The New England Journal of Statistics in Data Science Volume 1, 46–61 (2023)
DOI: https://doi.org/10.51387/22-NEJSDS5

Effects of stopping criterion on the growth of trees in regression random forests

Aryana ARSHAM*, Philip ROSENBERG, and Mark LITTLE

Abstract

Random forests are a powerful machine learning tool that capture complex relationships between independent variables and an outcome of interest. Trees built in a random forest are dependent on several hyperparameters, one of the more critical being the node size. The original algorithm of Breiman, controls for node size by limiting the size of the parent node, so that a node cannot be split if it has less than a specified number of observations. We propose that this hyperparameter should instead be defined as the minimum number of observations in each terminal node. The two existing random forest approaches are compared in the regression context based on estimated generalization error, bias-squared, and variance of resulting predictions in a number of simulated datasets. Additionally the two approaches are applied to type 2 diabetes data obtained from the National Health and Nutrition Examination Survey. We have developed a straightforward method for incorporating weights into the random forest analysis of survey data. Our results demonstrate that generalization error under the proposed approach is competitive to that attained from the original random forest approach when data have large random error variability. The R code created from this work is available and includes an illustration.

KEYWORDS AND PHRASES: Regression random forest, Node size, Generalization error.

1. INTRODUCTION

The prominence of the random forest (RF) algorithm as a machine learning tool is due to its ability to accurately model large and complex datasets and availability in many software packages. The non-parametric model is determined by three user specified parameters, one of the more critical being the stopping criterion node size. The node size regulates the model complexity of each tree in the forest and has implications on the statistical performance of the algorithm.

The original RF model introduced by [4] defines the node size parameter as the minimum number of observations required to further split the data, e.g., minimum size of parent nodes. We advocate that the node size should instead be defined as the minimum number of observations in the final subgroups, e.g., minimum size of the leaf nodes. The proposed approach guarantees that predictions from a tree within the forest are the mean outcome of at least node size observations. In contrast, the original RF allows for subgroups to contain as few as a single observation.

Most recent developments on the RF have been extensions of the original algorithm for specified analysis. Despite this, the original algorithm remains widely used for many health care and other applications. The popularity of the algorithm warrants further investigation into the effect of its hyperparameters on predictive performance, in particular node size. There has been some attention in the literature on the tuning of certain RF hyperparameters, in particular the node size and the maximum number of nodes as discussed in [14].

The multiplicity of new RF methods has been reviewed by [14]. Specifically, the authors highlight the need for further investigation on the effects of tuning hyperparameters for the RF, which is reviewed in greater depth by [16]. The importance among hyperparameters for various machine learning methods including the RF algorithm are explored in [17]. Their results show that node size and m_{try} are the two most influential hyperparameters for the RF algorithm, in particular the effect on the variability between trees in the RF model. The authors note an important distinction between the two differing definitions of node size:

> At first sight, the minimal samples per split and minimal samples per leaf hyperparameters seem quite similar, but at closer inspection they are not: logically, minimal samples per split is overshadowed by minimal samples per leaf. (p. 6)

In this work we investigate the effect of the two definitions of node size on the predictive performance of the RF model. Specifically, defining the node size as minimal samples per split versus minimal samples per leaf, and propose the later approach. Section 2 provides background on the regression framework. In Section 3 the bias-variance decomposition and corresponding estimators used to evaluate the performance of the competing approaches are detailed. The type 2 diabetes example obtained from The National Health and Nutrition Survey (NHANES) data is then introduced in Section 4 along with analysis and results of the data. The

*Corresponding author.

section also includes description of the survey design which [5] discuss in greater depth. Due to the survey design, this work modifies the two competing RF algorithms for the inclusion of individual participant weights. Section 5 details the simulation study consisting of 14 simulated datasets and results obtained from the simulation analysis. Section 6 summarizes the results from our analysis and the advantages and limitations of each approach.

2. BACKGROUND

An overview of the regression framework utilized in this paper is now discussed. Let the available dataset be denoted by $T = \{T_n, T_\kappa\}$, such that $T = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_{n+\kappa}, y_{n+\kappa})\}$. The portion of the sample T used for model testing and tuning of the random forest (RF) model is denoted as $T_\kappa = \{(\mathbf{x}_{n+1}, y_{n+1}), \ldots, (\mathbf{x}_{n+\kappa}, y_{n+\kappa})\}$ and is of size κ. The remaining sample T_n is the proportion of the data used to train the RF model. The training sample is denoted as $T_n = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ and is of size n. The input covariates are $\{\mathbf{x}_1, \ldots, \mathbf{x}_{n+\kappa}\}$, and each is a vector of size p assumed to lie in $\Re^p$. The continuous dependent variable $y = \{y_1, \ldots, y_{n+\kappa}\}$ are assumed to be generated by a model $y_i = f(\mathbf{x}_i) + \epsilon_i$, $i = 1, \ldots, n+\kappa$. The random errors ϵ_i, $i = 1, \ldots, n+\kappa$ are assumed to be independent and identically distributed random variables with mean zero, $E(\epsilon_i) = 0$, and constant variance, $Var(\epsilon_i) = \sigma^2$, $i = 1, \ldots, n+\kappa$. [2, pg. 2502] showed that when the function $f = E[Y|X = x]$ is almost everywhere continuous in x the random forest estimator is consistent. The aim of the RF algorithm is to estimate the unknown mean function f for a new observation having covariate vector $\mathbf{x}_0$, specifically $f(\mathbf{x}_0) = E[Y|\mathbf{X} = \mathbf{x}_0]$. For a detailed overview of the RF modelling technique see [3].

The random forest algorithm is governed by three user selected hyperparameters: M which is the number of trees in the forest, m_{try} which is the number of randomly selected features used to split each node, and node size which is the minimum size of nodes required for further splitting. The model complexity is primarily determined by m_{try} and node size. The recommended values for these hyperparameters in the regression framework are M = 500, $m_{try} = \frac{p}{3}$, node size = 5. In particular these are the default values implemented in the RF statistical package, `randomForest` in `R` software.

The documentation of `randomForest` states that the node size hyperparameter is the minimum size of the terminal nodes, e.g., leaf nodes (see [8]). We created two algorithms in `R` software, one implementing the original RF (that defines node size as the size of the parent node) and the other defining node size as the size of the leaf node. Results from the application of the two algorithms on a collection of simulated data sets suggests that the stopping criteria in `randomForest` package is defined by the size of the parent node. This misspecification prompted our investigation into the comparison of these two RF approaches in the regression context, based on their respective predictive performances for health care and simulated data.

As we shall demonstrate, this difference in node size definition has important implications on predictive performance. The node size hyperparameter controls the complexity of each tree in the forest by its depth. Smaller values of node size result in deeper trees allowing for predictions that usually attain smaller bias-squared. Larger values of node size result in shallower trees with larger bias-squared but smaller variability of resulting predictions. The original RF algorithm allows for sparse nodes that may contain as few as a single observation. The proposed definition of node size is the minimum number of observations defining the final nodes of each tree. The proposed stopping criterion is implemented in certain packages, including `randomForestSRC` by [7] in `R` software for the analysis of survival data.

For implementing the RF algorithm, T_n is the data that is randomly sampled with replacement in the construction of each tree in the forest. We apply a hold-out approach in which observations in the testing set, T_κ are used to obtain the generalization error performance. For the remainder of this paper, we will refer to the original RF approach (as implemented in the `randomForest` package) as the 'parent approach' and the proposed method as the 'leaf approach'. Development of the generalization error decomposition and predictive metrics used to assess the performance of the two competing algorithms are now be presented.

3. PREDICTIVE PERFORMANCE CRITERIA AND ESTIMATION

The generalization error based on the mean squared error loss function is the mean squared prediction error (MSPE). The comparison of the two random forest (RF) approaches is based on three criteria: MSPE, bias-squared, and variance of predictions for new observations. Assessment of the modelling approaches from these criteria is facilitated by a simulation study. A total of B simulated RF models are built and used to estimate MSPE, bias-squared, and variance. In particular, estimation of bias-squared requires a known mean function, necessitating a simulation study. Development of the estimators for each of the criterion are derived from the standard bias-variance decomposition of MSPE (see for example [6]) further details of which are provided next.

3.1 Bias-variance decomposition of the generalization error

To introduce the bias-variance decomposition of the MSPE let $\hat{f}_{rf}(\mathbf{x}_0)$ denote the prediction for a new observation with covariate vector $\mathbf{x}_0$, (in $\Re^p$) from the RF model. It follows that the MSPE of the RF model for a new observation $\mathbf{x}_0$ is,

$$MSPE[\hat{f}_{rf}(\mathbf{x}_0)] = \sigma^2 + Bias^2[\hat{f}_{rf}(\mathbf{x}_0)] + Var[\hat{f}_{rf}(\mathbf{x}_0)]. \tag{3.1}$$

The last two terms in (3.1) are estimated from B simulated RF predictions, denoted by $\tilde{f}^b_{rf}(\mathbf{x}_0)$, $b = 1, \ldots, B$ and taken from independent sets of forests. The first term is the intrinsic variance or the variance of the random error $\epsilon_i, i = 1, \ldots, n + \kappa$. The estimated prediction from the mth tree in the b^{th} simulated RF is denoted $\tilde{f}^b_m(\mathbf{x}_0)$, $m = 1, \ldots, M$, $b = 1, \ldots, B$. The b^{th} RF prediction $\tilde{f}^b_{rf}(\mathbf{x}_0)$ is the average of the M regression tree predictions, such that $\tilde{f}^b_{rf}(\mathbf{x}_0) = \frac{1}{M}\sum_{m=1}^{M} \tilde{f}^b_m(\mathbf{x}_0)$.

