5-2015

A Comparison of Random Forest-Based Methods for Racial/ Ethnic-Specific Classification of Obesity

Sun Young Jeon
Utah State University

A COMPARISON OF RANDOM FOREST-BASED METHODS FOR RACIAL/ETHNIC-SPECIFIC CLASSIFICATION OF OBESITY

by

Sun Young Jeon

A report submitted in partial fulfillment
of the requirements for the degree

of

MASTER OF SCIENCE

in

Statistics

Approved:

Dr. John R. Stevens Major Professor	Dr. Adele Cutler Committee Member

Dr. Eric N. Reither
Committee Member

UTAH STATE UNIVERSITY
Logan, Utah

2015

Abstract

A Comparison of Random Forest-Based Methods for Racial/Ethnic-Specific Classification of Obesity

by

Sun Young Jeon, Master of Science

Utah State University, 2015

Major Professor: Dr. John R. Stevens
Department: Mathematics and Statistics

Obesity is typically defined using body mass index (BMI) and its established cut-off. However, some studies have highlighted the importance of developing racial/ethnic-specific classifications of obesity that reflect different body compositions and fat distributions. Using National Health and Nutrition Examination Survey (NHANES) data and Random Forest classification, this paper attempts to identify important body measures and cut-offs for predicting obesity-related health risks among White, Hispanic and Black male populations in the U.S. In particular, this paper compares the performance of three Random Forest-based methods for dealing with class imbalance: weighted Random Forest (WRF), Random Forest with down-sampling (DS), and Random Forest with SMOTE. Of the three methods, the best performing one turned out to be different for each population in the given dataset. Thus, WRF for Whites, Random Forest with SMOTE for Hispanics, and Random Forest with DS for Blacks are used as the final models for the classification. The results show that BMI is indeed an important body measure for predicting obesity-related health risks among White males, but is considerably less informative for Hispanic and Black males. On the other hand, using waist circumference along with population-specific cut-offs turned out to be more useful in predicting obesity-related health risks for these two populations.

(54 pages)

Acknowledgments

Foremost, I would like express my sincerest gratitude to my committee professors, Drs. John R. Stevens, Adele Cutler, and Eric N. Reither for their insightful comments, feedback, and recommendations. I especially thank my major professors Drs. John R. Stevens and Eric. N. Reither for fully supporting the concurrent degree plan. Without their patience and encouragement, I could not have finished this project. Besides my committee professors, I am grateful to the faculty, staff, and graduate students in the Department of Mathematics and Statistics and the Department of Sociology.

Lastly, I would like to thank my family: Dad and Mom who are the most supportive parents in this world, my sister Jeeyoung and brother-in-law Seunggeun who are always my best friends, and my boyfriend Pedro who is a partner of my life and academic journey.

Sun Y. Jeon

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Obesity is a health condition of having excess body fat, which can elevate the risks of multiple health problems such as diabetes, hypertension, and cardiovascular diseases [1]. A typical way to define obesity is by using the body mass index (BMI) and its established cut-offs. The BMI of an individual is simply calculated as its weight in kilograms, divided by the square of its height in meters, and the cut-offs suggested by National Institutes of Health (NIH) for defining overweight and obesity are $25kg/m^2$ and $30kg/m^2$, respectively [2].

BMI can be an informative indicator of obesity, based on one fundamental assumption: a variation of relative weight to height is derived from the variation in the amount of body fat [1]. Consequently, BMI reflects changes of health risks. Although research has shown that this is a quite sound assumption based on the moderate correlation between BMI and body fat [3], there have been growing concerns about its limitations [4–6]. Recently, researchers have argued that using BMI and its cut-offs does not differentiate lean mass ("healthy weight") from fat mass [7], and does not explain concentrated fat in specific regions that may be particularly adverse for one's health (e.g., abdominal fats [8,9]). These arguments indicate that it may be important to take into account body composition rather than entire body weight, and distribution of body fats rather than total amount of fat in defining obesity.

In particular, when it comes to defining obesity for a diversity of populations, previous studies have come up with evidence that highlights the limitation of the BMI. Luke et al. [10] pointed out that Blacks tend to have a higher percentage of body fat at the same BMI as Whites, so the same BMI can represent significantly different amounts of body fat depending on population. Moreover, Lovejoy et al. [11] highlighted that what really matters in predicting health risks is the locations where body fats are concentrated, and because

the pattern of fat distribution among body parts differs by types of population, the body measures that accurately predict health risks can also be population-specific.

In order to not miss such distinctive racial/ethnic characteristics, it is important to develop a better racial/ethnic-specific classification of obesity. Previous studies have made efforts on this problem in two ways. Some studies have suggested using other body measures to complement the BMI [12], while other studies have suggested using BMI with population-specific cut-offs [13–15]. For example, Janssen et al. [12] showed that using waist circumference can be more accurate than using BMI for predicting obesity-related health risks in the Black population. On the other hand, a large number of studies have pointed out that using universal cut-offs for BMI is not appropriate to explain the obesity in different populations [14], and the cut-off should be adjusted to be population-specific [15].

This study adds to the existing approaches for developing a racial/ethnic-specific method of obesity classification by using various body measures and employing a unique statistical approach. The study tests if other body measures suggested by previous research such as skinfold thickness, and more direct measures of body fats that are estimated by Dual Energy X-ray Absorptiometry (DXA) scans [1,16], can be particularly informative for predicting obesity in a certain population. This requires an appropriate statistical approach to handle correlated but important measures.

When testing multiple body measures using a model such as multiple regression, one statistical issue that is likely to be observed is the high correlation between some of those measures. This is referred to as multi-collinearity, and can threaten study results. Since some of the highly correlated variables should be excluded from the model to obtain a valid result, it is challenging to compare the explanatory power of highly correlated but important body measures in one statistical model. This is a problem that happens with regression, and most previous studies use multiple regression, so they are limited to using a single measure or using only uncorrelated measures.

My approach in this study to solve the problem of multi-collinearity is to employ Random Forest (RF) classification [17], which is known to disperse the variable importance

by using sub-selection of predictor variables at each node of classification trees [18]. I also discuss one of the common problems in the application of RF to health data class imbalance, and how to handle it using weighted Random Forest (WRF) [19], down-sampling (DS) [20], and the synthetic minority over-sampling technique (SMOTE) procedures [21].

After evaluating the performance of each method, I choose one best performing method for each population to answer the two following research questions:

1. What are the important indicators in classification of obesity-linked diseases, and how do they differ by racial/ethnic groups?

2. What should be the racial/ethnic-specific cut-offs for the important indicators of obesity based on the probability of having obesity-linked diseases?

Chapter 2

Methods

2.1 Data and Variables

National Health and Nutrition Examination Surveys (NHANES) 2005–2006 data is used for this paper [22]. NHANES provides rich information about health status that is measured not only by questionnaires, but also by laboratory and examination. Data from 2005–2006 was selected because, currently, it is the latest survey that provides results of the whole body DXA scans, which include measures of percent and masses of body fat for the entire body or parts of a body. This study will particularly focus on male adults who are aged between 18 and 69.

Eleven body measures were obtained to predict obesity-related health risks. Six of them that were collected by trained health technicians are BMI, circumferences of waist, thigh, arm, and triceps, and subscapular skinfold thicknesses. The other five predictors that were measured by the DXA scan are percent fats of trunk, total body, arms and legs. The percent body fats for arms and legs were averaged scores of right/left arms and legs. Since the correlations were very high for left/right arms (Pearson's $r = 0.99$) and legs (Pearson's $r = 0.98$), it is considered unessential to include both of right/left measures in analysis. Other than the body measures, the models are controlled for age, considering that age is well known to be one of the most significant factors for predicting the health risks tested in this study [23, 24].

The health outcome in this study is whether an individual has at least one obesity-related health risk, such as diabetes, hypertension, and three cardiovascular diseases (i.e. coronary heart disease, congestive heart failures, and heart attack) [2]. If an individual has at least one of these diseases, he is coded as 1, and if he has none of them, he is coded as 0.

2.2 Methodology

I applied Random Forest (RF) classification [17] to the NHANES 2005–2006 data using the *randomForest* package [25] in R 3.2.1. RF is an ensemble method used for classification and regression. Multiple trees are fitted to a bootstrap sample of the original training dataset, and the final prediction is obtained by aggregating the votes of the ensemble. Previous studies have shown that RF significantly improves classification performance over single-tree methods such as CART [19]. The remainder of this paper will assume general familiarity with RF-related terminology; unfamiliar readers should benefit from Cutler et al. [26].

RF was selected for this study for two main reasons: (1) the way it handles multi-collinearity and (2) its established methods of application to imbalanced data. I will discuss these two issues in detail in this section.

2.2.1 Multi-collinearity and Random Forest

Most previous studies that have examined the relevance of body measures in obesity classification have used multiple regression models [3,12,27]. Although this can test whether a body measure has significant explanatory power in the prediction of obesity-related health outcomes, it is limited to using a single measure or measures that are not highly correlated. This is because the inclusion of highly correlated variables in a model can cause misleading results in regression analysis.

For example, BMI and waist circumference are the most frequently tested body measures in obesity classification. Figure 2.1 shows the correlations between various body measures in the NHANES 2005–2006 data, and the Pearson's r between BMI and waist circumference is 0.87. Because the correlation is quite high, including the two measures in one regression model should be avoided. Such a high correlation between variables results in over-inflated standard errors for regression coefficients, which can make the coefficient appear to be insignificant even when it is in fact significant.

Consequently, traditional procedures, such as stepwise or criterion-based variable selection, handle multi-collinearity by keeping a few of the highly correlated variables and

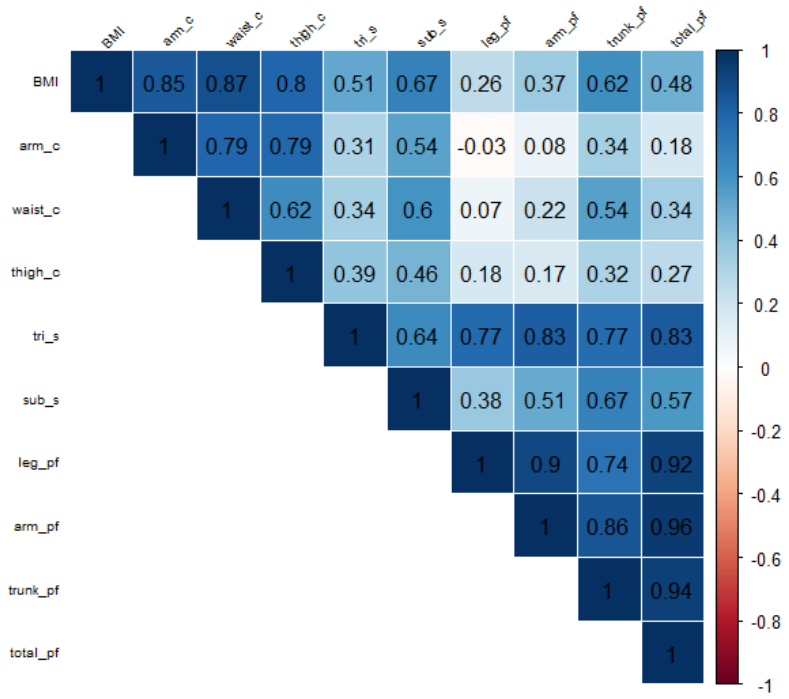


Fig. 2.1: Correlations between body measures in the NHANES 2005–2006 data

discarding the rest of them. Although this procedure helps prevent the over-inflation of standard errors, it makes it hard to compare the relative explanatory powers of those important but highly correlated variables.

On the other hand, RF takes care of multi-collinearity in a unique way that lets it keep all the variables. When deciding the best split at each node of every classification tree, RF randomly selects a subset of the variables instead of considering all of them. This procedure spreads [28] the variable importance across the variables that are highly correlated without discarding any of them.

Furthermore, RF provides the "mean decrease accuracy (MDA)" for all the variables as a measure of their importance in classification. The MDA is the normalized difference in the classification accuracy when the variable is included. It is obtained as follows: (1) As the out-of-bag (OOB) samples pass down the tree, the prediction accuracy is recorded; (2) the accuracy is measured repeatedly with the permutation of the values for variable x; and (3) the final MDA for variable x is obtained by averaging the records of the permutations

across the ensemble of trees [29].

