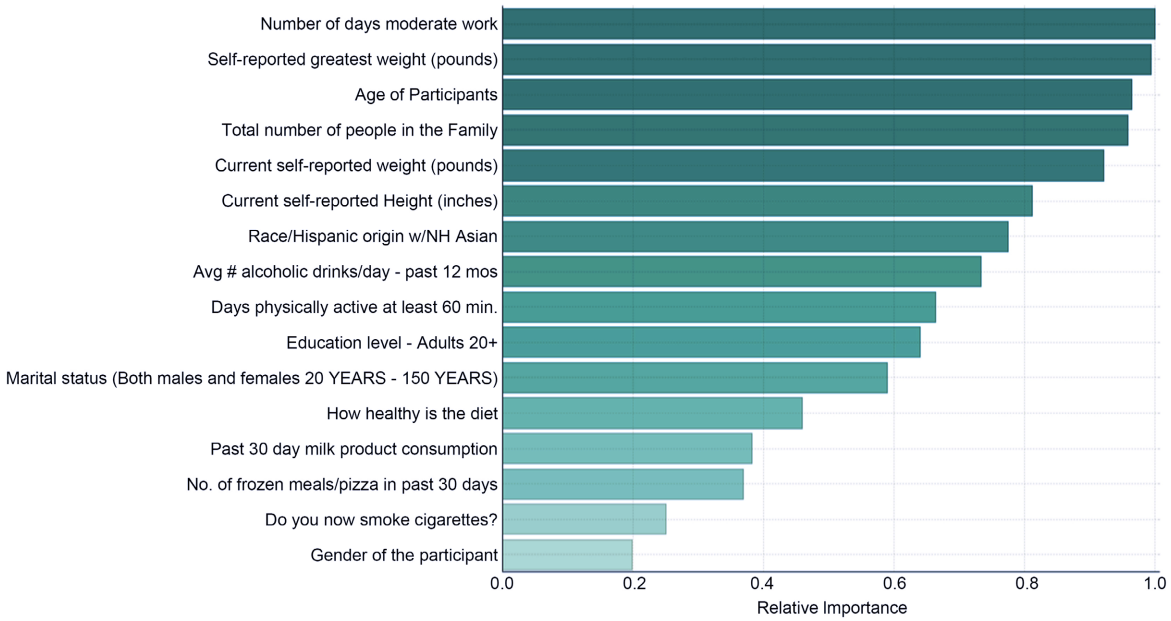

Figure 4. Ranking of important variables in the champion model (Forest).

The five most important factors are Number of days moderate work, *Self-reported greatest weight* (*pounds*), *Age of Participants*, *Total number of people in the Family*, *and Current self-reported weight* (*pounds*). The next chart below shows the Average Squared Error for the proposed model below. From Figure 5, it is very important to see that the ASE decreases as the model trees grow larger but after 20 trees it is convergent for all the training, validation and test data partitions remain flat through 50 trees that were the option used for the number of trees in the algorithm.