Let the sample variance of any single randomly grown tree for the new observation be $\phi^2(\mathbf{x}_0)$. The trees in the RF are identically distributed, but not independent. The sample correlation between any two trees for the new observation is assumed to be $\rho(\mathbf{x}_0)$. To estimate the terms in (3.1) we begin with the sample variance of any tree belonging to forest which is,

$$\begin{aligned}\phi^2(\mathbf{x}_0) = Var_{T_n}\Big[E_{\Omega|T_n}\{\hat{f}(\mathbf{x}_0, \Omega(T_n))\}\Big] \\ + E_{T_n}\Big[Var_{\Omega|T_n}\{\hat{f}(\mathbf{x}_0, \Omega(T_n))\}\Big].\end{aligned} \tag{3.2}$$

This leads to the variance estimate,

$$\begin{aligned}\widehat{\phi^2}(\mathbf{x}_0) = \frac{1}{B-1}\sum_{b=1}^{B}\Big[\tilde{f}^b_{rf}(\mathbf{x}_0) - \frac{1}{B}\sum_{b=1}^{B}\tilde{f}^b_{rf}(\mathbf{x}_0)\Big]^2 \\ + \frac{1}{B}\sum_{b=1}^{B}\Big[\frac{1}{M-1}\sum_{m=1}^{M}(\tilde{f}^b_m(\mathbf{x}_0) - \tilde{f}^b_{rf}(\mathbf{x}_0))^2\Big].\end{aligned} \tag{3.3}$$

Since the regression trees are identically distributed it follows that the correlation between any pair is estimated as,

$$\begin{aligned}\widehat{\rho}(\mathbf{x}_0) &= \frac{\widehat{Cov}_{T_n}\Big[E_{\Omega|T_n}\{\hat{f}(\mathbf{x}_0, \Omega(T_n))\}\Big]}{\widehat{\phi^2}(\mathbf{x}_0)} \\ &= \frac{\frac{1}{B-1}\sum_{b=1}^{B}\Big[\tilde{f}^b_{rf}(\mathbf{x}_0) - \frac{1}{B}\sum_{b=1}^{B}\tilde{f}^b_{rf}(\mathbf{x}_0)\Big]^2}{\widehat{\phi^2}(\mathbf{x}_0)}.\end{aligned} \tag{3.4}$$

The RF prediction is the average prediction from the M trees constituting the forest, that is the average of identically distributed random variables having correlation $\rho(\mathbf{x}_0)$. Therefore, the estimated variance of the RF prediction for a new observation is,

$$\widehat{Var}\Big[\hat{f}_{rf}(\mathbf{x}_0)\Big] = \widehat{\rho}(\mathbf{x}_0)\widehat{\phi^2}(\mathbf{x}_0) + \Big[1 - \widehat{\rho}(\mathbf{x}_0)\Big]\frac{\widehat{\phi^2}(\mathbf{x}_0)}{M}. \tag{3.5}$$

The bias of the RF prediction is estimated by,

$$\widehat{Bias}\Big[\hat{f}_{rf}(\mathbf{x}_0)\Big] = f(\mathbf{x}_0) - \frac{1}{B}\sum_{b=1}^{B}\tilde{f}^b_{rf}(\mathbf{x}_0). \tag{3.6}$$

The bias-squared and variance are estimated from the testing sample T_κ as,

$$\overline{\widehat{Bias^2}} = \frac{1}{\kappa}\sum_{i=n+1}^{n+\kappa}\widehat{Bias^2}[\hat{f}_{rf}(\mathbf{x}_i)], \tag{3.7}$$

$$\overline{\widehat{Var}} = \frac{1}{\kappa}\sum_{i=n+1}^{n+\kappa}\widehat{Var}[\hat{f}_{rf}(\mathbf{x}_i)], \tag{3.8}$$

respectively. The estimated MSPE is the average over the B simulated RF predictions, calculated as:

$$\widehat{MSPE}_b[\tilde{f}^b_{rf}(\mathbf{x}_i)] = \frac{1}{\kappa}\sum_{i=n+1}^{n+\kappa}\Big[y_i - \tilde{f}^b_{rf}(\mathbf{x}_i)\Big]^2. \tag{3.9}$$

We then average the B RF predictions for a fixed model (with fixed hyperparameters), resulting in the estimated MSPE,

$$\overline{\widehat{MSPE}} = \frac{1}{B}\sum_{b=1}^{B}\Big(\widehat{MSPE}_b[\tilde{f}^b_{rf}(\mathbf{x}_i)]\Big). \tag{3.10}$$

The type 2 diabetes application is now discussed.

4. NHANES APPLICATION AND WEIGHTING APPROACH

The National Health and Nutrition Examination Survey (NHANES) are large studies designed to assess the health and nutrition of adults in the United States. Among medical conditions prescribed for the analysis of the NHANES data is type 2 diabetes. A recommended biomarker for the study of diabetes from the NHANES is glycohemoglobin, also known as Hemoglobin A1C as described in [11] and [10]. We aim to assess the predictive performance of the two random forest (RF) models on the NHANES data for predicting glycohemoglobin. Additionally, our analysis identifies risk factors for diabetes amongst the United States population 20 years and older who have not been formally diagnosed with the condition. Qualitative comparison of our results with those of similar datasets is addressed in the analysis.

Data from NHANES are used to determine prevalence and risk factors associated with certain diseases. The data consist of information taken from participant interviews and physical examinations. The most recent surveys are taken over two-year periods. The interview portion of the survey includes information on demographics, socio-economic status, dietary habits, and general health-related questions.

The design of the sampling is intended to be representative of the United States population. However, certain demographics, including ethnic minorities and older individuals are oversampled in order to increase the accuracy and precision of estimates associated with these sub-populations. The sample design of the data consists of multi-year, stratified, clustered four-stage samples. First stage of sampling

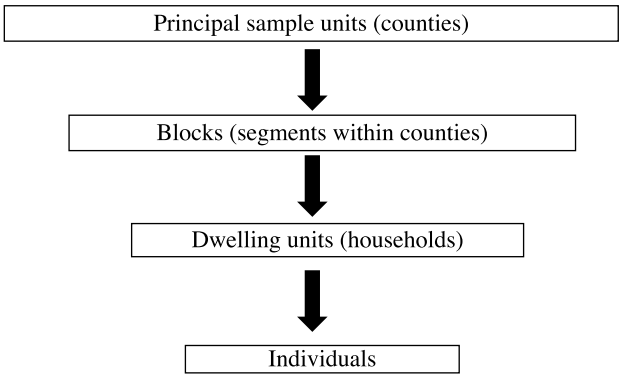


Figure 1: Four-stage sampling of NHANES 2015–2018 data.

is principal sample units (PSUs) consisting of counties. The second stage are blocks or segments within the counties. The third stage are dwelling units, which are households. The final sampling stage are individuals within the dwelling units. Figure 1 depicts the four-stage sampling of the NHANES sample design.

For our analysis we have utilized the cycles of 2015–2016 and 2017–2018, details of which are provided in [5]. The 2015–2016 data source is provided in [18] and 2017–2018 data source is provided in [19]. For these two cycles the oversampled sub-populations are: hispanics; non-hispanic blacks, non-hispanic non-black asians, non-hispanic whites and persons of other races and ethnicities at or below 185% of the federal poverty level; and non-hispanic whites and other races and ethnicities aged 0–11 years or 80 years and over. Weights are assigned to each participant in the survey and correspond to the number of persons in the United States population represented by each participant. The weights provided by NHANES account for the probabilities of selection, non-response in the survey, and differences between sample distribution and the population distribution. More information on the 2015–2018 survey design is detailed in [5].

Nineteen covariates are used in our analysis. Table 1 provides names of the variables and description provided by NHANES. In Table 1 the input variables are the covariates in the model. The outcome variable is glycohemoglobin measured as a percentage. Only samples undiagnosed with diabetes are utilized in the analysis, that is indicator of diabetes equal to one. The dataset used to train the RF algorithm is the 2015–2016 cycle having 4292 sampled units. The test dataset used to evaluate the predictive performance of the algorithms is the 2017–2018 cycle having 4051 sampled units.

Some of the input variables described in Table 1 have missing values. To conduct the analysis a straightforward imputation technique was applied. We created 14 subgroups based on gender and age. The subgroups are defined by combinations of binary gender (male and female) and the following age groups: [20–30), [30–40), [40–50), [50–60), [60–70), [70–80), [80+]. For each of the subgroups, missing values for any given variable were calculated as a function of non-missing training set values.

Table 1. NHANES variables and description.

Input Variable	**Description**
age	continuous (years from birth)
cholesterol	continuous (mg/dL)
diastolic	continuous (mm Hg)
gender	binary (0 = male, 1 = female)
glucose	continuous (mg/dL)
income	ordinal (1 to 15, annual household income grouped in intervals)
night urination	discrete (how many times urinate in night?)
race	nominal (1 = mexican american, 2 = other hispanic, 3 = non-hispanic white, 4 = non-hispanic black, 5 = other non-hispanic race)
risk for diabetes	binary (ever been told you have health risk for diabetes? 0 = yes, 1 = no)
salt	binary (doctor told you to reduce salt in diet? 0 = yes, 1 = no)
sedentary activity	continuous (minutes of sedentary activity in typical day)
stroke	binary (ever been told you had a stroke? 0 = yes, 1 = no)
systolic	continuous (mm Hg)
tobacco	binary (used any tobacco product in last 5 days? 0 = yes, 1 = no)
triglycerides	continuous (mg/dL)
urination	continuous (minutes between last urination)
weak kidneys	binary (ever been told you had weak/failing kidneys? 0 = yes, 1 = no)
weight	continuous (kg)
weight loss	binary (doctor told you to control/lose weight? 0 = yes, 1 = no)
Outcome Variable	**Description**
glycohemoglobin	continuous (percentage)
Excluding Variable	**Description**
indicator of diabetes	binary (doctor told you that you have diabetes? 0 = yes, 1 = no)

Missing values for any input variable in the 2015–2016 and 2017–2018 cycles were replaced by the corresponding measure of centrality for that variable from the 2015–2016 cycle. For continuous input variables the mean was used. The mode was utilized for binary and nominal variables, and the median for ordinal and discrete independent variables. For example, consider sedentary activity, for each subgroup the average sedentary activity is computed from the 2015–2016 dataset, the dataset used to construct the model. This estimated value replaces all missing sedentary activity values for the same subgroup. We refer the reader to the Appendix for summary statistics of the diabetes data.