In particular, the variable importance plot provided by the *randomForest* package in R [25] can help this study compare the relative importance of body measures in obesity classification. The plot presents the MDA for each variable in the analysis horizontally and sorts the variables vertically in descending order of MDA. Therefore, a variable located on the upper floor of the plot has the higher MDA and is more important than one on the lower floor. In this study, the importance of body measures in obesity classification is decided by the rank presented in the variable importance plot.

2.2.2 Imbalanced Data and Random Forest

In practical applications of classification algorithms, it is common to see imbalanced data. Data are imbalanced when they contain at least one class that represents a very small proportion of the data [19]. This class imbalance is found particularly often in health research, which usually uses data consisting mostly of the negative class (people who don't have a certain disease), with only a minority in the positive class (people who do).

Although it is not a rare problem, class imbalance needs to be dealt with when one is applying classification algorithms and evaluating a model's performance using predictive accuracy. Because high predictive accuracy can be achieved simply by identifying the negatives that are in the majority, algorithms often fail to correctly detect the minority of positives, resulting in a poor sensitivity (a low percentage of correctly classified positives) [30]. It is usually one of the goals of health research to understand the characteristics of the positive cases, and in this case the model does not achieve this goal.

There are two widely-used approaches to the class imbalance problem [19]. One is based on cost-sensitive learning that assigns a high cost to misclassification of members of the minority class. The other is a sampling technique that forces the classes into balance by under-sampling the majority class or over-sampling the minority class.

To deal with the class imbalance using RF, this study compares three different approaches: a cost-sensitive learning technique, weighted Random Forest (WRF) [19], and two sampling techniques, down-sampling (DS) [20] and the synthetic minority over-sampling

technique (SMOTE) [21]. The classification performances of those three methods are compared, and the best-performing method for each population is selected for the construction of variable importance plots and predicted probability plots.

Weighted Random Forest (WRF)

WRF is a cost-sensitive learning approach that imposes a heavier penalty for misclassification of the minority class. This penalty is taken into account twice in RF. First, during the process of finding the best split to induce the tree, the penalty is imposed on the Gini index. Second, at the terminal nodes of each tree, a weighted majority vote, which is the weight for that class times the number of cases for the class at the terminal node, determines the prediction of the classes. The final prediction of WRF is estimated via an aggregation of the weighted votes of each individual tree.

In this study, four different weights are tested: (5:1), (5:2), (5:3) and (5:4) using the *classwt* option in the *randomForest* function under R 3.2.1. R users should be aware that the *classwt* option in the current version of the *randomForest* package does not implement the complete WRF algorithm since the package reflects the original Fortran code, but not the new Fortran code. The complete WRF algorithm can be implemented using the open source Fortran 77 code [17].

For each weight, four different cut-offs (0.5, 0.6, 0.7, and 0.8) are also tested. Each tree gives a classification for each observation, which counts as a vote for the classification. By aggregating the votes from all the trees, RF decides the winning class as the one having the most votes for the observation. The cut-off, which is 0.5 by default when there are two classes, is used to adjust the votes. For example, suppose an observation has 209 votes for the first class and only 92 votes for the second class. If the default cut-off is used, the winning class is decided by comparing $209/0.5 = 408$ to $92/0.5 = 184$. Since 408 is greater than 194, the winning class is the first class. If a higher cut-off is used, say 0.7, then $209/0.7 = 298.6$ is compared to $92/(1 - 0.7) = 306.67$ and the winning class is the second class.

Down-sampling (DS)

In the Random Forest, when each tree is fitted to a bootstrap sample of the original training dataset, a class imbalance leads to there being only a small number of the minority class in the bootstrap sample, which results in poor predicting performance for the minority class (poor sensitivity). To alleviate this problem, the number of samples drawn from each class in RF can be tuned so that they are equal, which forces the classes to be balanced. Although this can be done by over-sampling the minority class or down-sampling the majority class, recent studies have shown that down-sampling outperforms over-sampling in spite of the possible loss of information [19].

In this study, down-sampling is conducted using the *sampsize* option in the *randomForest* function with four different cut-offs: 0.5, 0.6, 0.7, and 0.8. The numbers of samples drawn from the majority and minority classes are both set to the sample size of the minority class in the original dataset.

Synthetic Minority Over-sampling Technique (SMOTE)

SMOTE is another sampling technique that has more recently been suggested for dealing with class imbalance [21]. The basic idea of SMOTE is to combine over-sampling of the minority class with down-sampling of the majority class. While the down-sampling in SMOTE is done by random exclusion of samples in the majority class, what makes SMOTE unique is the way it over-samples: it creates synthetic samples of the minority class instead of drawing samples from the minority with replacement.

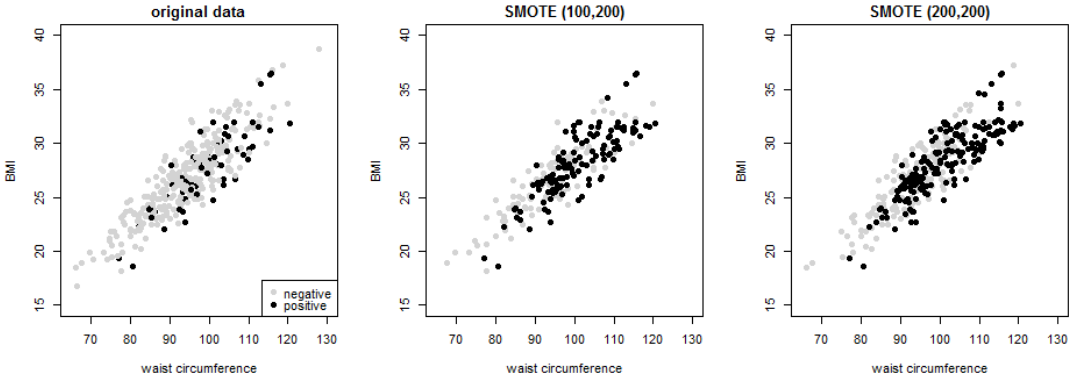
SMOTE generates synthetic samples as follows: For a given sample from a minority class, k neighbors belonging to the minority class are randomly selected. The default value of k is 5, but this can be adjusted. Then, depending on the over-sampling rate, n of k nearest neighbors are randomly selected. For each minority sample and its nearest neighbors, the difference between their feature vectors is then obtained. Finally, by multiplying a random number between 0 and 1 by the difference vector, SMOTE creates the synthetic sample.

The SMOTE procedure is carried out using the *DMwR* package [31] in R 3.2.1. I applied two different SMOTE models: one with an under-sampling rate of 200 and an

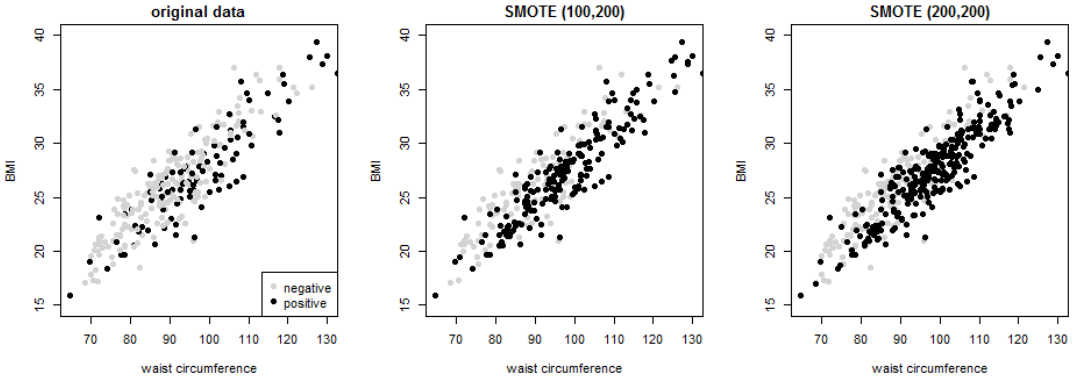
over-sampling rate of 100, the other with both rates at 200. The notation used to describe such models is SMOTE (r_{under}, r_{over}), with r_{under} and r_{over} being the rates of under- and over-sampling, respectively.

Because new synthetic samples are added, the data processed by the SMOTE procedure contains samples that do not exist in the original dataset. As shown in the scatter plots in Figure 2.2, while some samples of the majority class are randomly excluded, new samples of the minority classes are added by the SMOTE procedure. In this study, RF is fitted to the training data set processed by SMOTE, and predictive accuracy measures for comparison with other methods are estimated using the test data set containing only the original samples.

(a) Hispanics



(b) Blacks



(c) Whites

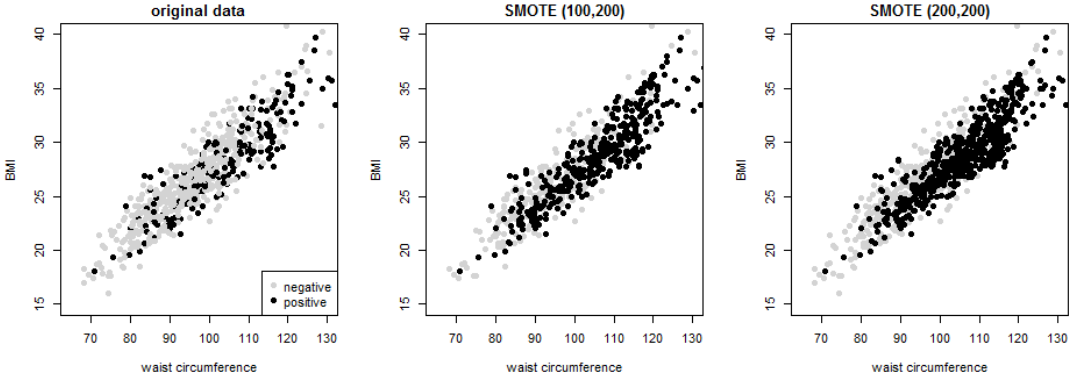


Fig. 2.2: Scatter plots of waist circumference and BMI for original data, SMOTE (100,200), and SMOTE (200,200), by population.

Chapter 3

Results

3.1 Performance Measurement

Table 3.1: Confusion Matrix

	Predicted Positive	Predicted Negative	Total
Actual Positive	True Positive (TP)	False Negative (FN)	P1.
Actual Negative	False Positive (FP)	True Negative (TN)	P2.
Total	P.1	P.2	1

Performance of the RF-based methods to be compared is evaluated using five different measures: percent correctly classified (PCC), sensitivity, specificity, kappa (adjusted PCC for the random chances), and area under the receiver operating characteristic curve (ROC AUC). The formulas for calculating the first four measures are presented in detail below based on elements of the confusion matrix in Table 3.1, and the AUC is obtained by measuring the area under the ROC curves using the *pROC* package in R.

$$PCC = \frac{TP + TN}{TP + FN + FP + TN}$$

$$Specificity = \frac{TP}{TP + FN}$$

$$Sensitivity = \frac{TN}{FP + TN}$$

$$Kappa = \frac{PCC - \text{random accuracy}}{1 - \text{random accuracy}}$$

$$(\text{where random accuracy} = P.1 * P1. + P.2 * P2.)$$

3.2 Performance Comparison

Table 3.2 compares the performances of RF, WRF, RF with DS, and RF with SMOTE.

When applying the RF without any other means of treating the class imbalance, the PCC was 0.82 for Hispanics, 0.71 for Blacks,and 0.73 for Whites. These PCCs are mostly gained by correctly classifying the majority negative. So although the specificities are good, the sensitivities are relatively poor – the poorest in Hispanic males (0.09), and slightly better but still inadequate among Whites (0.34) and Blacks (0.43). To better predict the minority positive class, better sensitivity is required.