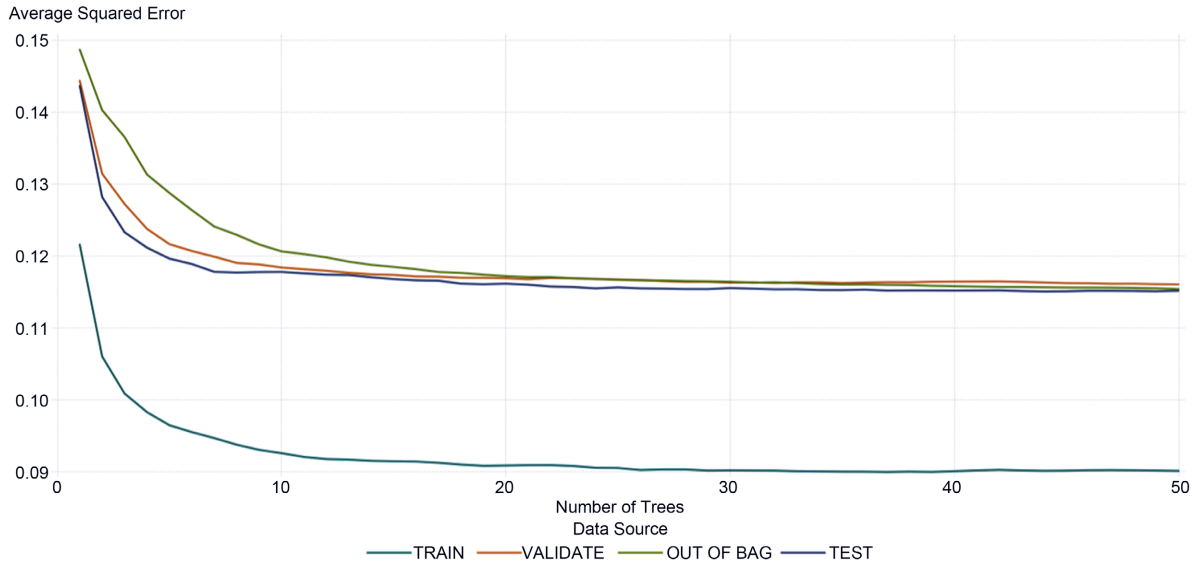

Figure 5. Average squared error for proposed model (Forest).

In the case of assessing the model through the ROC curve, it is a plot of sensitivity (the true positive rate) against 1-specificity (the false positive rate), which are both measures of classification based on the confusion matrix. These measures are calculated at various cutoff values. To help identify the best cutoff to use when scoring data, the KS Cutoff reference line is drawn at the value of 1-specificity where the greatest difference between sensitivity and 1-specificity is observed for the VALIDATE partition. The KS Cutoff line is drawn at the cutoff value 0.85, where the 1-specificity value is 0.574 and the sensitivity value is 0.687. Cutoff values range from 0 to 1, inclusive, in increments of 0.05. At each cutoff value, the predicted target classification is determined by whether the Risk of Prediabetes, which is the predicted probability of the event "2" (category - NO) for the target Risk/diab, is greater than or equal to the cutoff value. When P_Risk_diab2 (category - NO) is greater than or equal to the cutoff value, then the predicted classification is the event, otherwise, it is a non-event.

The confusion matrix for each cutoff value contains four cells that display the true positives for events that are correctly classified (TP), false positives for non-events that are classified as events (FP), false negatives for events that are classified as non-events (FN), and true negatives for non-events that are classified as non-events (TN). True negatives include non-event classifications that specify a

different non-event. Sensitivity is calculated as TP/(TP + FN). Specificity, the true negative rate, is calculated as TN/(TN + FP), so 1-specificity is FP/(TN + FP). The values of sensitivity and 1-specificity are plotted at each cutoff value. A ROC curve in Figure 6 that rapidly approaches the upper-left corner of the graph, where the difference between sensitivity and 1-specificity is the greatest, indicates a more accurate model. A diagonal line where sensitivity = 1-specificity indicates a random model. Captured response percentage is calculated by sorting each partition in descending order by the predicted probability of the target event P_Risk_diab2, which represents the predicted probability of the event "2" (category - NO) for the target Risk_diab. The data is divided into 20 quantiles (demi-deciles, with 5% of the data in each), and the number of events in each quantile is computed. Captured response percentage is the percentage of the total number of events that are in that quantile. With no model, it is expected that 5% of the events are in each quantile.

