Understanding Correlates of Obesity: Supervised and Unsupervised Learning Approaches

Katherine M. Prioli

December 05, 2019

Abstract

Background Increasing overweight and obesity represent an important national health concern and a significant, preventable driver of healthcare expenditure. Every two years, the National Health and Nutrition Examination Survey (NHANES) collects data about Americans' dietary and exercise behaviors along with demographics and relevant clinical characteristics. If these characteristics can be used to identify overweight and obesity at a population level, interventions designed to reduce excess adiposity could be appropriately targeted to those who could most benefit. This study applied supervised and unsupervised machine learning techniques to categorize cases by Body Mass Index (BMI) weight category. ***Methods*** Medical, demographic, and behavioral variables known or suspected to be correlated with obesity were selected from the 2015-2016 NHANES data. The dataset was subset to adult, missing data were imputed, and BMI weight category (four levels) was computed for each case. Random forest classification (supervised approach) was performed using $n \times k$-fold crossvalidation on both the initial dataset and a smaller dataset with dichotomous BMI categories. Hierarchical clustering (unsupervised approach) was performed on a 5% sample stratified by BMI category using both agglomerative (AGNES) and divisive (DIANA) methods, and percentages of cases in each BMI category were calculated for each cluster. ***Results*** Text ***Conclusion*** Text

Background

Overweight and obesity are growing public health concerns, with 71.6% of American adults meeting at least the criteria for overweight, and 39.8% qualifying as obese per the Centers for Disease Control and Prevention (CDC) as of 2016.[1] Defined as a body mass index (BMI) of 30 kg/m^2 or above, obesity is of especially high concern because it strongly correlates with many diseases and conditions that lead to significant morbidity and mortality, and thus to increased healthcare utilization and decreased productivity, both at high economic burden.[2-4]

The National Health and Nutrition Examination Survey (NHANES), a biennial health surveillance study performed by the CDC, collects data from a nationally representative sample of community-dwelling Americans pertaining to demographics, health, and nutritional and exercise habits through both self-report and physical examination.[5] NHANES has uncovered some demographic correlates of obesity − namely, that obesity is more prevalent among Hispanics and non-Hispanic blacks for both adults and children, that obesity is more prevalent among women across all racial groups, and that for all racial groups except blacks, education level appears inversely correlated with prevalence of obesity.[6,7] However, though much progress has been made toward understanding the socioeconomic and behavioral drivers of obesity, much remains unknown about the interactions of these characteristics and how constellations thereof could be used to detect obesity at the population level.

Reducing both the incidence and prevalence of obesity will be critical in managing obesity-related healthcare expenditure, especially for publicly funded programs such as Medicare and Medicaid. If constellations of obesity correlates can be identified, tailored health interventions that target these constellations to prevent or reduce obesity could be developed. Using the 2015-2016 NHANES data, the objective of this study was to apply both supervised and unsupervised machine learning approaches to find patterns in a carefully selected set of characteristics which are known or suspected to be correlated with obesity.

Methods

Variable Selection

A literature search of PubMed for currently known or suspected demographic, behavioral, and medical correlates of excess adiposity was used to inform variable selection. NHANES 2015-2016 variables chosen for the study are presented in Table 1, with the source denoted as "native" for variables native to the NHANES dataset or "derived" for variables calculated from native NHANES variables. Derived variables of note included the scored and categorized nine-item Patient Health Questionnaire (PHQ-9), a validated nine-item depression inventory, as well as `BP_cat`, representing clinical categories of blood pressure ranging from hypotension through hypertensive crisis, assigned based on systolic and diastolic blood pressure values. Additionally, many of the native categorical variables were refactored to group nonresponses (e.g., refusals to respond, "Don't know" responses, and missing values) into a single category.

Table 1. Variables in initial dataset.