Table 3.2: Comparing performances of RF, WRF and RF with DS by cut-off, and RF with SMOTE (estimates of 10-fold cross-validation)

(a) Hispanics (proportion of the minority = 18.0%)

No.	Method	PCC	specificity	sensitivity	kappa	AUC
1	RF, default (cutoff=0.5)	0.82	0.98	0.09	0.10	0.53
2	RF, cutoff=0.6	0.78	0.91	0.20	0.13	0.55
3	RF, cutoff=0.7	0.73	0.81	0.39	0.18	0.60
4	RF, cutoff=0.8	0.66	0.66	0.66	0.22	0.66
5	WRF, cutoff=0.5, weight=5:1	0.81	0.97	0.08	0.07	0.52
6	WRF, cutoff=0.6, weight=5:1	0.77	0.92	0.13	0.05	0.52
7	WRF, cutoff=0.7, weight=5:1	0.74	0.81	0.42	0.21	0.62
8	WRF, cutoff=0.8, weight=5:1	0.63	0.63	0.64	0.18	0.64
9	WRF, cutoff=0.5, weight=5:2	0.81	0.98	0.08	0.08	0.53
10	WRF, cutoff=0.6, weight=5:2	0.78	0.91	0.16	0.08	0.53
11	WRF, cutoff=0.7, weight=5:2	0.74	0.83	0.36	0.18	0.59
12	WRF, cutoff=0.8, weight=5:2	0.67	0.68	0.61	0.20	0.64
13	WRF, cutoff=0.5, weight=5:3	0.81	0.98	0.06	0.06	0.52
14	WRF, cutoff=0.6, weight=5:3	0.78	0.92	0.14	0.07	0.53
15	WRF, cutoff=0.7, weight=5:3	0.74	0.83	0.34	0.16	0.59
16	WRF, cutoff=0.8, weight=5:3	0.65	0.65	0.61	0.18	0.63
17	WRF, cutoff=0.5, weight=5:4	0.81	0.98	0.05	0.04	0.51
18	WRF, cutoff=0.6, weight=5:4	0.77	0.91	0.13	0.05	0.52
19	WRF, cutoff=0.7, weight=5:4	0.74	0.82	0.38	0.18	0.60
20	WRF, cutoff=0.8, weight=5:4	0.66	0.66	0.66	0.22	0.66
21	RF with DS, cutoff=0.5	0.69	0.72	0.58	0.22	0.65
22	RF with DS, cutoff=0.4	0.66	0.63	0.78	0.26	0.71
23	RF with DS, cutoff=0.3	0.56	0.48	0.89	0.20	0.69
24	RF with DS, cutoff=0.2	0.42	0.31	0.92	0.11	0.62
25	RF with SMOTE (100,200)	0.66	0.66	0.64	0.21	0.65
26	RF with SMOTE (200,200)	0.70	0.71	0.64	0.26	0.67
27	RF with SMOTE (250,300)	0.76	0.81	0.53	0.30	0.67

(b) Blacks (proportion of the minority = 33.0%)

No.	method	PCC	specificity	sensitivity	kappa	AUC
1	RF, default (cu-toff=0.5)	0.71	0.85	0.43	0.31	0.64
2	RF, cut-off=0.6	0.73	0.77	0.65	0.41	0.71
3	RF, cut-off=0.7	0.69	0.66	0.75	0.38	0.71
4	RF, cut-off=0.8	0.57	0.44	0.82	0.21	0.63
5	WRF, cut-off=0.5, weight=5:1	0.73	0.86	0.47	0.35	0.67
6	WRF, cut-off=0.6, weight=5:1	0.75	0.78	0.68	0.45	0.73
7	WRF, cut-off=0.7, weight=5:1	0.69	0.66	0.74	0.36	0.70
8	WRF, cut-off=0.8, weight=5:1	0.59	0.46	0.85	0.25	0.65
9	WRF, cut-off=0.5, weight=5:2	0.71	0.85	0.45	0.32	0.65
10	WRF, cut-off=0.6, weight=5:2	0.74	0.78	0.64	0.42	0.71
11	WRF, cut-off=0.7, weight=5:2	0.69	0.66	0.74	0.37	0.70
12	WRF, cut-off=0.8, weight=5:2	0.58	0.44	0.86	0.24	0.65
13	WRF, cut-off=0.5, weight=5:3	0.70	0.86	0.39	0.28	0.63
14	WRF, cut-off=0.6, weight=5:3	0.74	0.78	0.65	0.42	0.71
15	WRF, cut-off=0.7, weight=5:3	0.70	0.66	0.76	0.39	0.71
16	WRF, cut-off=0.8, weight=5:3	0.58	0.45	0.84	0.24	0.64
17	WRF, cut-off=0.5, weight=5:4	0.71	0.86	0.41	0.30	0.64
18	WRF, cut-off=0.6, weight=5:4	0.70	0.75	0.61	0.35	0.68
19	WRF, cut-off=0.7, weight=5:4	0.70	0.67	0.75	0.39	0.71
20	WRF, cut-off=0.8, weight=5:4	0.58	0.46	0.81	0.22	0.64
21	RF with DS, cut-off=0.5	0.70	0.74	0.62	0.35	0.68
22	RF with DS, cut-off=0.4	0.68	0.64	0.76	0.36	0.70
23	RF with DS, cut-off=0.3	0.60	0.47	0.84	0.25	0.65
24	RF with DS, cut-off=0.2	0.50	0.27	0.94	0.16	0.61
25	RF with SMOTE (100,200)	0.71	0.69	0.75	0.41	0.72
26	RF with SMOTE (200,200)	0.69	0.72	0.63	0.34	0.67
27	RF with SMOTE (250,300)	0.72	0.79	0.59	0.38	0.69

(c) Whites (proportion of the minority = 29.4%)

No	method	PCC	specificity	sensitivity	kappa	AUC
1	RF, default (cut-off=0.5)	0.73	0.89	0.34	0.26	0.62
2	RF, cut-off=0.6	0.70	0.79	0.48	0.27	0.64
3	RF, cut-off=0.7	0.64	0.66	0.58	0.22	0.62
4	RF, cut-off=0.8	0.57	0.46	0.81	0.21	0.64
5	WRF, cut-off=0.5, weight=5:1	0.72	0.88	0.34	0.24	0.61
6	WRF, cut-off=0.6, weight=5:1	0.70	0.79	0.49	0.28	0.64
7	WRF, cut-off=0.7, weight=5:1	0.65	0.67	0.61	0.24	0.64
8	WRF, cut-off=0.8, weight=5:1	0.55	0.44	0.81	0.19	0.62
9	WRF, cut-off=0.5, weight=5:2	0.73	0.90	0.34	0.27	0.62
10	WRF, cut-off=0.6, weight=5:2	0.71	0.81	0.48	0.30	0.64
11	WRF, cut-off=0.7, weight=5:2	0.65	0.67	0.63	0.26	0.65
12	WRF, cut-off=0.8, weight=5:2	0.55	0.44	0.80	0.19	0.62
13	WRF, cut-off=0.5, weight=5:3	0.73	0.89	0.34	0.27	0.62
14	WRF, cut-off=0.6, weight=5:3	0.71	0.81	0.47	0.29	0.64
15	WRF, cut-off=0.7, weight=5:3	0.66	0.68	0.60	0.25	0.64
16	WRF, cut-off=0.8, weight=5:3	0.57	0.48	0.80	0.21	0.64
17	WRF, cut-off=0.5, weight=5:4	0.73	0.90	0.31	0.24	0.61
18	WRF, cut-off=0.6, weight=5:4	0.70	0.80	0.44	0.25	0.62
19	WRF, cut-off=0.7, weight=5:4	0.66	0.68	0.60	0.25	0.64
20	WRF, cut-off=0.8, weight=5:4	0.58	0.49	0.80	0.23	0.65
21	RF with DS, cut-off=0.5	0.70	0.76	0.55	0.30	0.65
22	RF with DS, cut-off=0.4	0.64	0.61	0.70	0.26	0.66
23	RF with DS, cut-off=0.3	0.55	0.44	0.83	0.20	0.63
24	RF with DS, cut-off=0.2	0.42	0.21	0.92	0.08	0.56
25	RF with SMOTE (100,200)	0.64	0.64	0.64	0.24	0.64
26	RF with SMOTE (200,200)	0.65	0.69	0.57	0.24	0.63
27	RF with SMOTE (250,300)	0.67	0.76	0.46	0.22	0.61

One approach to improve sensitivity is to adjust the cut-off. As the cut-off changes, sensitivity is gained, but at the expense of specificity. Figure 3.1 displays the trade-off in detail. Although the higher values of the cut-off yield greater sensitivity, specificity and PCC deteriorate. This is also shown in the results in Table 3.2.

(a) Hispanics

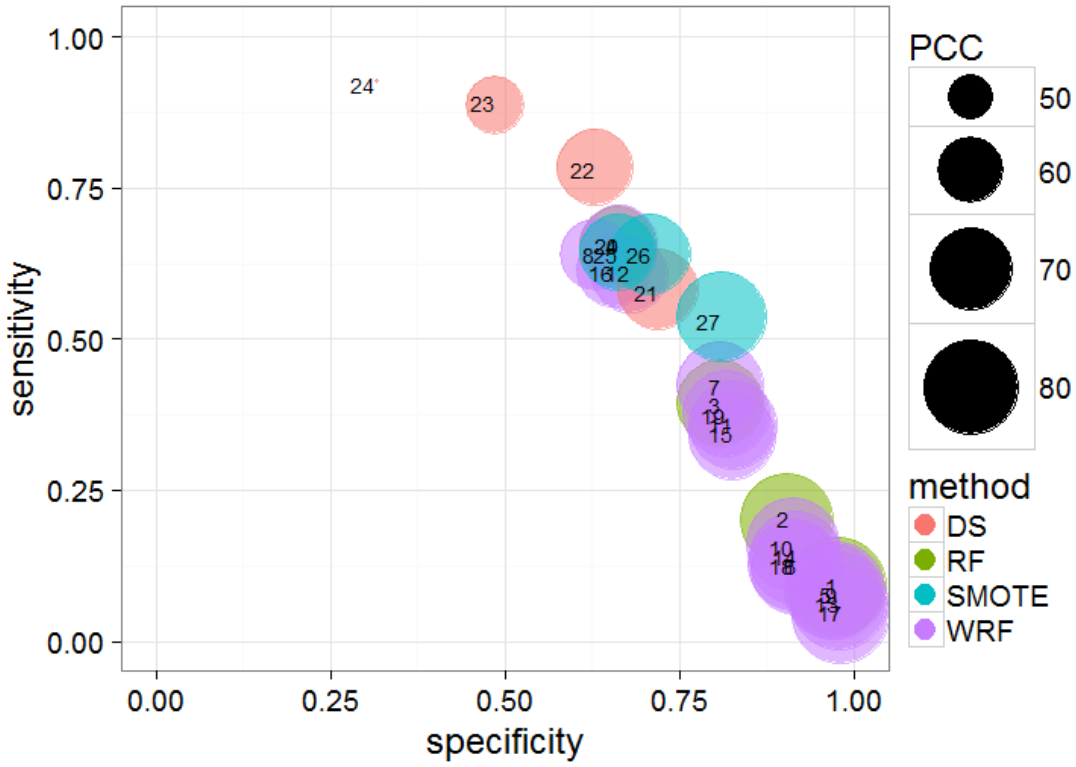


Fig. 3.1: Visualization of model performance results from Table 3.2. Bubble sizes are proportional to PCC, and bubble identifier numbers correspond to model numbers in Table 3.2.