The analysis of the NHANES dataset requires adjustment for individual participant weights. Rather than using the usual sum of square error and mean squared prediction error, we utilize weighted sum of squares for determining the optimal input variable and its corresponding cut-point for each node splitting in a regression tree. Additionally, for the generalization error the weighted MSPE is used. We will now detail the modifications for the weighted analysis.

4.1 Weighting random forest approach

We now describe how weighting is incorporated into the two RF approaches for an application of survey data. To begin let w_C^i be the weight of ith participant in the training set belonging to node C of a regression tree, such that $\sum_{i=1}^{n} w^i = N$, where N is the size of the population of interest. The outcome of ith sample belonging to node C is y_C^i, $i = 1, \ldots, n$. The weights assigned to each observation represent the number of units in the population which are represented by the observation. For example, survey weights provided by NHANES account for various factors so that the sample data is representative of the United States population. We refer to [5] for details. The average outcome of all samples belonging to the same node is denoted by $\bar{y}_C$, and the number of samples belonging to the node is n_C. The nodes produced from a split are noted as C_L for left node and C_R for right node. Let the notation for weighted analysis be denoted by a W superscript.

Adjusting the RF algorithm to incorporate weights involves three modifications. Firstly, we adjust the sum of squares deviation criterion to a sum of squares weighted deviation criterion, denoted as SSE_C^W. The criterion is defined as $SSE_C^W = \sum_{i=1}^{n_C} w_C^i (y_C^i - \bar{y}_C)^2$. Secondly, in the usual RF algorithm the predicted value from a given tree in the forest is the average of the outcomes from the training set belonging to the same terminal node as specified in [4]. For the weighted RF algorithm, the average is replaced by the weighted average, $\frac{\sum_{i=1}^{n_C} w_C^i y_C^i}{\sum_{i=1}^{n_C} w_C^i}$.

The variable importance of an input variable in the RF model represents the relative influence that variable has on predicting the outcome. Larger values signify greater influence compared to smaller values. For the third modification, the variable importances are calculated as a function of the weighted sum of squares criterion rather than the usual sum of squares criterion. Let the importance from a single split for variable j and value cut-point k be denoted as: $\tau_C^{W(j)}$ such that the reduction in the sum of squares is $\tau_C^{W(j)} = SSE_C^W - [SSE_{C_L}^W(j,k) + SSE_{C_R}^W(j,k)]$.

For a single tree the importance for variable j is the sum of all $\tau_C^{W(j)}$s corresponding to variable j, denoted as $\lambda_m^{W(j)}$, for $m = 1 \ldots M$ (m is the mth tree). Let $\bar{\lambda}^{W(j)}$ denote the average of the variable importances from all M trees in the forest for the jth variable, defined as $\bar{\lambda}^{W(j)} = \frac{\sum_{m=1}^{M} \lambda_m^{W(j)}}{M}$. The variable importance for variable j from the RF model is made into a relative importance by the following scaling, $\Gamma^{W(j)} = \frac{\bar{\lambda}^{W(j)}}{\sum_{j=1}^{p} \bar{\lambda}^{W(j)}}$.

The generalization error used to evaluate the model performance is the weighted MSPE, a function of the testing set outcome y_i, its predicted value from the RF model $\tilde{f}_{rf}^W(\mathbf{x}_i)$, and its associated weight w^i, for $i = n+1 \ldots, n+\kappa$. The final weighted MSPE based on the testing set is

$$\widehat{MSPE}^W[\tilde{f}_{rf}(\mathbf{x}_i)] = \frac{\sum_{i=n+1}^{n+\kappa} w^i \left[y_i - \tilde{f}_{rf}^W(\mathbf{x}_i)\right]^2}{\sum_{i=n+1}^{n+\kappa} w^i}. \quad (4.1)$$

The weighted analysis prescribed in this section is applied to the NHANES diabetes data, described in the previous section. We now present the performance results for the diabetes application.

4.2 Analysis of the diabetes NHANES data

The analysis of the National Health and Nutrition Examination Survey (NHANES) diabetes dataset follows. Performance results for the diabetes dataset are based on the estimated weighted MSPE. The use of a weighted RF approach is needed to account for the survey sampling as provided in Section 4.1. The weights utilized in the analysis are from the NHANES dataset as discussed in beginning of Section 4. The aim of our analysis is to predict presence of diabetes in the United States population for persons 20 years and older who have been undiagnosed with the illness. The diabetes biomarker glycohemoglobin is the outcome for the analysis. The dataset used to train the weighted RF algorithm is the 2015–2016 cycle data. The test dataset used to evaluate the predictive performance of the algorithms is the 2017–2018 cycle dataset. Description of the dataset, imputation for missing data, and weighted RF approach are detailed in the beginning of Section 4.

For the analysis a total of 49 RF models were fit consisting of 7 m_{try} values and 7 node size values. The m_{try} values are: 3, 6, 9, 12, 15, 18, 19, the node size values are: 1, 3, 5, 7, 10, 15, 20, and the number of trees in each forest is $M = 500$. The estimated weighted MSPE, $\widehat{MSPE}^W$, is provided in (4.1) and is computed for each of the 49 fitted models of the parent and leaf approaches, respectively. The resulting estimated weighted MSPE results for each of the fitted models under the parent approach are reported

Table 2. Weighted MSPE results from the weighted parent approach for the NHANES dataset.

node size	m_{try}						
	3	6	9	12	15	18	19
1	0.16854	0.14938	0.14329	0.14148	0.14063	0.14077	0.14041
3	0.16917	0.14920	0.14313	0.14119	0.14025	0.14084	0.14043
5	0.16950	0.14811	0.14333	0.14080	0.14060	0.14025	0.14056
7	0.16815	0.14963	0.14388	0.14083	0.14018	0.14043	0.14078
10	0.16791	0.14977	0.14297	0.14029	0.14027	0.14067	0.14047
15	0.16930	0.15037	0.14337	0.14131	0.14054	0.14031	0.14097
20	0.16992	0.15010	0.14368	0.14196	0.14077	0.14118	0.14095

Table 3. Weighted MSPE results from the weighted leaf approach for the NHANES dataset.

node size	m_{try}						
	3	6	9	12	15	18	19
1	0.1685	0.14938	0.14329	0.14148	0.14063	0.14077	0.14041
3	0.17188	0.15023	0.14312	0.14060	0.13986	0.14035	0.14020
5	0.17219	0.15045	0.14314	0.14049	0.14010	0.13996	0.14003
7	0.17182	0.15235	0.14403	0.14130	0.13994	0.14034	0.14034
10	0.17430	0.15390	0.14562	0.14178	0.14043	0.14076	0.14092
15	0.17935	0.15691	0.14968	0.14435	0.14272	0.14214	0.14224
20	0.18098	0.16344	0.15290	0.14816	0.14636	0.14561	0.14555

in Table 2, and Table 3 under the leaf approach. Results obtained for both weighted RF approaches are calculated from the algorithms we have programmed. The `R` code and the NHANES data used are provided in the Supplementary Material.

From the results, both approaches provide similar overall predictive performance. In general, each method provides greater performance for higher values of m_{try} and larger node size values. When comparing the two models under fixed hyperparameters the parent approach tends to outperform the leaf approach, especially as the node size increases and the m_{try} value decreases. However, when considering smaller node size and larger m_{try} values the leaf approach tends to perform competitively or better. Specifically, when the data have low to moderate random noise (evident from the low $\widehat{MSPE}^W$) and model hyperparameters are the same for both approaches, the parent approach outperforms the leaf approach with regard to $\widehat{MSPE}^W$.

The main interest is to compare the two RF approaches based on their respective optimal models, which are the models selected for obtaining predictions. From the parent approach the lowest estimated weighted MSPE occurs for m_{try} equal to 15 and node size equal to 7. The leaf approach attains optimal performance for m_{try} equals 15 and node size equals 3. The comparison of these two optimal models is facilitated by the percentage difference, defined as $\frac{(\hat{\theta}_P - \hat{\theta}_L)100\%}{(\frac{\hat{\theta}_P + \hat{\theta}_L}{2})}$, such that $\hat{\theta}_P$ is the estimate from parent approach and $\hat{\theta}_L$ is the estimate from leaf approach. The resulting percentage difference of $\widehat{MSPE}^W$ from the respective optimal models is 0.2292%, indicating that the optimal model from the proposed leaf approach outperforms that from the parent approach.

Additionally, the importance of diabetes as a serious chronic disease has prompted our investigation into variable importances. Typically, variable importances are calculated based on the reduction of sum of square errors from each split in the forest. For our analysis the associated weights of each participant are accounted for by utilizing the weighted sum of square errors, over the usual sum of squared errors as detailed in Section 4.1.

From our analysis, the order of the nineteen variables in regard to their predictive influence on glycohemoglobin are largely the same between the two approaches. The exception is that the leaf approach has salt and risk for diabetes as the 15th and 16th most influential. In contrast the parent approach has the rank of these two variables switched. The relative influence based on the optimal model from the leaf approach is provided in Figure 2, and that from the parent approach is similar.

As one may expect glucose is the most influential input variable in predicting glycohemoglobin of the United States population over 20 years of age. This is followed by age, weight, triglycerides, cholesterol and so on. A plot of predicted glycohemoglobin as a function of glucose for the testing data is provided in Figure 3, where predictions are obtained from the optimal leaf model and indicate a positive relationship.