Variable	Source	Description	Reference(s)
SEQN	native	Unique identifier	[8]
RIAGENDR	native	Gender	[8]
RIDAGEYR	native	Age	[8]
RIDRETH3	native	Race/ethnicity	[8]
DMDEDUC2	native	Education level	[8]
DMDMARTL	native	Marital status	[8]
INDFMIN2	native	Annual family income	[8]
INDFMPIR	native	Ratio of family income to poverty	[8]
DIQ010	native	History of diabetes	[9]
DIQ280	native	Last hemoglobin A1C (HbA1C) level	[9]
MCQ160c	native	History of coronary heart disease	[10]
MCQ160e	native	History of myocardial infarction	[10]
MCQ160m	native	History of thyroid disease	[10]
MCQ365a	native	Doctor has told to lose weight	[10]
MCQ365b	native	Doctor has told to exercise	[10]
BPXSY2	native	Systolic blood pressure ($mmHg$)	[11]
BPXDI2	native	Diastolic blood pressure ($mmHg$)	[11]
BP_cat	derived	Clinical blood pressure category	[12, 13]; native variables BPXSY2, BPXDI2
PAD615	native	Average minutes of vigorous physical work per day	[14]
PAD630	native	Average minutes of moderate physical work per day	[14]
PAD660	native	Average minutes of vigorous physical recreational activity per day	[14]
PAD675	native	Average minutes of moderate physical recreational activity per day	[14]
mins_activ	derived	Average minutes of moderate and/or vigorous activity per day	Native variables PAD615, PAD630, PAD660, PAD675
PAD680	native	Average minutes awake and sedentary per day	[14]
PFQ049	native	Unable to work due to impairment	[15]
PFQ061	native	Difficulty walking 1/4 mile	[15]
DPQ010 - DPQ090	native	Depression inventory subscores	[16]
PHQ9_score	derived	Depression inventory score	[17]; native variables DPQ010 - DPQ090
PHQ9_cat	derived	Depression inventory category	[17]; native variables DPQ010 - DPQ090
DBQ700	native	Self-perception of dietary healthiness	[18]
CBQ505	native	Ate fast food or pizza within past 12 months	[18]
CBQ540	native	Used nutritional information to choose fast foods	[18]
CBQ545	native	Would use nutritional information to choose fast foods	[18]
CBQ550	native	Ate at a restaurant with waitstaff in past 12 months	[18]
CBQ585	native	Used nutritional information to choose restaurant meal	[18]
CBQ590	native	Would use nutritional information to choose restaurant meal	[18]
DR1TKCAL	native	Dietary intake, Day 1 ($kcal$)	[19]
DR1300	native	Day 1 dietary intake compared to usual	[19]
DR1_32OZ	native	Water intake, Day 1 (g)	[19]
BMXBMI	native	Body Mass Index (kg/m^2)	[20]
BMI_cat	derived	BMI weight category	[2]; native variable BMXBMI

Because BMI categories differ for children and adults, the analysis was limited to ages ≥ 20, consistent with the NHANES definition of adult.[8] Additionally, since `BMI_cat` represents the data labels, all rows with null `BMI_cat` were excluded.

Descriptive statistics and frequency tables were generated for continuous and categorical variables respectively to understand data contents and to assess the degree of missingness among the continuous data. Missing values were imputed for continuous data. To avoid introducing bias through imputation, two-sided Wilcoxon Rank-Sum tests were performed for each variable on the pre- vs. post-imputation data to identify any statistically significant changes introduced by imputation at the $\alpha = 0.05$ level. Any variables having statistically different post-imputation data were removed from the dataset.

Supervised Approach

Using algorithms available in Scikit-learn, a random forest of decision tree classifiers was used with $n \times k$-fold crossvalidation to classify cases into BMI category, and descriptive statistics were generated to assess model accuracy. To improve performance, hyperparameters `n_repeats` (or n, number of times crossvalidation was performed) and `n_splits` (or k, number of folds) in `RepeatedKFold()` were tuned, along with hyperparameters `n_estimators` (number of decision trees in the forest), `min_samples_leaf` (minimum cases allowed per terminal node), and `max_depth` (maximum tree depth) in `RandomForestClassifier`.

To further improve accuracy, the labels representing BMI category were recast as dichotomous, grouping together the underweight and healthy weight categories, versus the overweight and obese categories. Additionally, the six comorbidity variables (pertaining to diabetes, coronary artery disease, myocardial infarction, thyroid disease, hypertension, and depression) were consolidated into one variable representing number of comorbidities, and variables which were suspected not to contribute much information to the model (due to low variation or large amount of missingness) were excluded. This modified dataset was then run through $N \times k$-fold crossvalidation with hyperparameter tuning and model accuracy assessment as previously described.