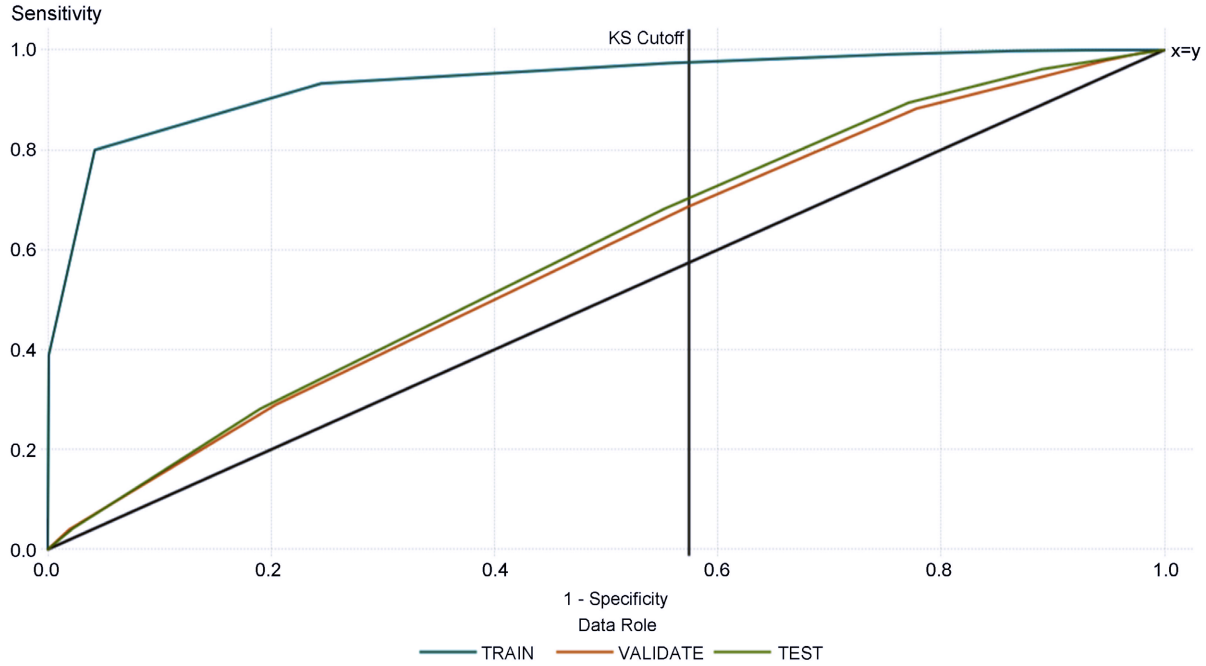

Figure 6. ROC for proposed model (Forest).

At the 5% quantile (depth of 5), the VALIDATE partition has a Captured response percentage of 5.4 (compared to the expected value of 5 for no model). The best possible value of Captured response percentage for this partition at depth 5 is 5.8. As shown in Figure 7 at the 5% quantile (depth of 5), the TRAIN partition has a Captured response percentage of 5.8 (compared to the expected value of 5 for no model). The best possible value of Captured response percentage for this partition at depth 5 is 5.79. At the 5% quantile (depth of 5), the TEST partition has a Captured response percentage of 5.4 (compared to the expected value of 5 for no model). The best possible value of Captured response percentage for this partition at depth 5 is 5.83.