(b) Blacks

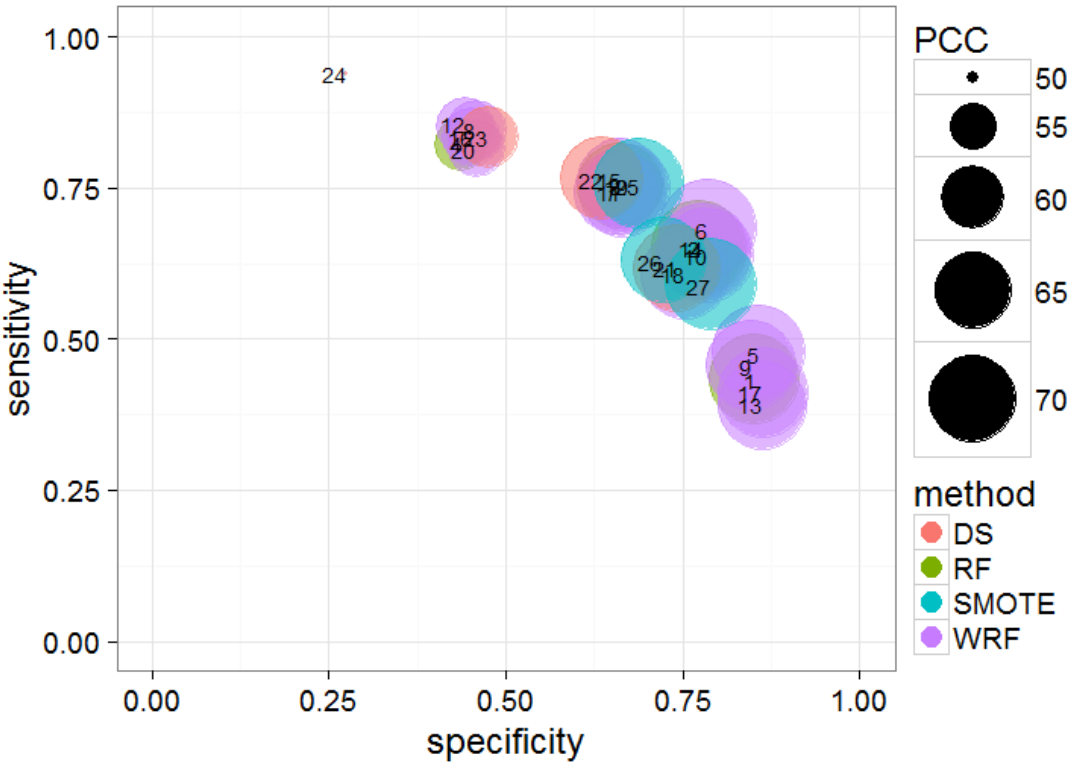


Fig. 3.1: (continued) Visualization of model performance results from Table 3.2.

(c) Whites

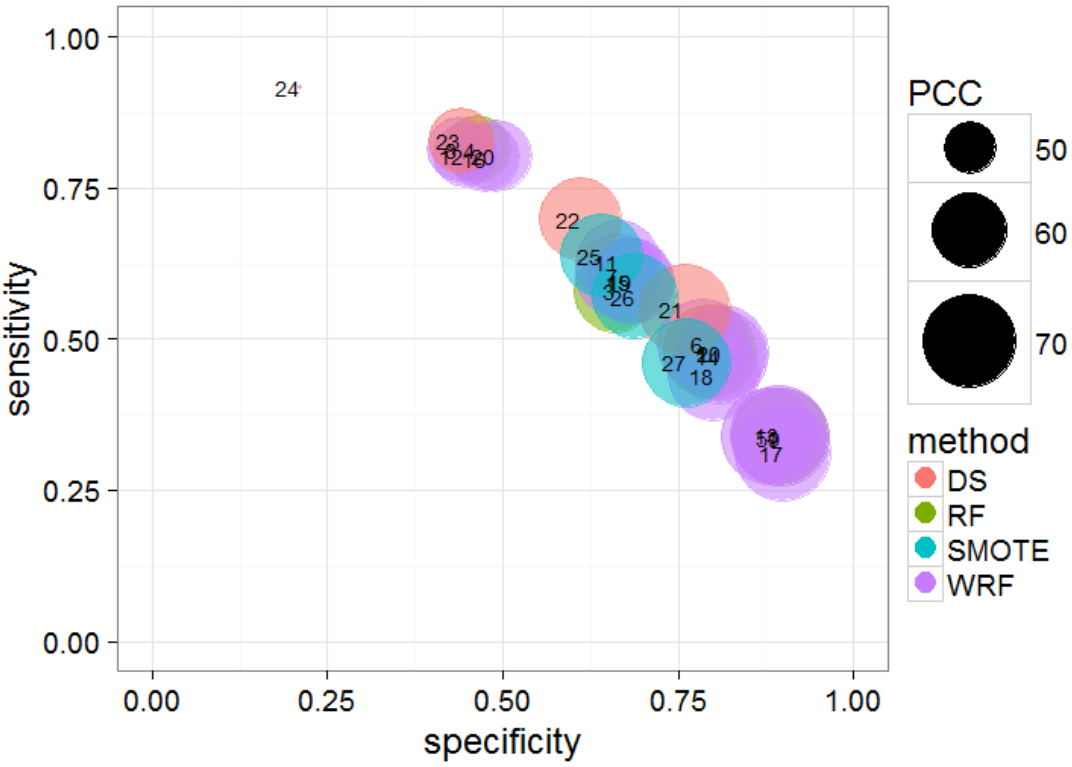


Fig. 3.1: (continued) Visualization of model performance results from Table 3.2.

For WRF, the four different weights appear to make negligible differences on the sensitivity. When holding the cut-off constant, the differences between the sensitivities for different weights turn out to be smaller than 0.05 for Hispanics and Blacks and smaller than 0.10 for Whites. This may be because WRF does not help balance the data in this study, or it may be due to the fact that the *classwt* option in the current version of the *randomForest* package in R does not implement WRF completely. As noted earlier, the *classwt* option in the current version of the *randomForest* package should be used with caution since it does not reflect the complete WRF algorithm. For the complete implementation of WRF, it is recommended to use the open source Fortran 77 code [17].

On the other hand, RF with DS works better for improving the sensitivity compared to WRF using the same cut-off. However, because improved sensitivity is acquired at the expense of specificity, the estimated specificity and PCC are worse than those estimated by WRF. A similar trend is seen in the cases of SMOTE. While it works well for improving sensitivity, PCC and specificity drop consequently.

There is no universal winner among the methods considered, since the best performing one turned out to vary by population. Based on Figure 3.1 and Table 3.2, RF with SMOTE (250, 300) for Hispanics, WRF with cut-off=0.6 and weight=5:1 for Blacks, and RF with DS with cut-off=0.5 for Whites are used as final models to test the importance of variables and to obtain predicted probability of having obesity-related diseases.

3.3 Application of Selected Final Model

Figure 3.2 presents the variable importance plots for Hispanic, Black, and White males in the U.S., constructed by RF using the final models for the classification of persons who have at least one obesity-related disease. For all the three populations, the most important variable is age. The most important body measure turned out to be waist circumference for Blacks and Hispanics, which is also the third important body measure for Whites. The most important body measure for Whites is BMI.

Other variables in the upper rank vary by population. Interestingly, BMI, which is the most important body measure for Whites, turns out to play a less critical role in the classification of Hispanic and Black males. In the case of Hispanics, percent fats of arm and leg are located in the upper ranks of the plot. In the case of Blacks, although there are not great differences in mean decrease accuracy between the body measures, arm circumference and the percent fats of trunk are among the top ranked. For Whites, thigh circumference is the second most important body measure.

To better illustrate how informative the BMI and its current cut-offs are for predicting obesity-related disease, a set of predicted probability plots has been constructed (Figure 3.3).

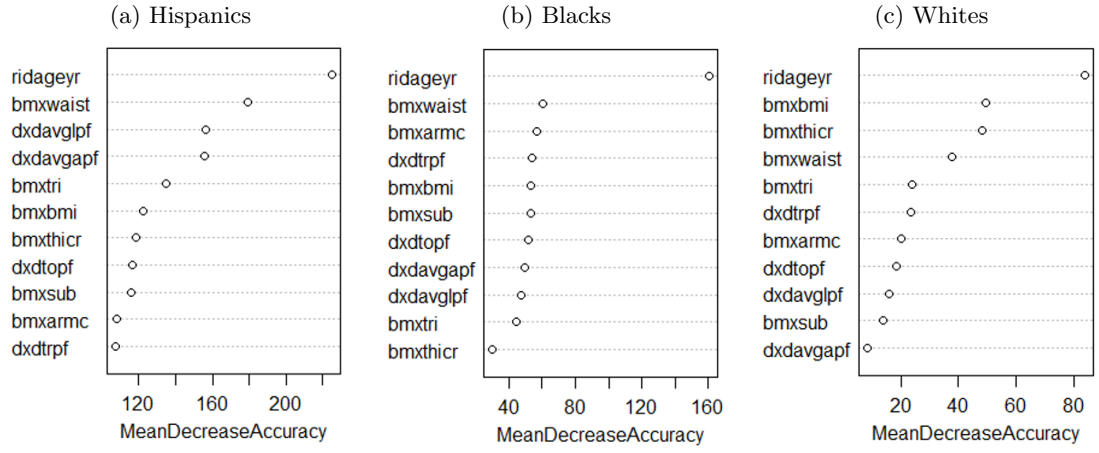


Fig. 3.2: Variable importance plot

(ridageyr: age; bmxarmc: arm circumference; bmxbmi: BMI; bmxsub: subscapular skinfold thickness; bmxthicr: thigh circumference; bmxtri: tricep skinfold thickness; bmxwaist: waist circumference; dxdavgapf: average percent fat of arms; dxdavglpf: average percent fat of legs; dxdtopf: percent fat of total body; dxdtrpf: percent fat of trunk)

The plots display changes in the predicted probability of having at least one obesity-related disease (z-axis) in relation to BMI (y-axis) and waist circumference (x-axis) when other body measures are held at the grand means of the three populations.

The plot for White males (Figure 3.3–(a)) shows cut-offs of both BMI and waist circumference at which the probability sharply increases. The probability jumps at the BMI between 25 and 30, and at the waist circumference between 100 and 110. If an individual has either high BMI or high waist circumference above those cut-offs, the probability of having at least one obesity-related disease rises. Especially if he has waist circumference greater than 110, the probability reaches 0.5 regardless of BMI level.

For Black males (Figure 3.3–(b)), the probability moderately increases from the BMI between 25 and 30, and jumps at the waist circumference around 100. When waist circumference is greater than 105, the probability exceeds 0.5. For White and Black males, these two body measures also interact. If an individual has high BMI *and* high waist circumference, the risk becomes the highest.

One the other hand, BMI lacks comparable explanatory power when it comes to Hispanic males (Figure 3.3–(c)). The predicted probability surface shows no substantial change of level across BMI, suggesting that BMI is a poor predictor of obesity-related disease in Hispanic males. Waist circumference better captures the probability changes. For Hispanic males, the probability gradually increases from the waist circumference of 90, reaching 0.5 at the waist circumference of 105.