The variables identified as being most influential agree with the literature on risk factors for diabetes. [12] investigated the relationship between sedentary activity and hemoglobin A1c using NHANES survey data from

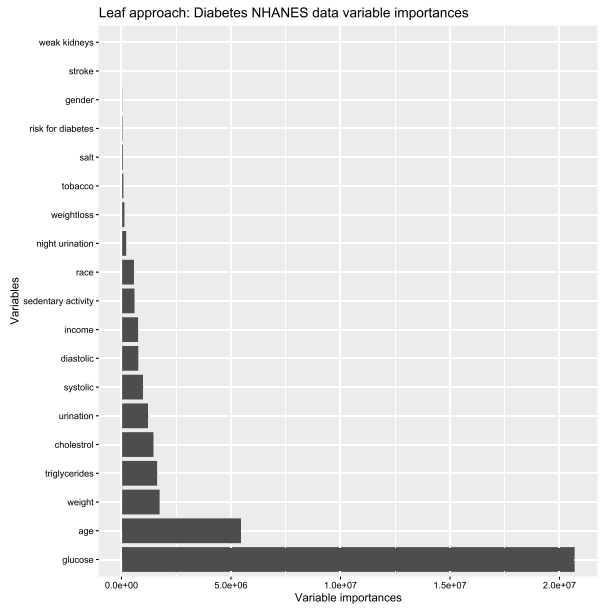


Figure 2: Variable importances for diabetes NHANES dataset under the optimal leaf approach model.

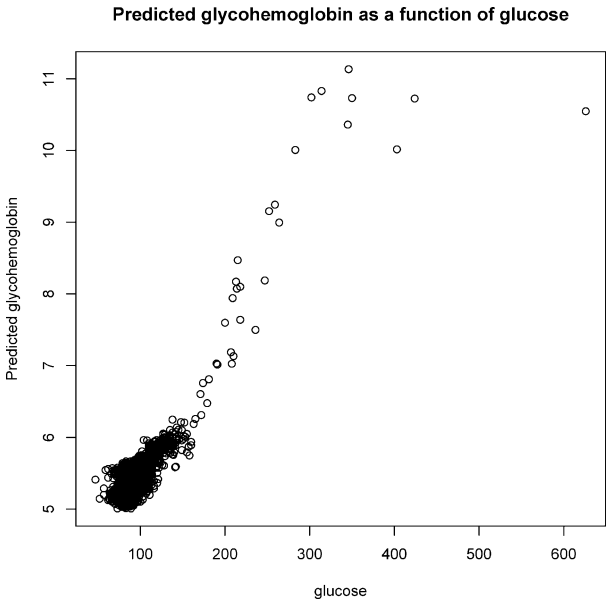


Figure 3: Predicted glycohemoglobin from optimal leaf model as a function of glucose from testing data.

2004–2005 and 2013–2014 and determined that prolonged sedentary activity is associated with elevated hemoglobin A1c levels. [13] analyzed NHANES data from 1999–2006 and found an association between increasing obesity and increased hemoglobin A1C levels. Analysis of the Third National Health and Nutrition Examination Survey (NHANES III) by [15] concluded that men diagnosed with diabetes had elevated odds of lower urinary tract symptoms. [1] note that hypertriglyceridemia is common in patients diagnosed with diabetes.

To summarize the results obtained for the NHANES data, the leaf and parent algorithms both identify the same variables as being most informative for predicting glycohemoglobin and the proposed leaf approach provides greater predictive performance. Further, each tree in the optimal leaf RF model uses at least three training set observations to obtain a tree prediction. In contrast the optimal parent model from the analysis does not have a minimum number of observations used for constructing each tree prediction, which may result in predictions based on only one training observation.

5. SIMULATION ANALYSIS

We shall now provide the design of the simulation analysis which will be used to compare the two competing random forest (RF) modelling approaches based on overall predictive performance, accuracy, and precision.

5.1 Design of simulation analysis

For the design of the simulation analysis a total of six true mean functions were utilized and are as follows:

True mean function 1:

$$
\begin{aligned}
&(X^{(1)},\ldots,X^{(20)}) \overset{i.i.d.}{\sim} Uniform(-5,5)\\
&f(\mathbf{X}) = \mathbb{1}[X^{(1)} \geq 3] + \mathbb{1}[X^{(2)} < 2.5] + \mathbb{1}[X^{(3)} \geq 1]\\
&\qquad + \mathbb{1}[X^{(4)} \geq 2.1] + \mathbb{1}[X^{(5)} < -2] + \mathbb{1}[X^{(6)} < -3.1]\\
&\qquad + \mathbb{1}[X^{(7)} \geq 4.3] + \mathbb{1}[X^{(8)} \geq 0.5] + \mathbb{1}[X^{(9)} < -0.7]\\
&\qquad + \mathbb{1}[X^{(10)} \geq 1.8] + \mathbb{1}[X^{(8)} \geq 1.5 \;\&\; X^{(10)} < 3.6]\\
&\qquad + \mathbb{1}[X^{(9)} < -0.5 \;\&\; X^{(10)} \geq 4.3] + \mathbb{1}[X^{(15)} < 3.6]\\
&\qquad + \mathbb{1}[0.5 \geq X^{(15)} < 4.6]
\end{aligned}
$$

True mean function 2:

$$
\begin{aligned}
&(X^{(1)},\ldots,X^{(10)}) \overset{i.i.d.}{\sim} Uniform(0,1)\\
&f(\mathbf{X}) = 10sin(\pi X^{(1)}X^{(2)}) + 20(X^{(3)} - 0.5)^2 + 10X^{(4)}\\
&\qquad + 5X^{(5)}
\end{aligned}
$$

True mean function 3:

$$
\begin{aligned}
&(X^{(1)},\ldots,X^{(5)}) \overset{i.i.d.}{\sim} Normal(\mu = 0, \sigma^2 = 1)\\
&f(\mathbf{X}) = sin(2\pi X^{(1)}) + cos(\pi X^{(2)})
\end{aligned}
$$

True mean function 4:

$$
(X^{(1)},\ldots,X^{(10)}) \overset{i.i.d.}{\sim} Uniform(0,1)
$$

$$
f(\mathbf{X}) = \begin{cases} 3 - 2X^{(1)} & \text{if } (X^{(2)} \geq 0.1)\\ 3 + X^{(4)} - 3X^{(2)} & \text{else if}(X^{(1)} \geq 0.3, X^{(2)} < 0.5)\\ 3 + 2X^{(5)} + X^{(3)} & \text{else} \end{cases}
$$

True mean function 5:

$$(X^{(1)},\ldots,X^{(10)}) \overset{i.i.d.}{\sim} Normal(\mu=0,\sigma^2=1)$$
$$f(\mathbf{X}) = 2X^{(1)} + (X^{(2)})^2 + 3(X^{(3)})^2 - 2X^{(4)}$$

True mean function 6:

$$(X^{(1)},\ldots,X^{(5)}) \overset{i.i.d.}{\sim} Uniform(0,1)$$
$$f(\mathbf{X}) = \begin{cases} 2 & X^{(1)} \leq 0.5 \;\&\; X^{(2)} \leq 0.5 \\ 4 & 0.5 < X^{(1)} \leq 0.75 \;\&\; 0.5 < X^{(2)} \leq 0.75 \\ 6 & o.w. \end{cases}$$

Table 4. Simulation analysis scenarios defined by combinations of $f(\mathbf{X})$ and σ^2.

Scenario	$f(\mathbf{X})$	σ^2	$\widehat{Var}(f(\mathbf{X}))$
1	1	0.25	2.82
2	2	1.00	25.00
3	3	0.05	1.01
4	4	0.05	0.60
5	5	1.00	26.90
6	6	0.25	2.95
7	1	1.00	2.82
8	2	5.00	25.00
9	3	0.25	1.01
10	4	0.25	0.60
11	5	5.00	26.90
12	6	1.00	2.95
13	1	5.00	2.82
14	3	5.00	1.01

From these six true mean functions a total of 14 simulated datasets are constructed, each using differing values of intrinsic variance. The simulation analysis is designed to investigate how differing levels of random error, i.e., variability of the random noise σ^2, effect the performance of the two approaches, such that $\epsilon_i \overset{i.i.d}{\sim} N(0,\sigma^2), i=1,\ldots,n+\kappa$.

The first six scenarios consist of small values of intrinsic variance and the second six consist of moderate values of the intrinsic variance. In addition to low and moderate levels of noise variability it is also of practical interest to compare the two random forest approaches when the variability of the random noise is larger than the variability of the true mean function. The later scheme corresponds to the last two scenarios. A table summarizing the scenarios as the combination of true mean function $f(\mathbf{X})$ and intrinsic variance σ^2 is shown in Table 4.

In Table 4 the estimated variance of the true mean function, $\widehat{Var}[f(\mathbf{X})]$, is computed from the entire dataset $T=\{T_n,T_\kappa\}$ such that $n=500$ and $\kappa=1000$. For each of the scenarios a total of 18 models were fit and the optimal model was selected based on the minimum estimated generalization error, MSPE defined by (3.10). A total of $M=1000$ regression trees were included each RF model.

Table 5. Hyperparameter combinations constituting 18 models fit for each scenario.

Scenario	m_{try} values	*node size* values
1	(5, 10, 15)	(1, 3, 5, 10, 15, 20)
2	(3, 5, 10)	(1, 3, 5, 10, 15, 20)
3	(2, 3, 5)	(1, 3, 5, 10, 15, 20)
4	(3, 5, 10)	(1, 3, 5, 10, 15, 20)
5	(3, 5, 10)	(1, 3, 5, 10, 15, 20)
6	(2, 3, 5)	(1, 3, 5, 10, 15, 20)
7	(5, 10, 15)	(1, 3, 5, 10, 15, 20)
8	(3, 5, 10)	(1, 3, 5, 10, 15, 20)
9	(2, 3, 5)	(1, 3, 5, 10, 15, 20)
10	(3, 5, 10)	(1, 3, 5, 10, 15, 20)
11	(3, 5, 10)	(1, 3, 5, 10, 15, 20)
12	(2, 3, 5)	(1, 3, 5, 10, 15, 20)
13	(5, 10, 15)	(1, 3, 5, 10, 15, 20)
14	(2, 3, 5)	(1, 3, 5, 10, 15, 20)

The 18 models constructed are based on a combination of three m_{try} values and six node size values. A summary table of the 18 models fit to each scenario is provided in Table 5.