Unsupervised Approach

For the unsupervised component of this analysis, data was subset to variables with $\geq 90\%$ non-missing values, then all cases with missing values were dropped. Since this approach involves visual analysis via dendrograms, the data was further subset to a random 5% sample within each BMI category to allow for sensible, interpretable dendrogram plots. The analytic dataset thus comprised 241 cases having 17 variables. To generate dendrograms for agglomerative clustering (AGNES) and divisive clustering (DIANA), a dissimilarity matrix was calculated based on Gower distance.[21] The Gower distance was chosen because the variables in this dataset are of mixed type (i.e., some continuous, some ordinal, and some nominal). Because the data is likely noisy, AGNES relied on complete (i.e., maximum distance) linkage between clusters.

Radial dendrograms for both AGNES and DIANA were plotted along with colored labels at each leaf to indicate the four BMI weight categories. For both AGNES and DIANA, cluster size was assessed and within-cluster sums of squares and average silhouette width were calculated based varying the number of clusters up to an upper limit of 15; these latter two metrics were used to generate scree and silhouette plots, which were used to determine an initial cluster number for each approach.[22] The radial dendrograms were replotted using these values to inform cluster coloration. Homogeneity of label colors in each cluster was assessed via both visual and numeric inspection, and the cluster number hyperparameter k was tuned as needed to improve homogeneity.

Technologies

Data import, wrangling, and visualization were performed in R using an RMarkdown notebook and relying primarily on the `tidyverse` ecosystem. The $n \times k$ crossvalidation random forest was implemented via Python code chunks in the notebook leveraging Scikit-Learn along with other common Python libraries as necessary (*e.g.*, `numpy` and `pandas`).[23-25] For the agglomerative and divisive clustering, R packages `cluster`, `fpc`, `ggdendro` and `dendextend` were used.[26-29] A public GitHub repository was established for this study and contains all raw data files and code along with project documentation.[30]

Results

The final analytic dataframe contained 5406 rows and 24 columns. Variables included in the dataset are presented in Table 2:

Table 2. Variables in final dataset.

Variable	Description
age	Age
BMI_cat	BMI weight category
dailykcal	Calories consumed on previous day
dailykcal_typical	Previous day's calorie consumption compared to usual

Variable	Description
dailywater	Previous day's water intake
diethealthy	Feels diet is healthy
educ	Level of education
famincome_cat	Family income category
fastfood_eat	Has eaten fast food in past 12 months
fastfood_usednutrit	Has used nutrition information to select fast food
fastfood_woulduse	Would use nutrition information to select fast food
gender	Gender
losewt_exer	Told by doctor to lose weight and/or exercise
marital	Marital status
mins_activ	Daily minutes of moderate to vigorous activity
mins_seden	Daily minutes spent sedentary
n_comorbid	Number of comorbidities
race	Race
restaur_eat	Has eaten at a restaurant in past 12 months
restaur_usednutrit	Has used nutrition information to select restaurant meal
restaur_woulduse	Would use nutrition information to select restaurant meal
seqn	Unique identifier (omitted from analyses)
walklim	Walking limitations
worklim	Work limitations