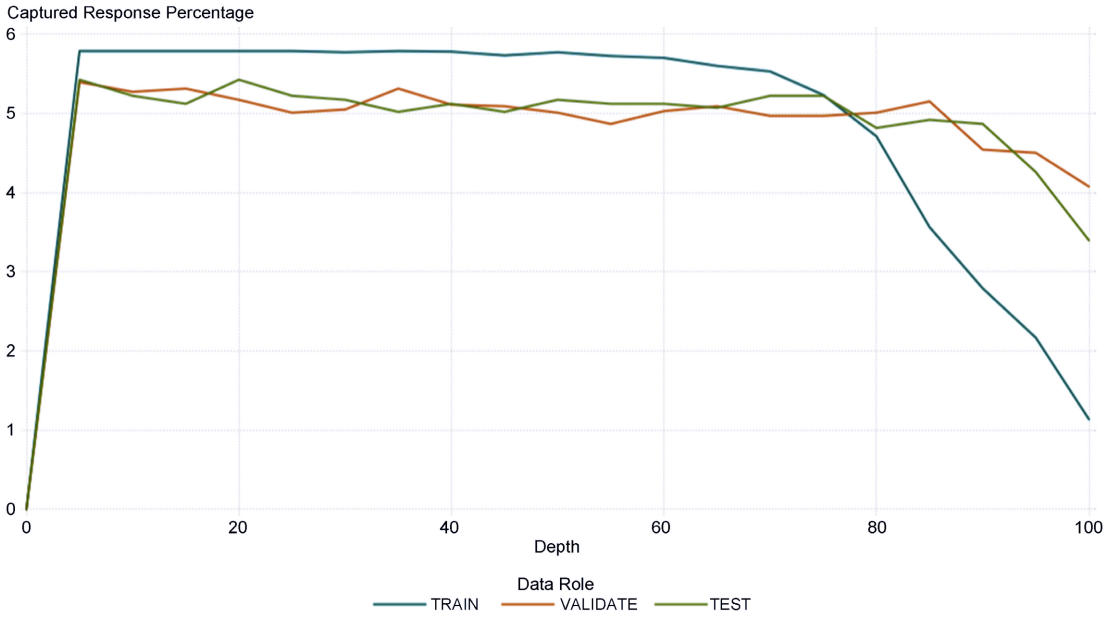

Figure 7. Captured response percentage for proposed model (Forest).

As part of model validation, we have used another plot called *Cumulative Lift* [28]. As shown in **Figure 8**, the cumulative lift for a particular quantile is the ratio of the number of events across all quantiles up to and including the current quantile to the number of events that would be there at random, or equivalently, the ratio of the cumulative response percentage to the baseline response percentage. The cumulative lift at depth 10 includes the top 10 percent of the data, which is the first 2 quantiles, which would have 10% of the events at random. Thus, cumulative lift measures how much more likely it is to observe an event in the quantiles than by selecting observations at random.