(a) Whites

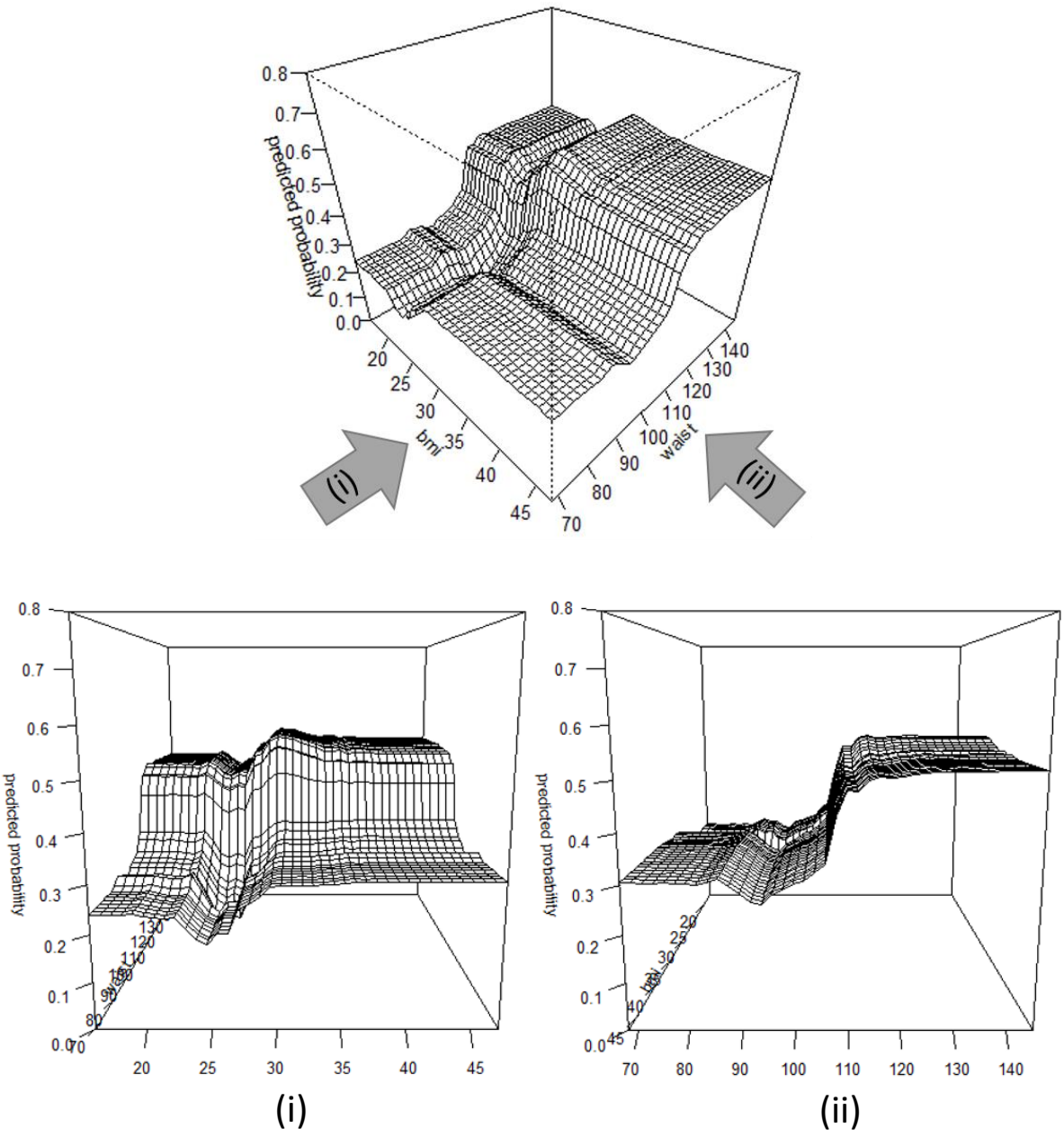


Fig. 3.3: Predicted probability plot by population

(b) Blacks

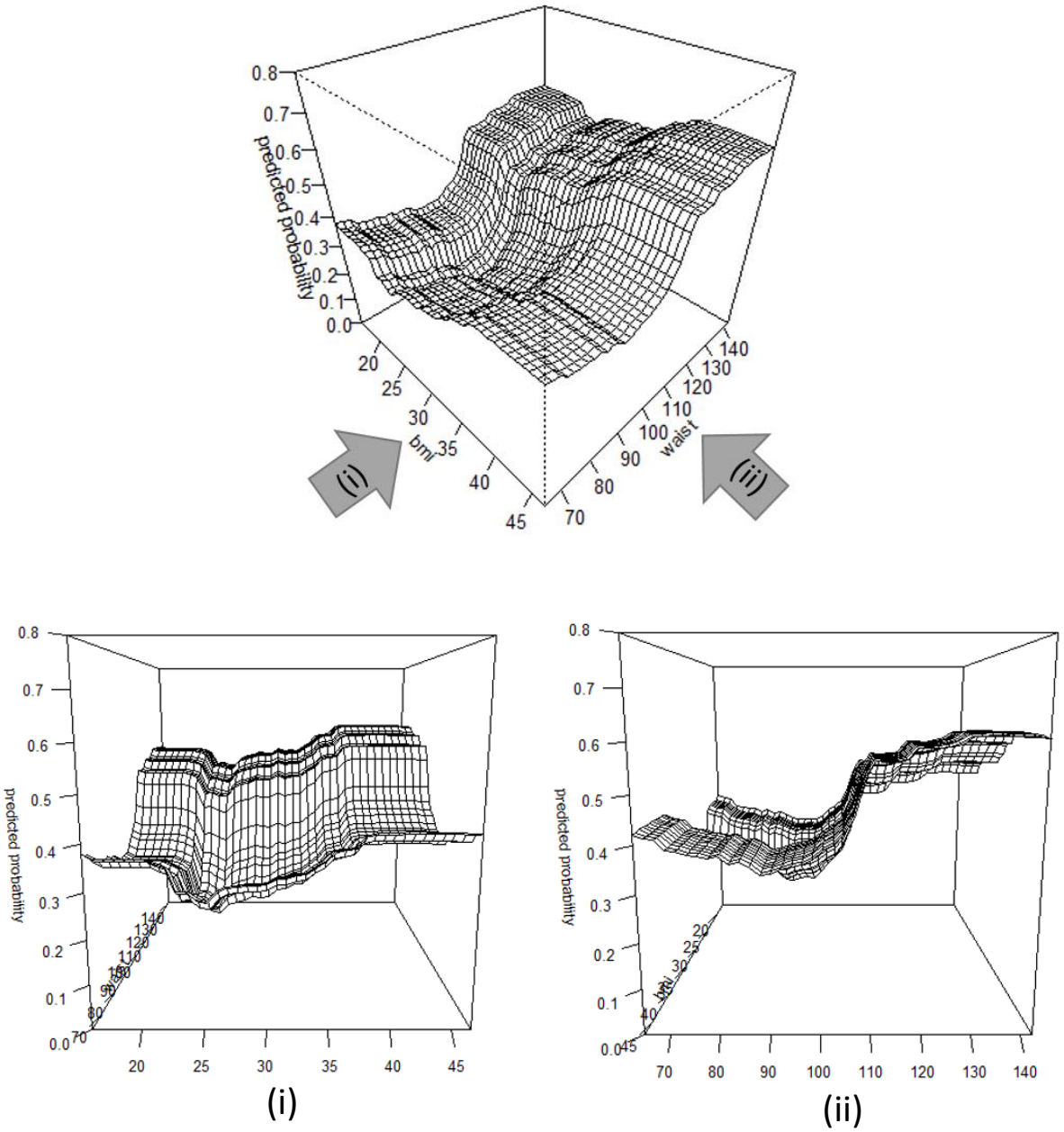


Fig. 3.3: (continued) Predicted probability plot by population

(c) Hispanics

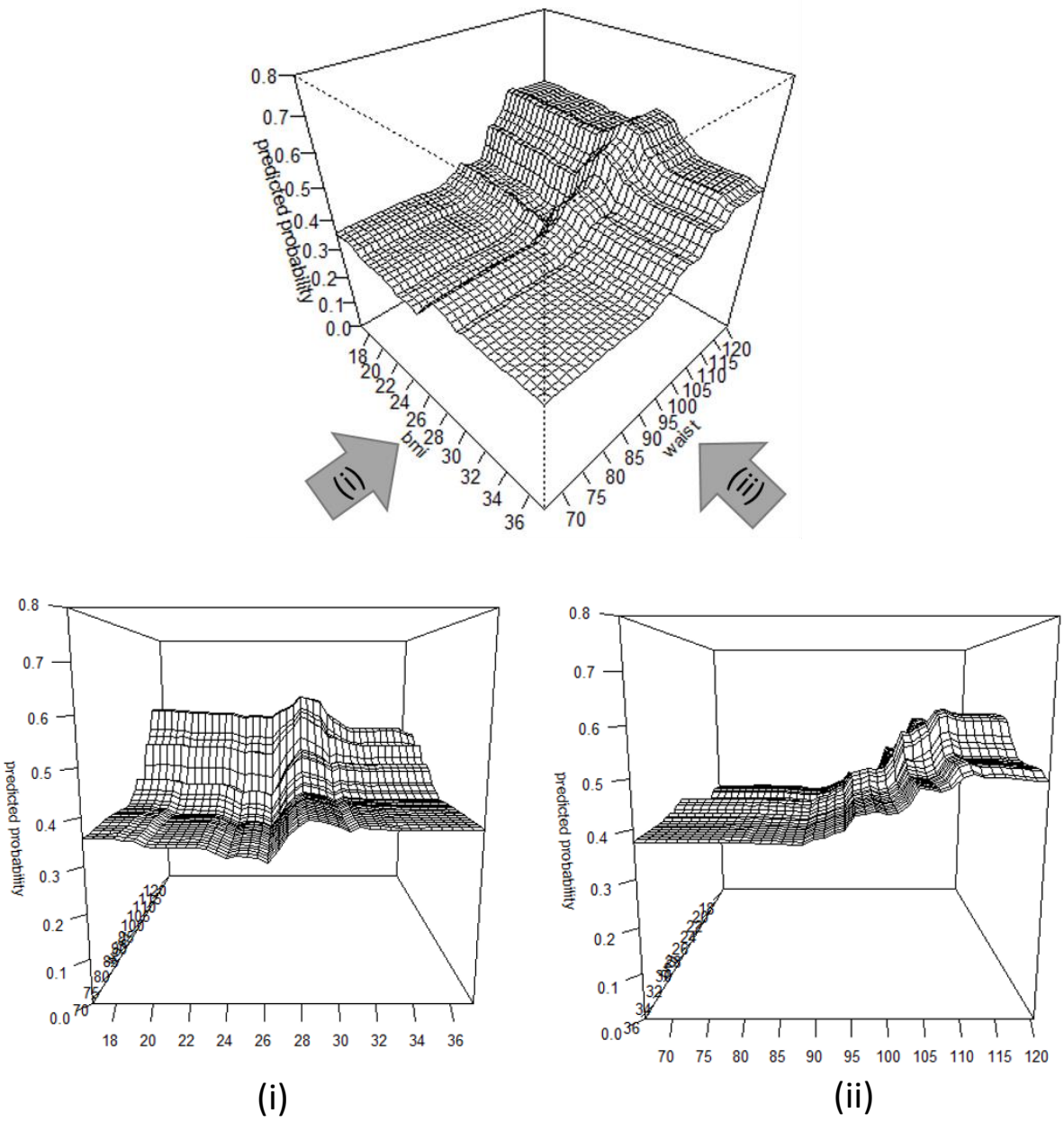


Fig. 3.3: (continued) Predicted probability plot by population

Chapter 4

Discussion and Future Studies

This paper compares three different Random Forest-based methods of handling class imbalance. While WRF imposes a heavier penalty for the misclassification of the minority class, DS and SMOTE use sampling techniques to force the classes to be balanced before inducing the trees. In this study, the best performing method turned out to vary by population. With this in mind, it is recommended for future studies to compare the performances of different balancing methods and choose the best performing method for the given data.

The two main obesity-related conclusions of this study confirm the stress laid by previous studies on the importance of developing racial- and ethnic-specific classification of obesity.

First, developing the body measures that work best for a certain racial or ethnic population will increase the accuracy of obesity classification of the entire diverse population insofar as these measures accurately reflect facts of body composition and fat distribution characteristics of the sub-population. According to the variable important plots (Figure 3.2), BMI turned out to be less informative, but waist circumference more informative, in classifying obesity in Hispanic and Black males in the United States.

The study also suggests that using several of the best indicators in conjunction provides a more accurate classification of obesity than using a single one alone. For example, both BMI and waist circumference proved to be important indicators of obesity for White males, and the risk of having obesity-related disease is affected by both of them. White males with the same BMI can have very different health risks. An individual with a BMI of 30 but a waist circumference below 100 has a probability around 0.3 of suffering an obesity-related disease, whereas another with the same BMI of 30 but a waist circumference above 115 has a markedly higher probability of doing so, around 0.6.

Second, this study supports the argument that risk factor cut-off should be population-specific as well. Although waist circumference turns out to be an important body measure among all the three population, which population is being studied affects the point at which probability sharply increases. For whites, the probability reaches 0.5 when waist circumference is higher than 110, while the probability exceeds 0.5 at waist circumference around 105 for Hispanic and Black males.

This study underscores the importance of considering predicted probability when setting the racial- or ethnic-specific cut-offs of indicators. If the cut-off is simply set at the point at which risk steeply increases, this lack of precision increases the chances of either overestimating or underestimating the risk of obesity. For example, although the probability of having obesity-related disease begins to increase at a waist circumference of 90 among White males, that probability remains quite low: around 0.3. It would be erroneous to classify persons with a relatively low probability of suffering an obesity-related disease as high-risk.