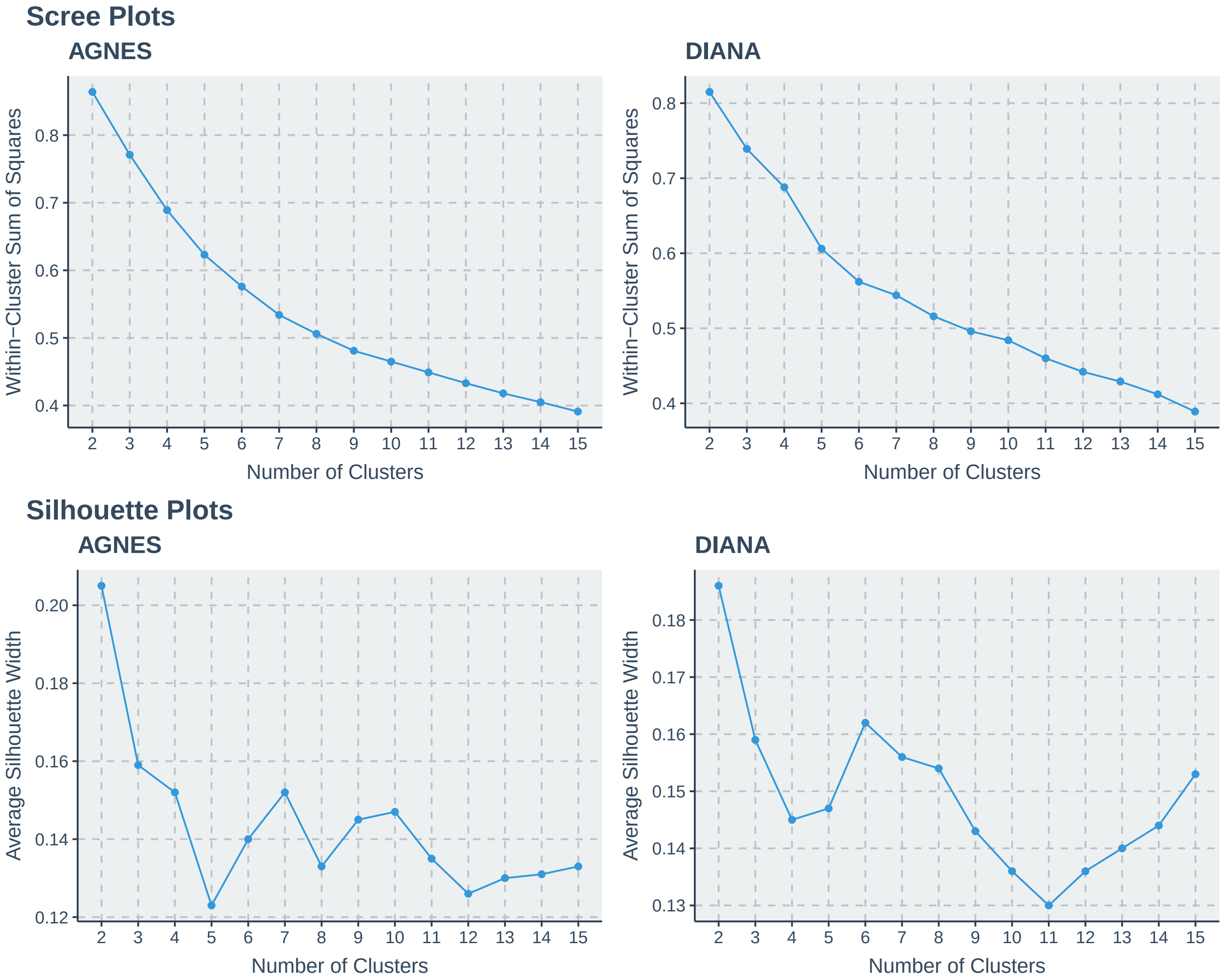
Supervised Approach

The initial model, based on the full dataset and using $k = 5$ folds, $n = 5$ repeats, and 100 trees in the ensemble without limitation on the number of cases per terminal node, yielded 52% mean accuracy (range: 49-55%). Variables included in the reduced dataset included age, gender, race, education level, marital status, family income category, number of comorbidities, number of daily minutes active and sedentary, having eaten fast food or at a restaurant in the past year, daily calories, daily water intake, and having been told by a doctor to lose weight and/or exercise, along with dichotomous BMI category. The final model based on this reduced dataset used $k = 3$ folds, $n = 10$ repeats, and 200 trees, and was limited to ≥ 25 cases per terminal node. This model yielded 75.6% mean accuracy (range: 74.4-77.4%), representing a meaningful improvement.

Unsupervised Approach

The scree and silhouette plots are presented in Figure 1 *a* and *b*. The AGNES scree plot shows a relatively smooth curve without strong elbows; possible but faint elbows appear at 5 and 7 clusters. For DIANA, two elbows are seen: one at 5 clusters and one at 6. Both the AGNES and DIANA silhouette plots show maximum average silhouette width at 2 clusters, which is unlikely to contain sufficient data to be meaningful. AGNES shows a local maximum at 7 clusters, and DIANA at 6. Considered together, scree and silhouette plots indicate that 7 clusters may be sufficient for AGNES, and 6 for DIANA.

Figure 3. Scree and Silhouette Plots



The initial radial dendrogram for AGNES is shown in Fig. 2 and the corresponding numeric analysis of BMI category by cluster number is presented in Table 3.