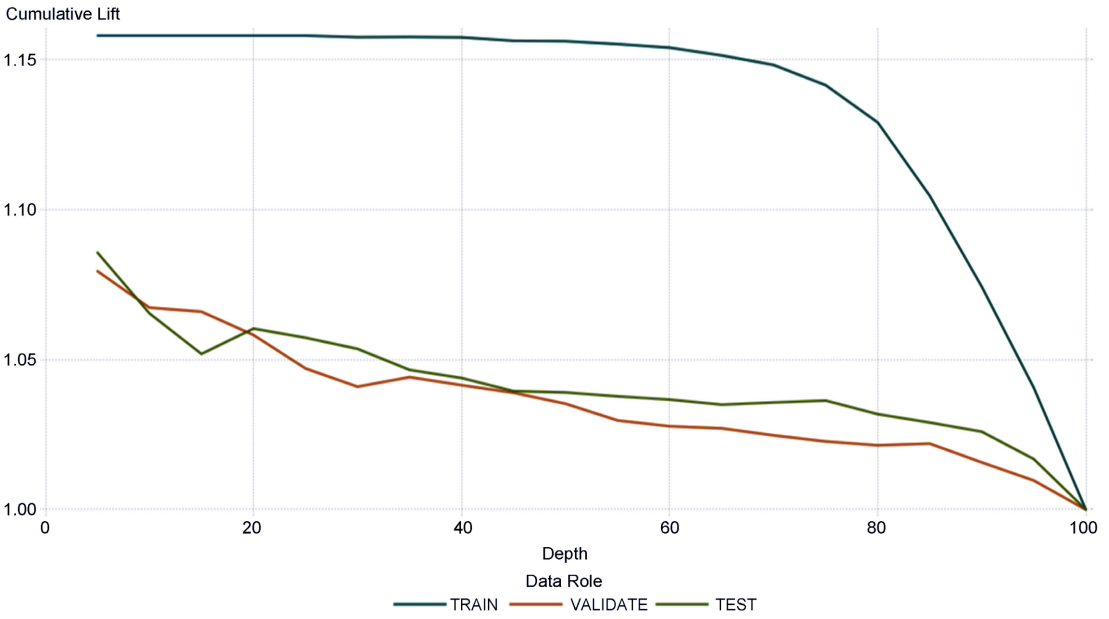

Figure 8. Cumulative lift for proposed model (Forest).

The VALIDATE partition has a Cumulative Lift of 1.07 in the 10% quantile (depth of 10) meaning there are about 1 times more events in the first two quantiles than expected by random (10% of the total number of events). Because this value is greater than 1, it is better to use your model to identify responders than no model, based on the selected partition. The TRAIN partition has a Cumulative Lift of 1.16 in the 10% quantile (depth of 10) meaning there are about 1 times more events in the first two quantiles than expected by random (10% of the total number of events). Because this value is greater than 1, it is better to use your model to identify responders than no model, based on the selected partition. The TEST partition has a Cumulative Lift of 1.07 in the 10% quantile (depth of 10) meaning there are about 1 times more events in the first two quantiles than expected by random (10% of the total number of events). Because this value is greater than 1, it is better to use your model to identify responders than no model, based on the selected partition. For our study, as shown in Figure 8, cumulative lift is calculated by sorting each partition in descending order by the predicted probability of the target event P_Risk_diab2, which represents the predicted probability of the event "2" (category - NO) for the target Risk_diab. The data is divided into 20 quantiles (demi-deciles, with 5% of the data in each), and the number of events in each quantile is computed.

In terms of model accuracy, we have plotted the accuracy plot for our proposed model as it is shown in Figure 9 below. Accuracy is the proportion of observations that are correctly classified as either an event or non-event, calculated at various cutoff values. As shown in Figure 9 below shows the accuracy of the proposed model. Cutoff values range from 0 to 1, inclusive, in increments of 0.05.