For each of the fixed models in Table 5 a total of $B=100$ simulated models are generated to calculate the criteria, $\overline{\widehat{Bias^2}}$, $\overline{\widehat{Var}}$, and $\overline{\widehat{MSPE}}$. Each of the B simulated models only differs by the random seed value used to generate it. The hold-out approach is used instead of the out-of-bag error estimation approach in order to facilitate a comparison in which the statistics reported are based on the same dataset T_κ.

5.2 Results

In this section we provide the main results obtained from implementing the two competing RF approaches on each of the fourteen scenarios considered in Table 4. We first begin with the results for optimal models attained under each approach of the simulation study.

5.2.1 Comparison of the optimal models from the leaf and parent approaches

For each scenario, 18 models were fit and $\overline{\widehat{MSPE}}$ was calculated. The optimal model is then selected as that which provides the minimum $\overline{\widehat{MSPE}}$. The hyperparameters corresponding to the optimal parent and leaf models are provided in Table 6. The resulting estimates of MSPE, bias-squared, and variance from each optimal model are reported in Table 7 and Table 8 for the parent approach and leaf approach, respectively. The results for the parent approach are attained by implementing the `randomForest` package in `R` software. Estimates for the leaf approach are attained by applying code we have developed form this work. The definitions of $\overline{\widehat{Bias^2}}$, $\overline{\widehat{Var}}$, $\overline{\widehat{MSPE}}$, are defined in (3.7), (3.8), and (3.10), respectively.

Table 6. Optimal hyperparameters for each scenario under the parent approach and leaf approach.

Scenario	Parent approach		Leaf approach	
	node size	m_{try}	*node size*	m_{try}
1	3	10	3	10
2	1	10	1	10
3	1	3	1	3
4	3	10	5	10
5	1	10	1	10
6	3	5	1	5
7	3	10	3	10
8	1	10	1	10
9	1	3	1	3
10	10	5	10	10
11	5	10	1	10
12	10	5	3	5
13	1	5	5	5
14	20	5	20	5

Table 7. Estimates of MSPE, bias-squared, and variance from the optimal parent approach for each scenario.

Scenario	$\overline{\widehat{MSPE}}$	$\overline{\widehat{Bias^2}}$	$\overline{\widehat{Var}}$	$\overline{\widehat{Bias^2}} + \overline{\widehat{Var}}$
1	1.1609	0.8879	0.0037	0.8916
2	4.6453	3.5673	0.01556	3.5829
3	0.3284	0.2777	0.0012	0.2789
4	0.0672	0.0156	0.0001	0.0157
5	3.8382	2.9523	0.0146	2.9670
6	0.3485	0.1017	0.0012	0.1029
7	2.0202	0.9615	0.0053	0.9668
8	8.7899	3.6388	0.0235	3.6623
9	0.5652	0.3088	0.0016	0.3105
10	0.2949	0.0449	0.0005	0.0454
11	7.8723	3.2245	0.0218	3.2463
12	1.1685	0.1713	0.0023	0.1736
13	6.3454	1.1591	0.0153	1.1743
14	5.7438	0.5839	0.0066	0.5905

Table 8. Estimates of MSPE, bias-squared, and variance from the optimal leaf approach for each scenario.

Scenario	$\overline{\widehat{MSPE}}$	$\overline{\widehat{Bias^2}}$	$\overline{\widehat{Var}}$	$\overline{\widehat{Bias^2}} + \overline{\widehat{Var}}$
1	1.1602	0.8867	0.0033	0.8899
2	4.6512	3.5706	0.0156	3.5863
3	0.3288	0.2780	0.0012	0.2792
4	0.0666	0.0155	0.0001	0.0155
5	3.8393	2.9527	0.0147	2.9674
6	0.3484	0.1017	0.0012	0.1029
7	2.0143	0.9563	0.0046	0.9609
8	8.7939	3.6391	0.0236	3.6628
9	0.5650	0.3087	0.0016	0.3104
10	0.2890	0.0385	0.0002	0.0388
11	7.8664	3.2157	0.0230	3.2388
12	1.1664	0.1716	0.00239	0.1739
13	6.3247	1.1428	0.0103	1.1531
14	5.6196	0.4953	0.0029	0.4982

Table 9. Percentage difference of estimated MSPE, bias-squared, variance, and bias-squared plus variance from the two approaches.

Scenario	$\overline{\widehat{MSPE}}_{pd}$	$\overline{\widehat{Bias^2}}_{pd}$	$\overline{\widehat{Var}}_{pd}$	$(\overline{\widehat{Bias^2}} + \overline{\widehat{Var}})_{pd}$
1	0.06445	0.1427	12.0032	0.1891
2	−0.1271	0.0922	−0.4865	−0.0939
3	−0.1204	−0.1373	−0.4537	−0.1387
4	0.9592	0.8804	34.9609	1.1081
5	−0.0288	−0.0135	−0.1020	−0.0140
6	0.0040	0.0586	−1.8965	0.0352
7	0.2951	0.5388	13.7005	0.6068
8	−0.0453	−0.0098	−0.5772	−0.0134
9	0.0315	0.0439	−0.7250	0.03982
10	2.0164	15.2582	72.4678	15.7669
11	0.0752	0.2711	−5.35903	0.2322
12	0.1791	−0.1415	−1.0484	−0.1537
13	0.3262	1.4178	38.5253	1.826
14	2.1854	16.4332	77.8676	16.9709

The first column of Table 6 are the scenarios corresponding to Table 4. The optimal node size and m_{try} hyperparameters for the parent approach are reported in the second and third columns, respectively. The optimal node size and m_{try} hyperparameters for the leaf approach are reported in the fourth and fifth columns, respectively. The optimal RF model is defined by these hyperparameter combinations and $M = 1000$ trees.

From Table 6, the optimal hyperparameters are the same between the two RF approaches for scenarios 1, 2, 3, 5, 7, 8, 9, 14. From this subset of scenarios: 2, 3, 5, 8, and 9 are identical models. This is because the later have the same m_{try} values and have node size equal to 1 which results in the same algorithm for the two approaches. For these models any difference in the estimated criteria of MSPE, bias-squared, and variance, reported in Tables 7 and 8, are solely due to the randomness of the RF algorithm.

From comparing the results reported in Tables 7 and 8, it is clear that the optimal models from each RF approach are similar. In order to compare these results, the percentage difference for each of the estimated criteria was determined for each scenario. Let the percentage difference be defined as $\frac{(\hat{\theta}_P - \hat{\theta}_L)100\%}{(\frac{\hat{\theta}_P + \hat{\theta}_L}{2})}$, such that $\hat{\theta}_P$ is the estimate from parent approach and $\hat{\theta}_L$ is the estimate from leaf approach. For each of the scenarios the percentage difference of $\overline{\widehat{MSPE}}$, $\overline{\widehat{Bias^2}}$, $\overline{\widehat{Var}}$, and $\overline{\widehat{Bias^2}} + \overline{\widehat{Var}}$ are provided in Table 9. This means that the percentage difference reported in Table 9 are determined using the corresponding results in Table 7 and Table 8.

In Table 9 the $\overline{\widehat{MSPE}}_{pd}$ is the percentage difference of $\overline{\widehat{MSPE}}$ from the parent and leaf approaches. The $\overline{\widehat{Bias^2}}_{pd}$,

$\overline{\widehat{Var}}_{pd}$, and $(\overline{\widehat{Bias^2}} + \overline{\widehat{Var}})_{pd}$ are similarly defined. For example, $\overline{\widehat{MSPE}}_{pd}$ corresponding to scenario 1 is computed as $\frac{(1.160947-1.160199)100\%}{(\frac{1.160947+1.160199}{2})} \approx 0.0645$.

From the results in Table 9 the percentage difference of $\overline{\widehat{MSPE}}$ is positive for all scenarios with the exception of scenarios 2, 3, 5, and 8. As mentioned previously the optimal models for these scenarios are identical, so that any differences are due to randomness of the algorithm. Scenarios 1, 4, 6, 7, 10–14 do not result in identical models, and these scenarios have positive percentage difference. Hence, the comparison of $\overline{\widehat{MSPE}}$ indicates that the proposed leaf approach provides generalization error that is competitive to or smaller than that from the parent approach.

The resulting estimates of the bias-squared are positive for 10 of the 14 scenarios. In fact, from the scenarios for which this quantity is negative, scenarios 3, 5, 8, and 12, only scenario 12 does not result in the same exact algorithm. Therefore, even with regards to the bias-squared criterion the leaf approach tends to have similar or slightly better performance compared to the original RF method.

The percentage difference results for the variability of predictions tend to be negative when the magnitude of the percentage difference is low. However, when the magnitudes are larger, namely scenarios 1, 4, 7, 10, 13, 14, the parent approach results in considerably larger variability of the resulting predictions. Interestingly, the magnitude of $\overline{\widehat{Var}}_{pd}$ tends to increase for a fixed mean function as the intrinsic variance increases. This demonstrates that the difference in prediction variability between the two RF approaches grows as the intrinsic variance increases. Clearly there are scenarios for which the prediction variability from the leaf approach is substantially smaller than those from the parent approach. Further, for any scenario, the magnitude of the percentage difference is larger for variance compared to the percentage difference of the bias-squared.