Figure 2. Initial Colored Radial Dendrogram, AGNES ($k = 7$)

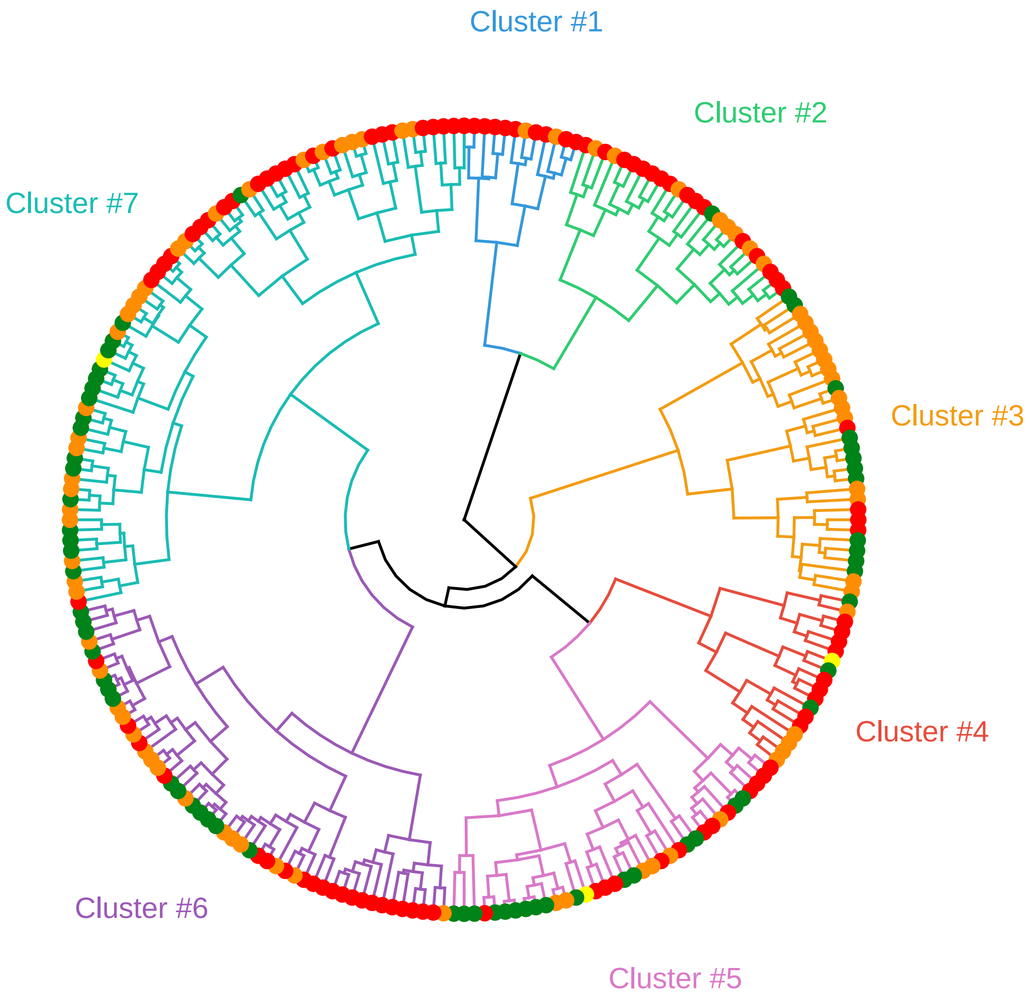
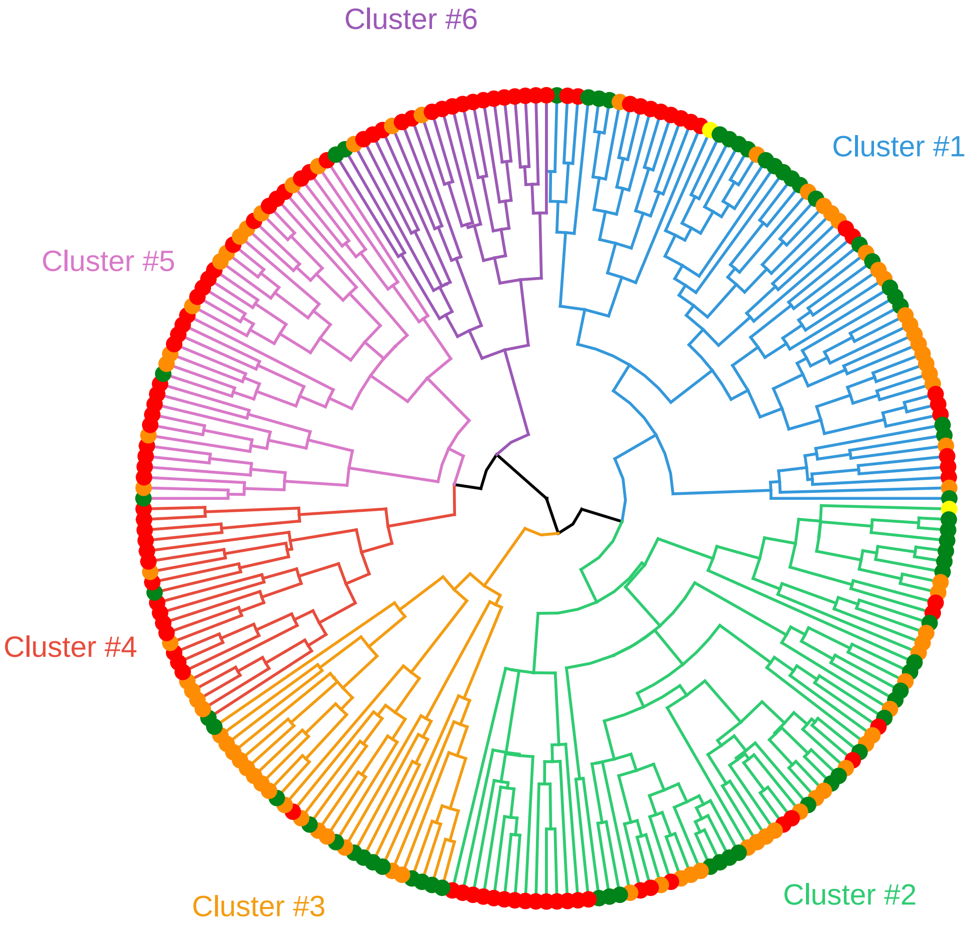