Future studies can build upon the results of the present investigation by applying a comparable range of body measures, with appropriate cut-offs, to other racial and ethnic populations; females and adolescents in various populations may also be studied. Furthermore, to obtain better prediction for obesity-related disease, it would be valuable to account for known risk behaviors associated with obesity-related diseases such as smoking status and alcohol intake along with the various body measures. As obesity emerges as one of the world's most important health problems, tailoring methods of obesity classification to the specific risk-affecting characteristics of specific population groups can help health professionals and policy makers more accurately assess who is at risk and to apply more effective treatment and policies.

References

[1] P. G. Kopelman, "Obesity as a medical problem." *Nature*, vol. 404, no. 6778, pp. 635–643, 2000.

[2] Panel, NHLBI Obesity Education Initiative Expert and others, "Clinical guidelines on the identification, evaluation, and treatment of overweight and obesity in adults." 1998 [Online]. Available: `http://www.ncbi.nlm.nih.gov/books/NBK2003/`.

[3] E. Evans, D. Rowe, S. Racette, K. Ross, and E. McAuley, "Is the current BMI obesity classification appropriate for black and white postmenopausal women?" *International Journal of Obesity*, vol. 30, no. 5, pp. 837–843, 2006.

[4] S. M. Garn, W. R. Leonard, and V. M. Hawthorne, "Three limitations of the body mass index." *The American Journal of Clinical Nutrition*, vol. 44, no. 6, pp. 996–997, 1986.

[5] M. Deurenberg-Yap, G. Schmidt, W. Van Staveren, and P. Deurenberg, "The paradox of low body mass index and high body fat percentage among Chinese, Malays and Indians in Singapore." *International Journal of Obesity*, vol. 24, pp. 1011–1017, 2000.

[6] V. Wickramasinghe, G. Cleghorn, K. Edmiston, A. Murphy, R. Abbott, and P. Davies, "Validity of BMI as a measure of obesity in Australian white Caucasian and Australian Sri Lankan children." *Annals of Human Biology*, vol. 32, no. 1, pp. 60–71, 2005.

[7] M. Micozzi and D. Albanes, "Three limitations of the body mass index." *The American Journal of Clinical Nutrition*, vol. 46, no. 2, pp. 376–377, 1987.

[8] I. Okosun, S. Tedders, S. Choi, G. Dever, *et al.*, "Abdominal adiposity values associated with established body mass indexes in White, Black and Hispanic Americans. A study from the Third National Health and Nutrition Examination Survey." *International Journal of Obesity*, vol. 24, no. 10, pp. 1279–1285, 2000.

[9] F. Xavier Pi-Sunyer, "Obesity: criteria and classification." *Proceedings of the Nutrition Society*, vol. 59, no. 04, pp. 505–509, 2000.

[10] A. Luke, R. Durazo-Arvizu, C. Rotimi, T. E. Prewitt, T. Forrester, R. Wilks, O. J. Ogunbiyi, D. A. Schoeller, D. McGee, and R. S. Cooper, "Relation between body mass index and body fat in black population samples from nigeria, jamaica, and the united states." *American Journal of Epidemiology*, vol. 145, no. 7, pp. 620–628, 1997.

[11] J. C. Lovejoy, A. Jacques, M. Klemperer, and R. Tulley, "Abdominal fat distribution and metabolic risk factors: effects of race." *Metabolism*, vol. 45, no. 9, pp. 1119–1124, 1996.

[12] I. Janssen, P. T. Katzmarzyk, and R. Ross, "Waist circumference and not body mass index explains obesity-related health risk." *The American Journal of Clinical Nutrition*, vol. 79, no. 3, pp. 379–384, 2004.

[13] WHO, Expert Consultation, "Appropriate body mass index for asian populations and its implications for policy and intervention strategies." *Lancet*, vol. 363, no. 9403, p. 157, 2004.

[14] P. Deurenberg, M. Yap, W. A. Van Staveren, *et al.*, "Body mass index and percent body fat: a meta analysis among different ethnic groups." *International Journal of Obesity*, vol. 22, pp. 1164–1171, 1998.

[15] C. N. Mascie-Taylor and R. Goto, "Human variation and body mass index: a review of the universality of BMI cut-offs, gender and urban-rural differences, and secular changes." *Journal of Physiological Anthropology*, vol. 26, no. 2, pp. 109–112, 2007.

[16] M. Ferland, J.-p. DesprÉS, A. Tremblay, S. Pinault, A. Nadeau, S. Moorjani, P. J. Lupien, G. ThÉriault, and C. Bouchard, "Assessment of adipose tissue distribution by computed axial tomography in obese women: association with body density and anthropometric measurements." *British Journal of Nutrition*, vol. 61, no. 02, pp. 139–148, 1989.

[17] L. Breiman, "Random forests." *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[18] R. Genuer, J.-M. Poggi, and C. Tuleau-Malot, "Variable selection using random forests." *Pattern Recognition Letters*, vol. 31, no. 14, pp. 2225–2236, 2010.

[19] C. Chen, A. Liaw, and L. Breiman, "Using random forest to learn imbalanced data." *University of California, Berkeley*, 2004.

[20] X.-Y. Liu, J. Wu, and Z.-H. Zhou, "Exploratory undersampling for class-imbalance learning." *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. 39, no. 2, pp. 539–550, 2009.

[21] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique." *Journal of Artificial Intelligence Research*, pp. 321–357, 2002.

[22] National Center for Health Statistics, "National Health and Nutrition Examination Survey Data." 2005-2006 [Online]. Available: `http://wwwn.cdc.gov/nchs/nhanes/search/nhanes05_06.aspx`.

[23] H. King, R. E. Aubert, and W. H. Herman, "Global burden of diabetes, 1995–2025: prevalence, numerical estimates, and projections," *Diabetes care*, vol. 21, no. 9, pp. 1414–1431, 1998.

[24] D. of Noncommunicable Diseases Prevention and H. Promotion, "Life course perspectives on coronary heart disease, stroke and diabetes," *Summary Report of a Meeting of Experts*, 2001.

[25] A. Liaw and M. Wiener, "Classification and regression by randomforest." *R News*, vol. 2, no. 3, pp. 18–22, 2002 [Online]. Available: `http://CRAN.R-project.org/doc/Rnews/`.

[26] A. Cutler, D. Cutler, and J. Stevens, "Random forests." in *Ensemble Machine Learning*, C. Zhang and Y. Ma, Eds. Springer US, 2012, pp. 157–175 [Online]. Available: http://dx.doi.org/10.1007/978-1-4419-9326-7_5.

[27] A. Must, G. E. Dallal, and W. H. Dietz, "Reference data for obesity: 85th and 95th percentiles of body mass index (wt/ht2) and triceps skinfold thickness." *The American Journal of Clinical Nutrition*, vol. 53, no. 4, pp. 839–846, 1991.

[28] D. R. Cutler, T. C. Edwards Jr, K. H. Beard, A. Cutler, K. T. Hess, J. Gibson, and J. J. Lawler, "Random forests for classification in ecology." *Ecology*, vol. 88, no. 11, pp. 2783–2792, 2007.

[29] M. L. Calle and V. Urrea, "Letter to the editor: Stability of random forest importance measures." *Briefings in Bioinformatics*, vol. 12, no. 1, pp. 86–89, 2011.

[30] N. V. Chawla, N. Japkowicz, and A. Kotcz, "Editorial: special issue on learning from imbalanced data sets." *ACM Sigkdd Explorations Newsletter*, vol. 6, no. 1, pp. 1–6, 2004.

[31] L. Torgo, *Data Mining with R, learning with case studies.* Chapman and Hall/CRC, 2010 [Online]. Available: http://www.dcc.fc.up.pt/~ltorgo/DataMiningWithR.