The results of the percentage difference pertaining to $\overline{\widehat{Bias^2}} + \overline{\widehat{Var}}$ show that all scenarios having negative percentage difference are cases when the models are identical algorithms, with the exception of scenario 12. In particular the sign of the last column is in agreement with the percentage differences of $\overline{\widehat{MSPE}}$ in the first column, apart from scenario 12. Therefore, the results of $\overline{\widehat{Bias^2}} + \overline{\widehat{Var}}$ (estimate of $MSPE$ without intrinsic variance) are in good agreement with the estimated $MSPE$. Further, the magnitude of the percentage difference of $\overline{\widehat{MSPE}}$ and $\overline{\widehat{Bias^2}} + \overline{\widehat{Var}}$ for scenario 12 is near zero, indicating the performance of the approaches is competitive for that scenario.

Summarizing the results in Table 9, the estimated $MSPE$ for the predictions is lower under the leaf approach (omitting the scenarios for which the two algorithms are identical) compared to the parent approach. For many scenarios the estimated bias-squared of predictions is also larger under the original RF algorithm. Furthermore, results from six of the fourteen scenarios indicate that the proposed leaf approach reduces prediction variability substantially.

Table 10. Suggested random forest modelling approach for each scenario.

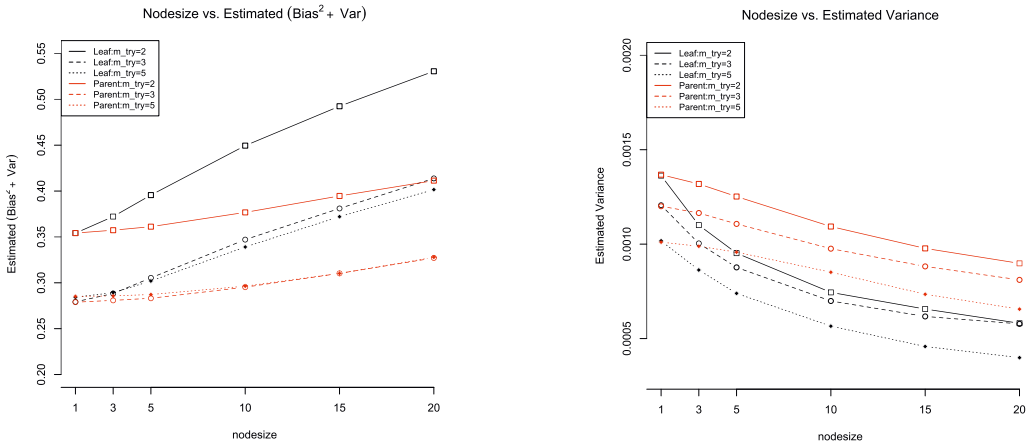
Scenario	Approach
1	Either
2	Either
3	Either
4	Leaf
5	Either
6	Either
7	Leaf
8	Either
9	Either
10	Leaf
11	Either
12	Either
13	Leaf
14	Leaf

To establish if the leaf algorithm provides substantial improvement over the parent algorithm one may take a conservative approach utilizing the results from Table 9. We note that the amount of conservatism is at the discretion of the analyst. All scenarios having $\overline{\widehat{Var}}_{pd} \geq 10\%$ and $\overline{\widehat{Bias^2}}_{pd} \geq 0.50\%$ also have $\overline{\widehat{MSPE}}_{pd} \geq 0.30\%$. Hence one may in practice determine that there is substantial gain by using the leaf approach when $\overline{\widehat{MSPE}}_{pd} \geq 0.30\%$. We use this conservative approach to determine when the leaf approach provides considerably better predictive performance overall, the results are provided in Table 10.

After taking a conservative approach, the RF that utilizes the leaf stopping criterion (based on the size of terminal nodes) performs as well or favorably to the parent stopping criterion (based on the size of the parent nodes). Additionally, as mentioned in Section 2, the leaf approach is also advantageous in ensuring that the resulting predictions from the model are obtained by a minimum number of observations from each tree model. In the next section the two approaches are compared under fixed models.

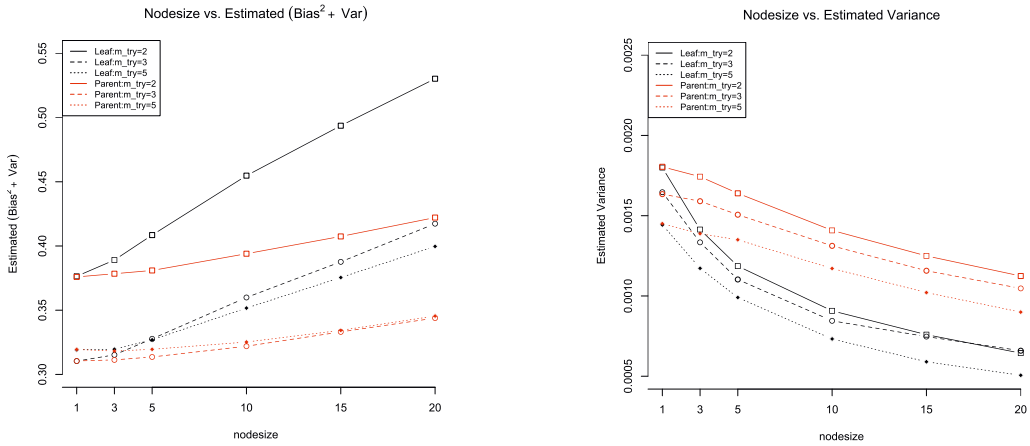
5.2.2 Effect of hyperparameters on predictive performance

We now explore the implications of the two RF method more broadly. Specifically, summary of all 18 models fit for each scenario are reported, not only for the optimal model. Based on the results from all scenarios, the estimates of $\overline{\widehat{Bias^2}} + \overline{\widehat{Var}}$ are dominated by $\overline{\widehat{Bias^2}}$. When all hyperparameters are fixed, scenarios having small and moderate values of intrinsic variance σ^2, have a bias-variance trade-off between the two RF approaches. As the node size parameter increases, the original RF approach provides predictions that have smaller bias-squared and larger variability compared to that obtained from the leaf approach.



(a) $\widehat{\overline{Bias^2}} + \widehat{\overline{Var}}$ *results* (b) $\widehat{\overline{Var}}$ *results*

Figure 4: Scenario 3 results of $\widehat{\overline{Bias^2}} + \widehat{\overline{Var}}$ and $\widehat{\overline{Var}}$ under two RF approaches for 18 models as function of node size.



(a) $\widehat{\overline{Bias^2}} + \widehat{\overline{Var}}$ *results* (b) $\widehat{\overline{Var}}$ *results*

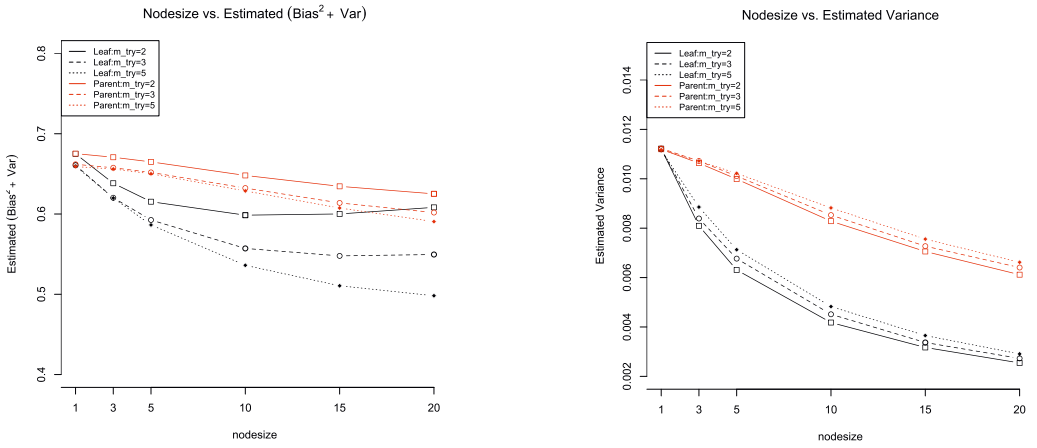
Figure 5: Scenario 9 results of $\widehat{\overline{Bias^2}} + \widehat{\overline{Var}}$ and $\widehat{\overline{Var}}$ under two RF approaches for 18 models as function of node size.

However, for the last two scenarios 13 and 14 which are consistent with very noisy data (the variability of the error is larger than that of the true mean function) the bias-variance trade-off does not hold. Rather, the leaf approach provides lower bias-squared and lower variance of predictions compared to the parent approach when the model hyperparameters are the same. Hence, the results indicate that the leaf approach provides more accurate and more precise results when applied to data with substantial error variance.

We present results for the third mean function corresponding to scenarios 3, 9, and 14. These scenarios have small, moderate, and large variance of the random noise, respectively. Results of the remaining scenarios follow similarly. The graphical results for scenarios 3, 9, and 14 correspond to Figure 4, Figure 5, Figure 6, respectively. The black trends are estimates obtained from the leaf approach and red from the parent approach. For each approach 18 models are considered corresponding to the three values of m_{try} and six values of node size provided in Table 5.

The variability of the true mean function corresponding to these scenarios is 1.01, as reported in Table 4. The intrinsic variance for scenarios 3, 9, and 14 are 0.05, 0.25, and 5.00, respectively. The bias-variance trade-off as a function of node size is present for scenarios 3 and 9, corresponding to Figure 4 and Figure 5. For this particular mean function, larger values of m_{try} result in lower bias-squared and variance of predictions.

Additional analysis was conducted on the resulting estimates of bias-squared and variance as a function of the parameter m_{try}. Table 11 and Table 12 summarize the ef-



(a) $\overline{\widehat{Bias^2}} + \overline{\widehat{Var}}$ *results* (b) $\overline{\widehat{Var}}$ *results*

Figure 6: Scenario 14 results of $\overline{\widehat{Bias^2}} + \overline{\widehat{Var}}$ and $\overline{\widehat{Var}}$ under two RF approaches for 18 models as function of node size.