Table 3. BMI Categories by Cluster, Initial AGNES Analysis.

Cluster	Underweight (n, %)	Healthy Weight (n, %)	Overweight (n, %)	Obese (n, %)
1	0 (0%)	0 (0%)	2 (16.7%)	10 (83.3%)
2	0 (0%)	1 (4%)	8 (32%)	16 (64%)
3	0 (0%)	12 (38.7%)	15 (48.4%)	4 (12.9%)
4	1 (5.6%)	3 (16.7%)	5 (27.8%)	9 (50%)
5	1 (2.8%)	16 (44.4%)	6 (16.7%)	13 (36.1%)
6	0 (0%)	14 (28%)	15 (30%)	21 (42%)
7	1 (1.4%)	17 (24.6%)	26 (37.7%)	25 (36.2%)

For DIANA, the initial radial dendrogram is shown in Fig. 3 with analysis of BMI category by cluster number in Table 4.

Figure 3. Initial Colored Radial Dendrogram, DIANA ($k = 6$)



Cluster	Underweight (n, %)	Healthy Weight (n, %)	Overweight (n, %)	Obese (n, %)
1	2 (3.3%)	22 (36.1%)	19 (31.1%)	18 (29.5%)
2	1 (1.4%)	23 (33.3%)	22 (31.9%)	23 (33.3%)
3	0 (0%)	12 (42.9%)	15 (53.6%)	1 (3.6%)
4	0 (0%)	2 (9.1%)	6 (27.3%)	14 (63.6%)
5	0 (0%)	2 (5.1%)	12 (30.8%)	25 (64.1%)
6	0 (0%)	2 (9.1%)	3 (13.6%)	17 (77.3%)

Discussion

For the supervised approach, the final decision model reached 75.6% mean accuracy. While this is an improvement over the initial model, it generally would be considered poor performance for any real-world implementation.

Strengths

Limitations

Because the Scikit-Learn `randomForestClassifier()` function was unable to handle missing data, missing values were handled via imputation (continuous variables) or by creating a "missing" category comprising all-cause nonresponses (categorical variables). Using a "missing" category in this manner may have introduced noise into the dataset, because it's not known missing values were missing completely at random or missing not at random. Future iterations of this analysis should include other random forest functions - e.g., those available in the `xgboost` Python library.