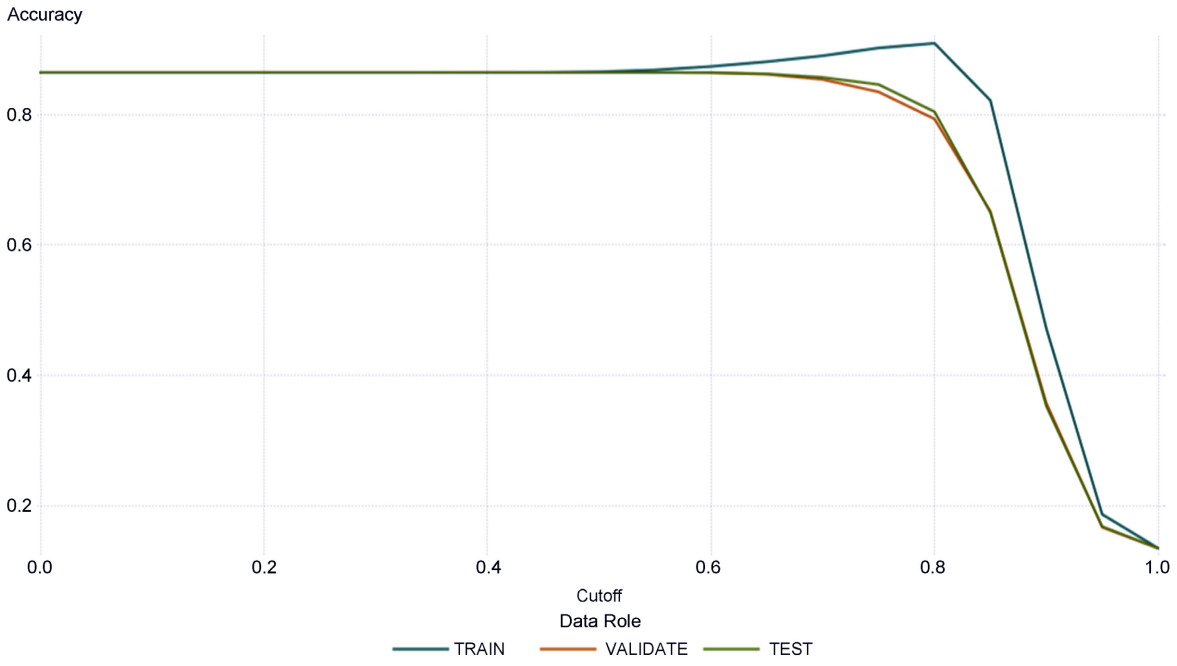

Figure 9. Accuracy plot for proposed model (Forest).

At each cutoff value, the predicted target classification is determined by whether P_Risk_diab2, which is the predicted probability of the event "2" (category - No) for the target Risk_diab, is greater than or equal to the cutoff value. When P_Risk_diab2 is greater than or equal to the cutoff value, then the predicted classification is the event, otherwise it is a non-event. When the predicted classification and the actual classification are both events (true positives) or both non-events (true negatives), the observation is correctly classified. If the predicted classification and actual classification disagree, then the observation is incorrectly classified. Accuracy is calculated as (true positives + true negatives)/(total observations).

5. Conclusions

From the results of the present study, it indicates that the Forest model that we have developed performed better than any other existing model used for classifying and predict the prevalence of Prediabetes condition among the US population. Although, ANN could be used as the best model for this study purpose which also suggested by some machine learning model for screening individuals for prediabetes developed by Choi *et al.* [29] where six risk factors were used to build the model from the Korean population on prediabetes, but considering 16 risk factors included in our study produced Forest is the best machine learning algorithm. On the other hand, Meng *et al.* [30] did comparative study among the performances of logistic regression, ANNs, and decision tree models for predicting diabetes as well as prediabetes in Chine population using common risk factors. Regarding Meng *et al.* study, the ANNs model was the least suggested model with the poorest performance in terms of accuracy. These indicate that our model is consistent with their machine learning model. Also, if any clinical study or research wants to use the model to classify the individual prediabetes state with more risk factors involved, then our proposed model will perform the best at a higher accuracy.

Although there are some common risk factors (covariates/attributes) that were included in our model and other machine learning models developed for prediction of prediabetes condition [30] [31], they have considered few risk factors. But in case of a large number of risk factors included in the model, the Forest model will perform relatively better.

In this present study, we have constructed a reasonably better model to classify prediabetes in the USA population. In the case of interested countries, they can implement a similar type of methodologies by government agencies, scholars and researchers might develop region or state-specific machine learning models. The development of such a model can be deployed as a web application with a user-friendly calculator program. This will enable the access of mass people including the health workers, medical scholars and professionals. Eventually, early diagnosis or preventive measures with correctly identified prediabetes state will impact the public health issue on this subject matter. Intern, it will help to reduce the incidence of other health issues related to prediabetes condition such

as heart disease, stroke, and obesity among early diagnosed and undiagnosed portion of the population.

The random forest model can be deployed in the medical and health centers to rank the most risk factors predicting the prediabetes conditions among individuals. Regarding the limitation of the algorithm, the ROC area could be improved if more relevant covariates were introduced at the stage of data acquisition and sanitation. This model can be generalized among the USA population only since the data are collected through the nationally established survey done in every year. To enhance the explanatory power and to validate the finding of this study in the future, it is advisable to the scholars to incorporate a robust variable selection method such as LASSO or RIDGE regression. Then implement machine learning algorithm to classify the pre-diabetic status of the individuals more accurately.