Table 11. Summary of the effects of m_{try} on bias^2+ variance and variance for the parent approach.

Scenario	$\overline{\widehat{Bias^2}} + \overline{\widehat{Var}}$ smallest → largest	$\overline{\widehat{Var}}$ smallest → largest
1	(10, 15, 5)	(15, 10, 5)
2	(10, 5, 3)	(10, 5, 3)
3	(3, 5, 2)	(5, 3, 2)
4	(10, 5, 3)	(10, 5, 3)
5	(10, 5, 3)	(10, 5, 3)
6	(5, 3, 2)	(5, 3, 2)
7	(10, 5, 15)	(15, 10, 5)
8	(10, 5, 3)	(10, 5, 3)
9	(3, 5, 2)	(5, 3, 2)
10	(10, 5, 3)	(10, 5, 3)
11	(10, 5, 3)	(10, 5, 3)
12	(5, 3, 2)	(5, 3, 2)
13	(15, 10, 5)	(15, 10, 5)
14	(5, 3, 2)	(2, 3, 5)

Table 12. Summary of the effects of m_{try} on bias^2+ variance and variance for the leaf approach.

Scenario	$\overline{\widehat{Bias^2}} + \overline{\widehat{Var}}$ smallest → largest	$\overline{\widehat{Var}}$ smallest → largest
1	(15, 10, 5)	(15, 10, 5)
2	(10, 5, 3)	(10, 5, 3)
3	(5, 3, 2)	(5, 3, 2)
4	(10, 5, 3)	(10, 5, 3)
5	(10, 5, 3)	(10, 5, 3)
6	(5, 3, 2)	(5, 3, 2)
7	(10, 15, 5)	(15, 10, 5)
8	(10, 5, 3)	(10, 5, 3)
9	(5, 3, 2)	(5, 3, 2)
10	(10, 5, 3)	(10, 5, 3)
11	(10, 5, 3)	(10, 5, 3)
12	(5, 3, 2)	(5, 3, 2)
13	(5, 10, 15)	(15, 10, 5)
14	(5, 3, 2)	(2, 3, 5)

fect m_{try} on the bias^2 + variance and variance of predictions from the parent and leaf approach, respectively. The trends for bias-squared are similar to those of bias^2+variance, as bias-squared dominates this quantity.

Each cell in Table 11 and Table 12 lists the sorted m_{try} resulting in the smallest to largest value of the corresponding criterion for each scenario. For example, consider the estimate $\overline{\widehat{Bias^2}} + \overline{\widehat{Var}}$ for scenario 1 in Table 11, $m_{try} = 10$ results in the smallest $\overline{\widehat{Bias^2}} + \overline{\widehat{Var}}$ followed by $m_{try} = 15$, and the largest estimate results from $m_{try} = 5$ for the parent approach. What emerges from the results is that in all but one scenario increasing m_{try} produces lower variance. The exception is for scenario 14, corresponding to the true mean function 3 and having large intrinsic variance. For this scenario the smallest values of m_{try} provided the smallest variability of predictions for both approaches.

In regard to $\overline{\widehat{Bias^2}} + \overline{\widehat{Var}}$, the default to larger values of m_{try} resulted in the lower estimates. For only one of the scenarios, scenario 13, did the smallest value of m_{try} result in a lower estimate of bias^2+ variance. This scenario was one of the two in the simulation study having large intrinsic variance. Therefore, the general conclusion is that moderate to large values of m_{try} result in lower prediction variability and bias^2, and this appears to hold for data having small and moderate values of intrinsic variance. However, data having large intrinsic variance appear to diverge from these trends.

The notion that a regression RF model may require a larger m_{try} value compared to a classification RF has been

noted by [4, pg. 27]. Our results support this statement generally for reasonable quality data that are not dominated by random error. Considering that the $MSPE$ of predictions from the RF algorithm are largely due to the bias term one may choose moderate to larger m_{try} values. Further, none of the optimal models in Table 6 are defined by small values of m_{try}.

6. CONCLUSION

The increasing popularity of the regression random forest (RF) algorithm necessitates examination of its predictive performance based on the hyperparameters defining the model. There has been relatively little investigation in the literature thus far on the effect of the node size hyperparameter in particular, despite the fact that it largely determines the predictive performance of the algorithm as indicated by [14]. In the present paper we have compared two existing RF approaches with differing definitions of node size. Importantly, we have presented an investigation on how the definition of node size affects predictive performance.

In summarizing the comparison of the two approaches under fixed hyperparameters, the original RF approach provides larger prediction variability and smaller predicted squared bias than the proposed approach. This trend is evident as the node size hyperparameter increases and when data have levels of random noise that are low to moderate in comparison to the variability of the mean function. However, our results indicate that the bias-variance trade-off does not hold for poor quality data, defined as data with substantial intrinsic variability. In these later scenarios the simulation results indicate that the leaf approach outperformed the original random forest approach in both variability and squared bias.

In contrast, the bias-variance trade-off is not present when comparing the two algorithms based on their respective optimal models. The estimated generalization error from the optimal model of the parent approach was consistently larger than that from the leaf approach in our simulation analysis. Further, the bias-squared from the leaf approach was often smaller than that obtained under the original random forest. Additionally, when the prediction variability was smaller under the leaf algorithm it tended to be substantially so in comparison to scenarios for which the prediction variability was smaller under the original algorithm. Summarizing the results pertaining to the optimal models, the predictive performance under the leaf approach is increasingly competitive to that of the original approach as the ratio of random noise variability to outcome variability increases.

Additionally, our results suggest that the value of m_{try} resulting in the lowest generalization error tends to range from the default value or larger, confirming other research on this hyperparameter. However, when the data have high levels of noise these general trends may not hold. This result underscores the need for hyperparameter tuning of the model used for prediction.

The application of data-driven algorithms, such as the RF algorithms, to investigate health related research in the presence of large and complex data is becoming increasingly popular. Incorporating participant weights in the presence of survey data was prompted by the analysis of the National Health and Nutrition Examination Survey (NHANES) diabetes dataset. The application of the two approaches indicates that there is marginal predictive performance gained by application of the leaf approach over the parent approach. Our results demonstrate that both algorithms identify known risk factors of diabetes as most influential.

Our contribution from this work is two-fold. Firstly, we have demonstrated that the leaf node size approach is increasingly competitive to the original RF model with respect to predictive performance as the ratio of random noise variability to outcome variability increases. Conclusions are based on three statistical criteria: accuracy measured by bias-squared, precision measured by variance, and overall predictive performance measured by the generalization error. Secondly, we have developed a straight-forward method for incorporating participant weights in the presence of survey data, which is of considerable practical relevance for application of the RF algorithm.

Results presented here pertain to the regression context when number of input features is considerably less than the number of sampled units in the dataset. Extensions of this work can provide greater understanding on the complex relationship between predictive performance of the regression random forest model and its hyperparameters. Importantly the sensitivity of the random forest algorithm to various imputation techniques for missing covariate values is a topic of future research. In particular, one may consider comparison of single and multiple imputation methods on the accuracy and precision of the random forest algorithm. Additionally, generalizations of the criteria for stopping regression tree expansion may be considered. Some alternative node size stopping criterion based on dispersion measures of the outcome variable are explored in [9]. Another relevant extension for future work is the comparison of predictive performances from competing regression random forest algorithms on stratified sub-group analysis applications.

SUPPLEMENTARY MATERIAL

The supplementary material provides code in R software for implementing the algorithms developed in this work. The National Health and Nutrition Examination Survey data utilized in the paper are also provided.

APPENDIX

In this appendix we include descriptive statistics of input variables and the outcome glycohemoglobin from the diabetes NHANES application, followed by plots for model assessment.

Table A.1. NHANES variables and summary statistics for 2017–2018 data.

Input Variable	**Measure of centrality**	**Measure of dispersion**	**% Missing**
age	48.92	17.63	0.00
cholesterol	191.05	40.04	2.25
diastolic	72.41	13.33	10.98
income	9.00	3.00	9.08
gender	1.00	N/A	0.00
glucose	94.6617	20.54	2.30
night urination	1.00	1.00	6.74
race	3.00	N/A	0.00
risk for diabetes	1.00	N/A	0.00
salt	1.00	N/A	0.049
sedentary activity	323.86	196.17	0.67
stroke	1.00	N/A	0.12
systolic	125.21	19.03592	10.98
tobacco	1.00	N/A	6.44
triglycerides	139.57	109.99	2.30
urination	147.09	109.29	13.28
weak kidneys	1.00	N/A	0.15
weight	81.33	21.98	1.23
weight loss	1.00	N/A	0.02
Outcome Variable	**Measure of centrality**	**Measure of dispersion**	**% Missing**
glycohemoglobin	5.52	0.59	0.00

Table A.2. NHANES variables and summary statistics for 2015–2016 data.