Another limitation is that variable selection is difficult with continuous variables. Principal Component Analysis (PCA) is a common method used for choosing variables when considering continuous data, but is not commonly used when data is categorical. There is, however, evidence in the literature to suggest that categorical variables can be treated via a modified PCA approach, but the packages currently available for this either have limited documentation or are not built on a recent version of R. [31-33] In the absence of more sophisticated tools, variable selection was based on those characteristics expected to be most closely correlated with weight.

Conclusion

References

1. Selected health conditions and risk factors, by age: United States, selected years 1988-1994 through 2015-2016. Centers for Disease Control and Prevention. Health, United States, 2017: Trend Tables. https://www.cdc.gov/nchs/data/hus/2017/053.pdf. Accessed October 20, 2019.
2. Defining Adult Overweight and Obesity. Centers for Disease Control and Prevention. https://www.cdc.gov/obesity/adult/defining.html. Updated April 11, 2017. Accessed October 22, 2019.
3. Hruby A, Hu FB. The Epidemiology of Obesity: A Big Picture. *Pharmacoeconomics.* 2015;33(7):673-89. doi: 10.1007/s40273-014-0243-x.
4. Li Q, Blume SW, Huang JC, Hammer M, Ganz ML. Prevalence and healthcare costs of obesity-related comorbidities: evidence from an electronic medical records system in the United States. *J Med Econ.* 2015;18(12):1020-8. doi: 10.3111/13696998.2015.1067623.
5. National Health and Nutrition Examination Survey. National Center for Health Statistics. https://www.cdc.gov/nchs/nhanes/index.htm. Updated September 24, 2019. Accessed October 20, 2019.
6. Hales CM, Carroll MD, Fryar CD, Ogden CL. Prevalence of Obesity Among Adults and Youth: United States, 2015-2016. NCHS Data Brief No. 288. National Center for Health Statistics. 2017. https://www.cdc.gov/nchs/data/databriefs/db288.pdf. Accessed October 20, 2019.
7. Ogden CL, Fakhouri TH, Carroll MD, et al. Prevalence of Obesity Among Adults, by Household Income and Education – United States, 2011-2014. *Morb Mortal Wkly Rep.* 2017;66:1369-1373. doi: http://dx.doi.org/10.15585/mmwr.mm6650a1.
8. National Health and Nutrition Examination Survey: Demographic Variables and Sample Weights (DEMO_I). National Center for Health Statistics. https://wwwn.cdc.gov/Nchs/Nhanes/2015-2016/DEMO_I.XPT. Updated September 2017. Accessed October 27, 2019.
9. National Health and Nutrition Examination Survey: Questionnaire Data - Diabetes (DIQ_I). National Center for Health Statistics. https://wwwn.cdc.gov/Nchs/Nhanes/2015-2016/DIQ_I.XPT. Updated September 2017. Accessed October 29, 2019.
10. National Health and Nutrition Examination Survey: Questionnaire Data - Medical (MCQ_I). National Center for Health Statistics. https://wwwn.cdc.gov/Nchs/Nhanes/2015-2016/MCQ_I.XPT. Updated September 2017. Accessed October 29, 2019.
11. National Health and Nutrition Examination Survey: Examination Data - Blood Pressure (BPX_I). National Center for Health Statistics. https://wwwn.cdc.gov/Nchs/Nhanes/2015-2016/BPX_I.XPT. Updated September 2017. Accessed October 29, 2019.
12. Whelton PK, Carey RM, Aronow WS, et al. 2017 ACC/AHA/AAPA/ABC/ACPM/AGS/APhA/ASH/ASPC/NMA/PCNA Guideline for the Prevention, Detection, Evaluation, and Management of High Blood Pressure in Adults: A Report of

the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. *J Am Coll Cardiol.* 2018;71(19):e127-e248. doi: 10.1016/j.jacc.2017.11.006.

13. Low Blood Pressure. US National Library of Medicine: MedlinePlus. https://medlineplus.gov/ency/article/007278.htm. Updated November 06, 2019. Accessed November 10, 2019.

14. National Health and Nutrition Examination Survey: Questionnaire Data - Physical Activity (PAQ_I). National Center for Health Statistics. https://wwwn.cdc.gov/Nchs/Nhanes/2015-2016/PAQ_I.XPT. Updated September 2017. Accessed October 29, 2019.