Appendices

Appendix A

R Codes

```
set.seed(123)
#install.packages("randomForest")
#install.packages("verification")
#install.packages("corrplot")
#install.packages("DMwR")
#install.packages("shape")
#install.packages("plyr")
#install.packages("pROC")
#install.packages("ROCR")
library(randomForest)
library(verification)
library(corrplot)
library(DMwR)
library(shape)
library(plyr)
library(pROC)
library(ROCR)



data = read.csv("data0915.csv")
data$dxdavgapf = (data$dxdlapf+data$dxdrapf)/2
data$dxdavglpf = (data$dxdllpf+data$dxdrlpf)/2


# =================
# OUTCOME VARIABLES
# =================


# remove observations with missing ridageyr
total = data[!is.na(data$ridageyr), ]
```

```
rm(data)
# remove observations with ridageyr <= 17
total = total[total$ridageyr>17, ]

# remove observations with emiss == 0
total = total[total$emiss==0, ]

# the following variables must be complete for the analysis to work:
total1 = total[,c("bmxbmi","bmxarmc", "bmxwaist", "bmxthicr", "bmxtri", "bmxsub"
    ,
                  "dxdavglpf", "dxdavgapf", "dxdtrpf","dxdtopf", "diq010", "
                      bpq020",
                  "mcq160b", "mcq160c", "mcq160e", "ridreth1", "riagendr") ]
sum(is.na(total1)) # just to check
rm(total1)
#--Total :: At least one of Diabetes, hypertension or heart diseases
total$total = 0
total$total[total$diq010==1 | total$diq010==3 | total$bpq020==1 | total$mcq160c
    ==1 |
              total$mcq160b==1 | total$mcq160e==1] = 1

# ============================
# Fig 2.1: Correlation Heatmap
# ============================
total1 = total[,c("bmxbmi","bmxarmc", "bmxwaist",
                  "bmxthicr", "bmxtri", "bmxsub", "dxdavglpf", "dxdavgapf",
                  "dxdtrpf","dxdtopf") ]
# names are for the corrplot only
colnames(total1) = c("BMI", "arm_c", "waist_c", "thigh_c", "tri_s", "sub_s",
                     "leg_pf", "arm_pf", "trunk_pf", "total_pf")
cor = cor(total1)
corrplot(cor, order="original", method = "shade",tl.pos="lt", type="upper",
         tl.col="black", tl.cex=0.6, tl.srt=45,
         addCoef.col="black"
)
```

```
rm(total1)

# ============================
# RACE/ETHNICITY STRATIFICATION
# ============================
total.hispm = total[total$ridreth1 <= 2 & total$riagendr == 1, ]
total.hispm$total = as.factor(total.hispm$total)

total.whitem = total[total$ridreth1 == 3 & total$riagendr == 1, ]
total.whitem$total = as.factor(total.whitem$total)

total.blackm = total[total$ridreth1 == 4 & total$riagendr == 1, ]
total.blackm$total = as.factor(total.blackm$total)

#%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
kappa = function(x){
  n = sum(x)
  pobs = (x[1,1]+x[2,2])/n
  pexp = (sum(x[1,])*sum(x[,1])+sum(x[2,])*sum(x[,2]))/n^2
  kappa = (pobs-pexp)/(1-pexp)
  return(kappa)
}

class.sum = function(truth, predicted, co){
  predicted[predicted >= (1-co)] = 1
  predicted[predicted < (1-co)] = 0
  auc = auc(truth, predicted)
  xt = table(truth, predicted)
  if(nrow(xt) != 2 | ncol(xt) != 2) stop("empty␣cells")
  pcc = 100*sum(diag(xt))/sum(xt)
  spec = xt[1,1]/sum(xt[1,])
  sens = xt[2,2]/sum(xt[2,])
  kap = kappa(xt)
  list(pcc, spec, sens, kap, auc)
}
```

```
##############################
# Fig 2.2: SMOTE Scatter Plot
##############################


total.hispm.s = SMOTE(total ~ bmxbmi  + bmxarmc + bmxwaist
                      + bmxthicr + bmxtri + bmxsub +dxdavglpf + dxdavgapf +
                         dxdtrpf + dxdtopf,
                      data=total.hispm, perc.over =100, perc.under=200)
set.seed(1234)
total.hispm.s2 = SMOTE(total ~  bmxbmi  + bmxarmc + bmxwaist
                       + bmxthicr + bmxtri + bmxsub +dxdavglpf + dxdavgapf +
                          dxdtrpf + dxdtopf,
                       data=total.hispm, perc.over =200, perc.under=200)
set.seed(1234)
total.hispm.s3 = SMOTE(total ~ bmxbmi  + bmxarmc + bmxwaist
                       + bmxthicr + bmxtri + bmxsub +dxdavglpf + dxdavgapf +
                          dxdtrpf + dxdtopf,
                       data=total.hispm, perc.over =250, perc.under=300)


set.seed(1234)
total.whitem.s = SMOTE(total ~  bmxbmi  + bmxarmc + bmxwaist
                       + bmxthicr + bmxtri + bmxsub +dxdavglpf + dxdavgapf +
                          dxdtrpf + dxdtopf,
                       data=total.whitem, perc.over =100, perc.under=200)
set.seed(1234)
total.whitem.s2 = SMOTE(total ~ bmxbmi  + bmxarmc + bmxwaist
                        + bmxthicr + bmxtri + bmxsub +dxdavglpf + dxdavgapf +
                           dxdtrpf + dxdtopf,
                        data=total.whitem, perc.over =200, perc.under=200)


set.seed(1234)
total.whitem.s3 = SMOTE(total ~ bmxbmi  + bmxarmc + bmxwaist
```

```
                        + bmxthicr + bmxtri + bmxsub +dxdavglpf + dxdavgapf +
                            dxdtrpf + dxdtopf,
                        data=total.whitem, perc.over =250, perc.under=300)

set.seed(1234)
total.blackm.s = SMOTE(total ~ bmxbmi  + bmxarmc + bmxwaist
                       + bmxthicr + bmxtri + bmxsub +dxdavglpf + dxdavgapf +
                           dxdtrpf + dxdtopf,
                       data=total.blackm, perc.over =100, perc.under=200)
set.seed(1234)
total.blackm.s2 = SMOTE(total ~  bmxbmi  + bmxarmc + bmxwaist
                        + bmxthicr + bmxtri + bmxsub +dxdavglpf + dxdavgapf +
                            dxdtrpf + dxdtopf,
                        data=total.blackm, perc.over =200, perc.under=200)
set.seed(1234)
total.blackm.s3 = SMOTE(total ~ bmxbmi  + bmxarmc + bmxwaist
                        + bmxthicr + bmxtri + bmxsub +dxdavglpf + dxdavgapf +
                            dxdtrpf + dxdtopf,
                        data=total.blackm, perc.over =250, perc.under=300)

#%%%%%%%%%MODIFIABLE VAVIRALBES%%%%%%%%%%%%%
data=list()
##HISP
#data[[1]] = total.hispm #---------RACE
#data[[2]] = total.hispm.s
#data[[3]] = total.hispm.s2
##WHITE
#data[[1]] = total.whitem
#data[[2]] = total.whitem.s
#data[[3]] = total.whitem.s2
##BLACK
data[[1]] = total.blackm
data[[2]] = total.blackm.s
data[[3]] = total.blackm.s2
#%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
```

```
bcols=NULL
ll = lapply(seq_along(data),function(x){
  data[[x]]$total.n = as.numeric(data[[x]]$total)
  bcols = cut(data[[x]]$total.n, 2, labels=c("lightgrey", "black"))
  plot(data[[x]]$bmxwaist, data[[x]]$bmxbmi, col=as.character(bcols), pch=16)
} )



# ============================
# TABLE 3.2
# ============================

rf.10cv = function(data, co=0.5, classwt=weight, samp=FALSE, smote=FALSE,
                   perc.over, perc.under){
  if(missing(classwt)) weight = table(data$total)
  probs = rep(0, nrow(data))
  set.seed(1234)
  xvs = rep(1:10, length=nrow(data))
  xvs = sample(xvs)
  for(i in 1:10) {
    test = data[xvs==i, ]
    train = data[xvs!=i, ]
    if(smote){
      # apply SMOTE to the training data:
      train = SMOTE(total ~ ridageyr+ bmxbmi + bmxarmc + bmxwaist + bmxthicr +
          bmxtri +
                      bmxsub +dxdavglpf + dxdavgapf + dxdtrpf + dxdtopf,
                    data=train, perc.over=perc.over, perc.under=perc.under)
    }
    if(samp){ # stratified sampling for the two groups, size of smaller group in
        each stratum
      sampsize = rep(min(table(train$total)), 2)
      strata = train$total
    } else { # stratified sampling with all observations in the same stratum
      # (this is the same as unstratified sampling)
```

```
      sampsize = nrow(train)
      strata = factor(rep(1, nrow(train)))
    }
    set.seed(1234)
    glub = randomForest(as.factor(total) ~ ridageyr+ bmxbmi + bmxarmc + bmxwaist
        + bmxthicr +
                          bmxtri + bmxsub + dxdavglpf + dxdavgapf + dxdtrpf +
                              dxdtopf,
                        classwt=weight, cutoff=c(co, 1-co),
                        strata=strata, sampsize=sampsize, data=train)
    probs[xvs==i] = predict(object=glub, newdata=test, type="prob")[, 2]
  }
  acc = class.sum(data$total, probs, co)
  list(acc=acc, probs=probs)
}

#-------Data
data = list()
data[[1]] = total.hispm
data[[2]] = total.blackm
data[[3]] = total.whitem

table1 = list()
table1[[1]] = table1[[2]] = table1[[3]] = matrix(0, nrow=27, ncol=5)

for (x in 1:3) {

  table1[[x]][1, ] = unlist(rf.10cv(data[[x]], co=0.5)$acc)
  table1[[x]][2, ] = unlist(rf.10cv(data[[x]], co=0.6)$acc)
  table1[[x]][3, ] = unlist(rf.10cv(data[[x]], co=0.7)$acc)
  table1[[x]][4, ] = unlist(rf.10cv(data[[x]], co=0.8)$acc)

  weight = c(1, 0.2)
  table1[[x]][5, ] = unlist(rf.10cv(data[[x]], co=0.5, classwt=weight)$acc)
  table1[[x]][6, ] = unlist(rf.10cv(data[[x]], co=0.6, classwt=weight)$acc)
```

```
  table1[[x]][7, ] = unlist(rf.10cv(data[[x]], co=0.7, classwt=weight)$acc)
  table1[[x]][8, ] = unlist(rf.10cv(data[[x]], co=0.8, classwt=weight)$acc)

  weight = c(1, 0.4)
  table1[[x]][9, ] = unlist(rf.10cv(data[[x]], co=0.5, classwt=weight)$acc)
  table1[[x]][10, ] = unlist(rf.10cv(data[[x]], co=0.6, classwt=weight)$acc)
  table1[[x]][11, ] = unlist(rf.10cv(data[[x]], co=0.7, classwt=weight)$acc)
  table1[[x]][12, ] = unlist(rf.10cv(data[[x]], co=0.8, classwt=weight)$acc)

  weight = c(1, 0.6)
  table1[[x]][13, ] = unlist(rf.10cv(data[[x]], co=0.5, classwt=weight)$acc)
  table1[[x]][14, ] = unlist(rf.10cv(data[[x]], co=0.6, classwt=weight)$acc)
  table1[[x]][15, ] = unlist(rf.10cv(data[[x]], co=0.7, classwt=weight)$acc)
  table1[[x]][16, ] = unlist(rf.10cv(data[[x]], co=0.8, classwt=weight)$acc)

  weight = c(1, 0.8)
  table1[[x]][17, ] = unlist(rf.10cv(data[[x]], co=0.5, classwt=weight)$acc)
  table1[[x]][18, ] = unlist(rf.10cv(data[[x]], co=0.6, classwt=weight)$acc)
  table1[[x]][19, ] = unlist(rf.10cv(data[[x]], co=0.7, classwt=weight)$acc)
  table1[[x]][20, ] = unlist(rf.10cv(data[[x]], co=0.8, classwt=weight)$acc)

  table1[[x]][21, ] = unlist(rf.10cv(data[[x]], co=0.5, samp=TRUE)$acc)
  table1[[x]][22, ] = unlist(rf.10cv(data[[x]], co=0.6, samp=TRUE)$acc)
  table1[[x]][23, ] = unlist(rf.10cv(data[[x]], co=0.7, samp=TRUE)$acc)
  table1[[x]][24, ] = unlist(rf.10cv(data[[x]], co=0.8, samp=TRUE)$acc)

  table1[[x]][25, ] = unlist(rf.10cv(data[[x]], co=0.5, smote=TRUE, perc.over
      =100, perc.under=200)$acc)
  table1[[x]][26, ] = unlist(rf.10cv(data[[x]], co=0.5, smote=TRUE, perc.over
      =200, perc.under=200)$acc)
  table1[[x]][27, ] = unlist(rf.10cv(data[[x]], co=0.5, smote=TRUE, perc.over
      =250, perc.under=300)$acc)

}
```

```
#======================
# Fig 3.1. Bubble Plot
#======================

#x=1 #Hispanic
x=2 #Black
#x=3 #White

table2 = list()
  table2[[x]] = cbind(table1[[x]], c(1:27))