Input Variable	**Measure of centrality**	**Measure of dispersion**	**% Missing**
age	47.25	17.44	0.00
cholesterol	195.71	41.92	0.96
diastolic	70.16	12.53	5.34
income	9.00	3.00	6.24
gender	1.00	N/A	0.00
glucose	95.75	22.81	0.89
night urination	1.00	1.00	9.67
race	3.00	N/A	0.00
risk for diabetes	1.00	N/A	0.00
salt	1.00	N/A	0.049
sedentary activity	367.38	196.44	0.93
stroke	1.00	N/A	0.07
systolic	124.33	17.92	5.34
tobacco	1.00	N/A	9.48
triglycerides	152.12	130.26	0.96
urination	143.81	99.24	9.41
weak kidneys	1.00	N/A	0.09
weight	80.36	21.04	0.75
weight loss	1.00	N/A	0.00
Outcome Variable	**Measure of centrality**	**Measure of dispersion**	**% Missing**
glycohemoglobin	5.55	0.60	0.00

Summary statistics

An overview of The National Health and Nutrition Examination Survey (NHANES) data for the prediction of the diabetes biomarker glycohemoglobin percentage is provided in Section 4 of the paper. For our analysis 19 input variables were utilized and described in Table 1 of the paper. The dataset used to construct the model was the 2015–2016 cycle and for model assessment the 2017–2018 cycle was utilized. The analyses conducted pertain to the United States population who have been undiagnosed with diabetes and are over 20 years of age. The following two tables are the summary statistics for the input variables and outcome variable for each cycle. The 19 input variables consist of continuous, discrete, ordinal, and nominal data (including binary data). The measure of centrality reported for continuous data is the arithmetic mean, for discrete and ordinal data it is the median, and for categorical data is the mode. For measure of dispersion the standard deviation is used for continuous variables, median absolute deviation (MAD) for discrete and ordinal data. The estimated MAD is defined as: $median(|x_i - q_{0.5}|)$, such that $q_{0.5}$ is the median of $x_i, i = 1, \ldots, n$. Additionally, the percentage of missing data are also reported in each table. The estimated quantities in the following tables are based on data prior to imputation.

From the summary statistics reported in Table A.1 and Table A.2 we find that the datasets used for training and testing are largely similar.

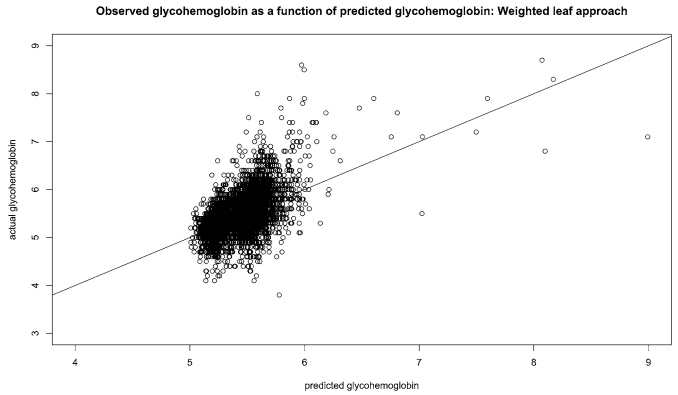


Figure A.1: Observed glycohemoglobin as a function of predicted glycohemoglobin from the optimal leaf model.

Graphical results

The optimal model for modelling the NHANES diabetes data was the leaf approach having hyperparameter values: $M = 500$, $m_{try} = 15$ and node size = 3. The following plot shows the observed outcome glycohemoglobin as a function of its predicted values from the optimal leaf model. Plot for the optimal parent model is similar to that shown in Figure A.1.

Clearly, predicted glycohemoglobin tend to range over a shorter interval than the observations they estimate, a common result when applying random forest models. This is due to the fact that the random forest algorithm estimates ex-

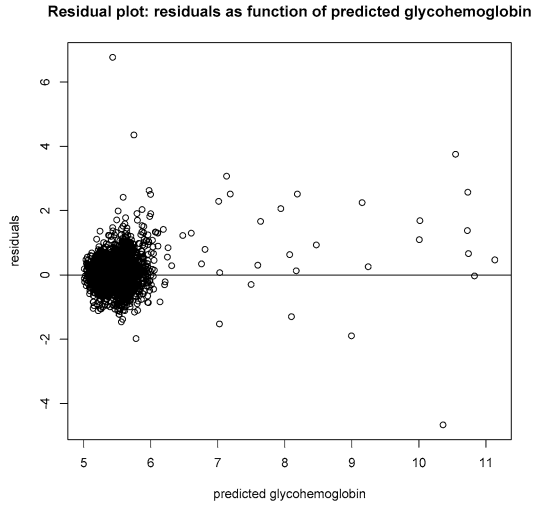


Figure A.2: Residuals as a function of predicted glycohemoglobin from optimal leaf model.

treme values by the average outcomes, which are closer to the mean response f. As a result, large values of the mean function f tend to be underestimated while small values of f tend to be overestimated. To further examine model fit we provide the residual plot in Figure A.2.

As seen in Figure A.2, majority of the predicted values tend to range from 5 to 6 with some values being larger. In general, the condition of constant variance, homoscedasticity, holds with the exception of one small predicted value near 5.5 and one large value around 10.5.

ACKNOWLEDGEMENTS

We are grateful to the comments and suggestions from the reviewers that have enhanced the paper.

FUNDING

This work was supported by the Intramural Research Program of the National Institutes of Health, National Cancer Institute, Division of Cancer Epidemiology and Genetics and utilized the computational resources of the NIH HPC Biowulf cluster. (http://hpc.nih.gov)

The Intramural Research Program of the National Institutes of Health, the National Cancer Institute, Division of Cancer Epidemiology and Genetics supported the work of all authors.

Accepted 18 July 2022

REFERENCES

[1] Alexopoulos, A., Qamar, A., Hutchins, K., Crowley, M. J., Batch, B. C. and R., G. J. (2019). Triglycerides: Emerging Targets in Diabetes Care? Review of Moderate Hypertriglyceridemia in Diabetes. *Current diabetes reports* **19** 13.

[2] Biau, G. and Devroye, L. (2010). On the layered nearest neighbour estimate, the bagged nearest neighbour estimate and the random forest method in regression and classification. *Journal of Multivariate Analysis* **101** 2499–2518. https://doi.org/10.1016/j.jmva.2010.06.019. MR2719877

[3] Biau, G. and Scornet, E. (2016). A random forest guided tour. *TEST* **25** 197–227. https://doi.org/10.1007/s11749-016-0481-7. MR3493512

[4] Breiman, L. (2001). Random Forests. *Machine Learning* **45** 5–32. MR3874153

[5] Chen, T. C., Clark, J., Riddles, M. K., Mohadjer, L. K. and Fakhouri, T. H. I. (2020). National Health and Nutrition Examination Survey, 2015–2018: Sample Design and Estimation Procedures. *Vital Health Stat 2* **184** 1–35.

[6] Hastie, T., Tibshirani, R. and Friedman, J. H. (2017) *The elements of statistical learning: Data mining, inference, and prediction*. Springer, New York. https://doi.org/10.1007/978-0-387-84858-7. MR2722294

[7] Ishwaran, H. and Kogalur, U. B. (2021). Package 'randomForestSRC'. https://doi.org/cran.r-project.org/web/packages/randomForestSRC/randomForestSRC.pdf.

[8] Liaw, A. and Wiener, M. (2018). Package 'randomForest'. https://doi.org/cran.r-project.org/web/packages/randomForest/randomForest.pdf.

[9] Liitle, M., Rosenberg, P. and Arsham, A. (2022). Alternative stopping rules to limit tree expansion for random forest models. *Scientific Reports* (Resubmitted).

[10] Makris, K. and Spanou, L. (2011). Is there a relationship between mean blood glucose and glycated hemoglobin? *J Diabetes Sci Technol* **5**(6) 1572–1583.

[11] Mendola, N. D., Chen, T. q. C., Gu, Q., Eberhardt, M. S. and Saydah, S. (2018). Prevalence of Total, Diagnosed, and Undiagnosed Diabetes Among Adults: United States, 2013–2016. *NCHS Data Brief* 1–8.

[12] Nicolo, M. L., Shewokis, P. A., Boullata, J., Sukumar, D., Smith, S., Compher, C. and Volpe, S. L. (2019). Sedentary behavior time as a predictor of hemoglobin A1c among adults, 40 to 59 years of age, living in the United States: National Health and Nutrition Examination Survey 2003 to 2004 and 2013 to 2014. *Nutrition and Health* **25** 275–279. https://doi.org/10.1002/sim.7049. MR3569919

[13] Ninh, T., Nguyen, X. T., Lane, J. and Wang, P. (2011). Relationship between obesity and diabetes in a US adult population: findings from the National Health and Nutrition Examination Survey, 1999–2006. *Obesity Surgery* **21** 351–355.

[14] Probst, P., Wright, M. N. and Boulesteix, A. (2019). Hyperparameters and tuning strategies for random forest. *WIREs Data Mining and Knowledge Discovery* **9** 1301.

[15] Rohrmann, S., Smit, E., Giovannucci, E. and Platz, E. A. (2005). Association between markers of the metabolic syndrome and lower urinary tract symptoms in the Third National Health and Nutrition Examination Survey (NHANES III). *International journal of obesity* **29** 310–316.

[16] Scornet, E. (2018). Tuning parameters in random forests. *ESAIM: Proceedings and Surveys* **60** 144–162. https://doi.org/10.1051/proc/201760144. MR3772478

[17] van Rijn, J. N. and Hutter, F. (2018). Hyperparameter importance across datasets. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

[18] *Centers for Disease Control and Prevention (CDC). National Center for Health Statistics (NCHS). National Health and Nutrition Examination Survey Data. Hyattsville, MD: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, [2017–2018].* http://wwwn.cdc.gov/Nchs/Nhanes/continuousnhanes/default.aspx?BeginYear=2015.

[19] *Centers for Disease Control and Prevention (CDC). National Center for Health Statistics (NCHS). National Health and Nutrition Examination Survey Data. Hyattsville, MD: U.S. Depart-*

ment of Health and Human Services, Centers for Disease Control and Prevention, [2015–2016]. http://wwwn.cdc.gov/Nchs/Nhanes/continuousnhanes/default.aspx?BeginYear=2017.

Aryana Arsham. Center for Data, Mathematical & Computational Sciences, Integrative Data Analytics, Goucher College, USA. E-mail address: aryana.arsham@goucher.edu

Philip Rosenberg. Biostatistics Branch, National Cancer Institute, USA. E-mail address: rosenbep@exchange.nih.gov

Mark Little. Radiation Epidemiology Branch, National Cancer Institute, USA. E-mail address: mark.little@nih.gov