15. National Health and Nutrition Examination Survey: Questionnaire Data - Physical Functioning (PFQ_I). National Center for Health Statistics. https://wwwn.cdc.gov/Nchs/Nhanes/2015-2016/PFQ_I.XPT. Updated September 2017. Accessed October 29, 2019.

16. National Health and Nutrition Examination Survey: Questionnaire Data - Mental Health: Depression Screener (DPQ_I). National Center for Health Statistics. https://wwwn.cdc.gov/Nchs/Nhanes/2015-2016/DPQ_I.XPT. Updated December 2017. Accessed October 29, 2019.

17. Kroenke K, Spitzer RL. The PHQ-9: A new depression diagnostic and severity measure. *Psychiatric Annals.* 2002;32(9):1-7 doi: 10.3928/0048-5713-20020901-06.

18. National Health and Nutrition Examination Survey: Questionnaire Data - Diet Behavior and Nutrition (DBQ_I). National Center for Health Statistics. https://wwwn.cdc.gov/Nchs/Nhanes/2015-2016/DBQ_I.XPT. Updated November 2018. Accessed October 29, 2019.

19. National Health and Nutrition Examination Survey: Questionnaire Data - Dietary Interview: Total Nutrient Intakes, First Day (DR1TOT_I). National Center for Health Statistics. https://wwwn.cdc.gov/Nchs/Nhanes/2015-2016/DR1TOT_I.XPT. Updated July 2018. Accessed October 29, 2019.

20. National Health and Nutrition Examination Survey: Examination Data - Body Measures (BMX_I). National Center for Health Statistics. https://wwwn.cdc.gov/Nchs/Nhanes/2015-2016/BMX_I.XPT. Updated September 2017. Accessed October 29, 2019.

21. Gower JC. A General Coefficient of Similarity and Some of Its Properties. *Biometrics.* 1971;27(4):857-71. doi: 10.2307/2528823.

22. Rousseeuw PJ. Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. *Journal of Computational and Applied Mathematics.* 1987(20): 53-65. https://doi.org/10.1016/0377-0427(87)90125-7.

23. sklearn.ensemble.RandomForestClassifier. Scikit-learn documentation. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html. Accessed November 10, 2019.

24. sklearn.model_selection.RepeatedKFold. Scikit-learn documentation. https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.RepeatedKFold.html. Accessed November 10, 2019.

25. sklearn.model_selection.cross_val_score. Scikit-learn documentation. https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.cross_val_score.html. Accessed November 10, 2019.

26. Maechler M, Rousseeuw P, Struyf A, Hubert M, Hornik K. cluster: Cluster Analysis Basics and Extensions. R package version 2.1.0. 2019. https://cran.r-project.org/package=cluster.

27. Hennig C. fpc: Flexible Procedures for Clustering. R package version 2.2-3. 2019. https://cran.r-project.org/package=fpc.

28. de Vries A, Ripley BD. ggdendro: Create Dendrograms and Tree Diagrams Using ‘ggplot2’. R package version 0.1-20. 2016. https://cran.r-project.org/package=ggdendro.

29. Galili T. dendextend: an R package for visualizing, adjusting, and comparing trees of hierarchical clustering. *Bioinformatics.* 2015;31(22):3718-20. doi: 10.1093/bioinformatics/btv428.

30. Prioli KM. CSC_8515_Final_Project. https://github.com/kprioliPROF/CSC_8515_Final_Project.

31. Niitsuma H, Okada T. Covariance and PCA for Categorical Variables. Advances in Knowledge Discovery and Data Mining. Ho TB, Cheung D, Liu H, eds. *Advances in Knowledge Discovery and Data Mining.* Berlin: Springer; 2005. doi: 10.1007/11430919_61. arXiv: 0711.4452 [cs.LG].

32. princals: Categorical principal component analysis (PRINCALS). Gifi Multivariate Analysis with Optimal Scaling documentation. https://cran.r-project.org/web/packages/Gifi/index.html. Updated June 25, 2019. Accessed November 17, 2019.

33. Chavent M, Kuentz-Simonet V, Labenne A, Saracco J. Multivariate Analysis of Mixed Data: The R Package PCAmixdata. arXiv:1411.4911v4 [stat.CO] Updated December 8, 2017. Accessed November 17, 2019.