Acknowledgements

Sincere thanks to the research team members of University of South Florida (USF), department of Mathematics & Statistics graduate Dr. Mahdi Goudarzi to support the data acquisition, and mentor/advisor, Prof. Chris P. Tsokos for their professional guide and support, and special thanks to managing editor *Delia Zhu* for providing logistic materials and a rare attitude of high quality.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Buysschaert, M. and Bergman, M. (2011) Definition of Prediabetes. *Medical Clinics of North America*, **95**, 289-297. https://doi.org/10.1016/j.mcna.2010.11.002

[2] CDC: Centers for Disease Control and Prevention (2019) What Causes Predia betes? https://www.cdc.gov/diabetes/prevention-type-2/prediabetes-prevent-type-2.html?CDC_AAref_Val=https://www.cdc.gov/diabetes/basics/prediabetes.html

[3] Tabák, A.G., Herder, C., Rathmann, W., Brunner, E.J. and Kivimäki, M. (2012) Prediabetes: A High-Risk State for Diabetes Development. *The Lancet*, **379**, 2279-2290. https://doi.org/10.1016/s0140-6736(12)60283-9

[4] Coutinho, M., Gerstein, H.C., Wang, Y. and Yusuf, S. (1999) The Relationship between Glucose and Incident Cardiovascular Events. A Metaregression Analysis of Published Data from 20 Studies of 95,783 Individuals Followed for 12.4 Years. *Diabetes Care*, **22**, 233-240. https://doi.org/10.2337/diacare.22.2.233

[5] Port, S.C., Goodarzi, M.O., Boyle, N.G. and Jennrich, R.I. (2005) Blood Glucose: A Strong Risk Factor for Mortality in Nondiabetic Patients with Cardiovascular Disease. *American Heart Journal*, **150**, 209-214. https://doi.org/10.1016/j.ahj.2004.09.031

[6] Tian, L., Zhu, J., Liu, L., Liang, Y., Li, J. and Yang, Y. (2013) Prediabetes and Short-Term Outcomes in Nondiabetic Patients after Acute St-Elevation Myocardial

Infarction. *Cardiology*, **127**, 55-61. https://doi.org/10.1159/000354998

[7] Herman, W.H., Hoerger, T.J., Brandle, M., Hicks, K., Sorensen, S., Zhang, P., *et al.* (2005) The Cost-Effectiveness of Lifestyle Modification or Metformin in Preventing Type 2 Diabetes in Adults with Impaired Glucose Tolerance. *Annals of Internal Medicine*, **142**, 323-332. https://doi.org/10.7326/0003-4819-142-5-200503010-00007

[8] Yoo, T.K., Kim, S.K., Kim, D.W., Choi, J.Y., Lee, W.H., Oh, E., *et al.* (2013) Osteoporosis Risk Prediction for Bone Mineral Density Assessment of Postmenopausal Women Using Machine Learning. *Yonsei Medical Journal*, **54**, 1321-1330. https://doi.org/10.3349/ymj.2013.54.6.1321

[9] National Health and Nutrition Examination Survey: NHANES 2015-2016 Questionnaire Data. https://wwwn.cdc.gov/nchs/nhanes/ContinuousNhanes/Default.aspx?BeginYear=2015

[10] Mitchell, T.M. (1997) Learning, Machine. McGraw-Hill.

[11] Schölkopf, B. and Smola, A.J. (2001). Learning with Kernels: Support Vector Machines, Regularization, Optimization, and beyond. The MIT Press. https://doi.org/10.7551/mitpress/4175.001.0001

[12] Friedman, J.H. (2001) Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics*, **29**, 1189-1232. https://doi.org/10.1214/aos/1013203451

[13] Segal, M.R. (2004) Machine Learning Benchmarks and Random Forest Regression. Center for Bioinformatics and Molecular Biostatistics, University of California, San Francisco.