  table2[[x]] = data.frame(table2[[x]], method=c(rep("RF",4), rep("WRF", 16),
      rep("DS",4), rep("SMOTE",3)))
  colnames(table2[[x]]) = c("PCC", "specificity", "sensitivity", "kappa", "ROC",
       "no.", "method")

  ggplot(table2[[x]], aes(x=table2[[x]][,2], y=table2[[x]][,3], label=table2[[x
      ]][,6]))+
    geom_jitter(aes(size = PCC, colour = method, alpha=0.3), show.legend = T) +
    geom_text(hjust = 1, size = 5) +
    scale_size(range = c(1,30)) +
    scale_x_continuous(name="specificity", limits=c(0.25,1))+
    scale_y_continuous(name="sensitivity", limits=c(0,1))+
    theme_bw(base_size=24)+
    guides(size=guide_legend(order=1), color=guide_legend(order=2))+
    guides(colour=guide_legend(override.aes=list(size=6)))+
    guides(alpha=FALSE)

################################################
# Fig 3.2 Variable Importance Plot
################################################

#-- Hispanic, SMOTE (250,300)
set.seed(1234)
hisp.smote = SMOTE(total ~ ridageyr + bmxbmi  + bmxarmc + bmxwaist
```

```
                                        + bmxthicr + bmxtri + bmxsub +dxdavglpf +
                                            dxdavgapf +dxdtrpf + dxdtopf,
                                        data=total.hispm, perc.over =250, perc.
                                            under=300)
set.seed(1234)
rf.hisp = randomForest(as.factor(total) ~  ridageyr+bmxbmi  + bmxarmc + bmxwaist
                     + bmxthicr + bmxtri + bmxsub +dxdavglpf + dxdavgapf +dxdtrpf
                          + dxdtopf,
                     ntree=10000, proximity=TRUE, importance=TRUE, data=hisp.
                         smote)
varImpPlot(rf.hisp)

#-- Black, WRF 5:1, cutoff = 0.6

set.seed(1234)
rf.black = randomForest(as.factor(total) ~ ridageyr+ bmxbmi + bmxarmc + bmxwaist
     + bmxthicr +
                       bmxtri + bmxsub + dxdavglpf + dxdavgapf + dxdtrpf +
                           dxdtopf,
                      classwt=c(1,0.2), cutoff=c(0.6, 0.4),
                      strata=factor(rep(1, nrow(total.blackm))), sampsize=nrow(
                          total.blackm),
                      ntree=10000, proximity=TRUE, importance=TRUE, data=total.
                          blackm)
varImpPlot(rf.black)

#-- White, DS cutoff=0.5

set.seed(1234)
rf.white1 = randomForest(as.factor(total) ~ ridageyr+ bmxbmi + bmxarmc +
    bmxwaist + bmxthicr +
                       bmxtri + bmxsub + dxdavglpf + dxdavgapf + dxdtrpf +
                           dxdtopf,
                       sampsize = rep(min(table(total.whitem$total)), 2),
                       strata = total.whitem$total,
```

```
                    ntree=10000, proximity=TRUE, importance=TRUE, data=total.
                        whitem)
varImpPlot(rf.white1)


########################################
# FIg 3.2: Predicted Probability Plot
########################################
library(shape)
library(plyr)

data.p = list()
data.p[[1]] = hisp.smote
data.p[[2]] = total.blackm
data.p[[3]] = total.whitem


#=====HISPANIC

x=1
var1_vals.wm = seq(from = min(data.p[[x]]$bmxwaist), to = max(data.p[[x]]$
    bmxwaist),
                    by = (max(data.p[[x]]$bmxwaist)-min(data.p[[x]]$bmxwaist))/
                        39)

  var2_vals.wm = seq(from = min(data.p[[x]]$bmxbmi), to = max(data.p[[x]]$bmxbmi
      ),
                    by = (max(data.p[[x]]$bmxbmi)-min(data.p[[x]]$bmxbmi))/39)
  two_vals.wm=NULL
  two_vals.wm = expand.grid(var1_vals.wm, var2_vals.wm)
  two_vals.wm = arrange(two_vals.wm, Var1, Var2)
  colnames(two_vals.wm) = c("bmxwaist", "bmxbmi")
  two_vals.wm$ridageyr = (mean(data.p[[1]]$ridageyr)+mean(data.p[[2]]$ridageyr)+
      mean(data.p[[3]]$ridageyr))/3
```

```
#two_vals.wm$bmxbmi = (mean(total.hm.s$bmxbmi) + mean(total.bm.s$bmxbmi) +
    mean(total.wm.s$bmxbmi))/3
two_vals.wm$bmxarmc = (mean(data.p[[1]]$bmxarmc) +mean(data.p[[2]]$bmxarmc) +
    mean(data.p[[3]]$bmxarmc))/3
two_vals.wm$dxdavgapf = (mean(data.p[[1]]$dxdavgapf)+mean(data.p[[2]]$
    dxdavgapf)+mean(data.p[[3]]$dxdavgapf))/3
two_vals.wm$dxdavglpf = (mean(data.p[[1]]$dxdavglpf)+mean(data.p[[2]]$
    dxdavglpf)+mean(data.p[[3]]$dxdavglpf))/3
two_vals.wm$dxdtopf = (mean(data.p[[1]]$dxdtopf)+mean(data.p[[2]]$dxdtopf)+
    mean(data.p[[3]]$dxdtopf))/3
two_vals.wm$bmxthicr = (mean(data.p[[1]]$bmxthicr)+mean(data.p[[2]]$bmxthicr)+
    mean(data.p[[3]]$bmxthicr))/3
two_vals.wm$dxdtrpf = (mean(data.p[[1]]$dxdtrpf) +mean(data.p[[2]]$dxdtrpf)+
    mean(data.p[[3]]$dxdtrpf))/3
two_vals.wm$bmxtri = (mean(data.p[[1]]$bmxtri)+mean(data.p[[2]]$bmxtri) + mean
    (data.p[[3]]$bmxtri))/3
two_vals.wm$bmxsub = (mean(data.p[[1]]$bmxsub)+mean(data.p[[2]]$bmxsub)+mean(
    data.p[[3]]$bmxsub))/3

k.wm= predict(object = rf.hisp, newdata=two_vals.wm,type="prob")[,2]
two_vals.wm = cbind(two_vals.wm, k.wm)
k.mx.wm = matrix(k.wm, ncol=40, nrow=40)
ylim = range(k.mx.wm)

#---2D plot
par(cex=0.8, oma=c(1,1,1,1)+0.1, mar=c(1,1,1,1)+1)
persp(var2_vals.wm, var1_vals.wm,  k.mx.wm, theta=45, phi=35, ticktype="
    detailed", nticks=8, xlab="bmi", ylab="waist",
      zlim=c(.0,0.8), zlab="predicted␣probability",
      main= "male")
persp(var2_vals.wm, var1_vals.wm,  k.mx.wm, theta=0, phi=10, ticktype="
    detailed", nticks=8, xlab="bmi", ylab="waist",
      zlim=c(.0,0.8), zlab="predicted␣probability",
      main= "male")
```

```
  persp(var2_vals.wm, var1_vals.wm,  k.mx.wm, theta=90, phi=10, ticktype="
      detailed", nticks=8, xlab="bmi", ylab="waist",
        zlim=c(.0,0.8), zlab="predicted probability",
        main= "male")

#===== Black

x=2
  var1_vals.wm = seq(from = min(data.p[[x]]$bmxwaist), to = max(data.p[[x]]$
      bmxwaist),
                     by = (max(data.p[[x]]$bmxwaist)-min(data.p[[x]]$bmxwaist))/
                         39)

  var2_vals.wm = seq(from = min(data.p[[x]]$bmxbmi), to = max(data.p[[x]]$bmxbmi
      ),
                     by = (max(data.p[[x]]$bmxbmi)-min(data.p[[x]]$bmxbmi))/39)
  two_vals.wm=NULL
  two_vals.wm = expand.grid(var1_vals.wm, var2_vals.wm)
  two_vals.wm = arrange(two_vals.wm, Var1, Var2)
  colnames(two_vals.wm) = c("bmxwaist", "bmxbmi")
  two_vals.wm$ridageyr = (mean(data.p[[1]]$ridageyr)+mean(data.p[[2]]$ridageyr)+
      mean(data.p[[3]]$ridageyr))/3
  #two_vals.wm$bmxbmi = (mean(total.hm.s$bmxbmi) + mean(total.bm.s$bmxbmi) +
      mean(total.wm.s$bmxbmi))/3
  two_vals.wm$bmxarmc = (mean(data.p[[1]]$bmxarmc) +mean(data.p[[2]]$bmxarmc) +
      mean(data.p[[3]]$bmxarmc))/3
  two_vals.wm$dxdavgapf = (mean(data.p[[1]]$dxdavgapf)+mean(data.p[[2]]$
      dxdavgapf)+mean(data.p[[3]]$dxdavgapf))/3
  two_vals.wm$dxdavglpf = (mean(data.p[[1]]$dxdavglpf)+mean(data.p[[2]]$
      dxdavglpf)+mean(data.p[[3]]$dxdavglpf))/3
  two_vals.wm$dxdtopf = (mean(data.p[[1]]$dxdtopf)+mean(data.p[[2]]$dxdtopf)+
      mean(data.p[[3]]$dxdtopf))/3
  two_vals.wm$bmxthicr = (mean(data.p[[1]]$bmxthicr)+mean(data.p[[2]]$bmxthicr)+
      mean(data.p[[3]]$bmxthicr))/3
```

```
  two_vals.wm$dxdtrpf = (mean(data.p[[1]]$dxdtrpf) +mean(data.p[[2]]$dxdtrpf)+
      mean(data.p[[3]]$dxdtrpf))/3
  two_vals.wm$bmxtri = (mean(data.p[[1]]$bmxtri)+mean(data.p[[2]]$bmxtri) + mean
      (data.p[[3]]$bmxtri))/3
  two_vals.wm$bmxsub = (mean(data.p[[1]]$bmxsub)+mean(data.p[[2]]$bmxsub)+mean(
      data.p[[3]]$bmxsub))/3

  k.wm= predict(object = rf.black, newdata=two_vals.wm,type="prob")[,2]
  two_vals.wm = cbind(two_vals.wm, k.wm)
  k.mx.wm = matrix(k.wm, ncol=40, nrow=40)
  ylim = range(k.mx.wm)

  #---2D plot
  par(cex=0.8, oma=c(1,1,1,1)+0.1, mar=c(1,1,1,1)+1)
  persp(var2_vals.wm, var1_vals.wm,  k.mx.wm, theta=45, phi=35, ticktype="
      detailed", nticks=8, xlab="bmi", ylab="waist",
        zlim=c(.0,0.8), zlab="predicted probability",
        main= "male")
  persp(var2_vals.wm, var1_vals.wm,  k.mx.wm, theta=0, phi=10, ticktype="
      detailed", nticks=8, xlab="bmi", ylab="waist",
        zlim=c(.0,0.8), zlab="predicted probability",
        main= "male")
  persp(var2_vals.wm, var1_vals.wm,  k.mx.wm, theta=90, phi=10, ticktype="
      detailed", nticks=8, xlab="bmi", ylab="waist",
        zlim=c(.0,0.8), zlab="predicted probability",
        main= "male")


#===== WHITE
  x=3
  var1_vals.wm = seq(from = min(data.p[[x]]$bmxwaist), to = max(data.p[[x]]$
      bmxwaist),
                     by = (max(data.p[[x]]$bmxwaist)-min(data.p[[x]]$bmxwaist))/
                         39)
```

```
var2_vals.wm = seq(from = min(data.p[[x]]$bmxbmi), to = max(data.p[[x]]$bmxbmi
    ),
                   by = (max(data.p[[x]]$bmxbmi)-min(data.p[[x]]$bmxbmi))/39)
two_vals.wm=NULL
two_vals.wm = expand.grid(var1_vals.wm, var2_vals.wm)
two_vals.wm = arrange(two_vals.wm, Var1, Var2)
colnames(two_vals.wm) = c("bmxwaist", "bmxbmi")
two_vals.wm$ridageyr = (mean(data.p[[1]]$ridageyr)+mean(data.p[[2]]$ridageyr)+
    mean(data.p[[3]]$ridageyr))/3
#two_vals.wm$bmxbmi = (mean(total.hm.s$bmxbmi) + mean(total.bm.s$bmxbmi) +
    mean(total.wm.s$bmxbmi))/3
two_vals.wm$bmxarmc = (mean(data.p[[1]]$bmxarmc) +mean(data.p[[2]]$bmxarmc) +
    mean(data.p[[3]]$bmxarmc))/3
two_vals.wm$dxdavgapf = (mean(data.p[[1]]$dxdavgapf)+mean(data.p[[2]]$
    dxdavgapf)+mean(data.p[[3]]$dxdavgapf))/3
two_vals.wm$dxdavglpf = (mean(data.p[[1]]$dxdavglpf)+mean(data.p[[2]]$
    dxdavglpf)+mean(data.p[[3]]$dxdavglpf))/3
two_vals.wm$dxdtopf = (mean(data.p[[1]]$dxdtopf)+mean(data.p[[2]]$dxdtopf)+
    mean(data.p[[3]]$dxdtopf))/3
two_vals.wm$bmxthicr = (mean(data.p[[1]]$bmxthicr)+mean(data.p[[2]]$bmxthicr)+
    mean(data.p[[3]]$bmxthicr))/3
two_vals.wm$dxdtrpf = (mean(data.p[[1]]$dxdtrpf) +mean(data.p[[2]]$dxdtrpf)+
    mean(data.p[[3]]$dxdtrpf))/3
two_vals.wm$bmxtri = (mean(data.p[[1]]$bmxtri)+mean(data.p[[2]]$bmxtri) + mean
    (data.p[[3]]$bmxtri))/3
two_vals.wm$bmxsub = (mean(data.p[[1]]$bmxsub)+mean(data.p[[2]]$bmxsub)+mean(
    data.p[[3]]$bmxsub))/3

k.wm= predict(object = rf.white1, newdata=two_vals.wm,type="prob")[,2]
two_vals.wm = cbind(two_vals.wm, k.wm)
k.mx.wm = matrix(k.wm, ncol=40, nrow=40)
ylim = range(k.mx.wm)

#---2D plot
par(cex=0.8, oma=c(1,1,1,1)+0.1, mar=c(1,1,1,1)+1)
```

```
persp(var2_vals.wm, var1_vals.wm,  k.mx.wm, theta=45, phi=35, ticktype="
    detailed", nticks=8, xlab="bmi", ylab="waist",
      zlim=c(.0,0.8), zlab="predicted probability",
      main= "male")
persp(var2_vals.wm, var1_vals.wm,  k.mx.wm, theta=0, phi=10, ticktype="
    detailed", nticks=8, xlab="bmi", ylab="waist",
      zlim=c(.0,0.8), zlab="predicted probability",
      main= "male")
persp(var2_vals.wm, var1_vals.wm,  k.mx.wm, theta=90, phi=10, ticktype="
    detailed", nticks=8, xlab="bmi", ylab="waist",
      zlim=c(.0,0.8), zlab="predicted probability",
      main= "male")
```