[14] Hansen, L.K. and Salamon, P. (1990) Neural Network Ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **12**, 993-1001. https://doi.org/10.1109/34.58871

[15] (2019) SAS Viya Data Mining and Machine Learning: Procedures Guide. The TREESPLIT Procedure—Variable Importance. https://documentation.sas.com/?docsetId=casmlocsetTar-get=viyaml_treesplit_details20.htmdocsetVersion=3.0locale=en

[16] Fernandez, G. (2010) Statistical Data Mining Using SAS Applications. CRC Press.

[17] Rubin, D.B. (2004) Multiple Imputation for Nonresponse in Surveys. John Wiley & Sons.

[18] Guyon, I. and Elisseeff, A. (2003) An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research*, **3**, 1157-1182.

[19] Wallach, D. and Goffinet, B. (1989) Mean Squared Error of Prediction as a Criterion for Evaluating and Comparing System Models. *Ecological Modelling*, **44**, 299-306. https://doi.org/10.1016/0304-3800(89)90035-5

[20] Hanley, J.A. and McNeil, B.J. (1982) The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve. *Radiology*, **143**, 29-36. https://doi.org/10.1148/radiology.143.1.7063747

[21] Fluss, R., Faraggi, D. and Reiser, B. (2005) Estimation of the Youden Index and Its Associated Cutoff Point. *Biometrical Journal*, **47**, 458-472. https://doi.org/10.1002/bimj.200410135

[22] Hastie, T., Tibshirani, R., Friedman, J.H. and Friedman, J.H. (2009) The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.

[23] Gareth, J., Daniela, W., Trevor, H. and Robert, T. (2013) An Introduction to Statistical Learning: With Applications in R. Spinger.

[24] Raquibul Bashar, A.K.M. (2019) Probabilistic Modeling of Democracy, Corruption, Hemophilia A and Prediabetes Data. Ph.D. Thesis, University of South Florida.

[25] Bashar, A.K.M.R. and Tsokos, C.P. (2019) Statistical Parametric Analysis on Democracy Data. *Open Access Library Journal*, **06**, 1-18. https://doi.org/10.4236/oalib.1105828

[26] Raquibul Bashar, A.KM. and Tsokos, C.P. (2017) Parametric Analysis of Factor 8 (F8) Hemophilia A. *International Journal of Mathematical Sciences in Medicine* (*IJMSM*), **1**, 1-10.

[27] Raquibul Bashar, A.K.M. and Tsokos, C.P. (2019) Statistical Classification of Democracy Index Scores of Countries of the World. *Scholars Journal of Arts, Humanities and Social Sciences*, **7**, 773-784.

[28] Festus Ayetiran, E. and Barnabas Adeyemo, A. (2012) A Data Mining-Based Response Model for Target Selection in Direct Marketing. *International Journal of Information Technology and Computer Science*, **4**, 9-18. https://doi.org/10.5815/ijitcs.2012.01.02

[29] Choi, S.B., Kim, W.J., Yoo, T.K., Park, J.S., Chung, J.W., Lee, Y.H., Kang, E.S. and Kim, D.W. (2014) Screening for Prediabetes Using Machine Learning Models. *Computational and Mathematical Methods in Medicine*, **2014**, Article ID: 618976.

[30] Meng, X., Huang, Y., Rao, D., Zhang, Q. and Liu, Q. (2012) Comparison of Three Data Mining Models for Predicting Diabetes or Prediabetes by Risk Factors. *The Kaohsiung Journal of Medical Sciences*, **29**, 93-99. https://doi.org/10.1016/j.kjms.2012.08.016

[31] Lee, Y., Bang, H., Kim, H.C., Kim, H.M., Park, S.W. and Kim, D.J. (2012) A Simple Screening Score for Diabetes for the Korean Population: Development, Validation, and Comparison with Other Scores. *Diabetes Care*, **35**, 1723-1730. https://doi.org/10.2337/dc11-2347