Are you what you eat?

Allie Collins & Erika Tyagi

Recall: our aim was to explore the nutritional profiles of Americans and their implications

We used dimension reduction and clustering techniques to group nutritional profiles, which we used to inform answering the following questions:

- How does nutrition relate to socio-demographic profiles?
- How does nutrition relate to health outcomes?

Recall: we leveraged NHANES survey data from the CDC

The National Health and Nutrition Examination Survey (NHANES) is a biannual, nationwide survey intended to study the health and nutritional status of adults and children in the United States.

It involves five components:

Demographic

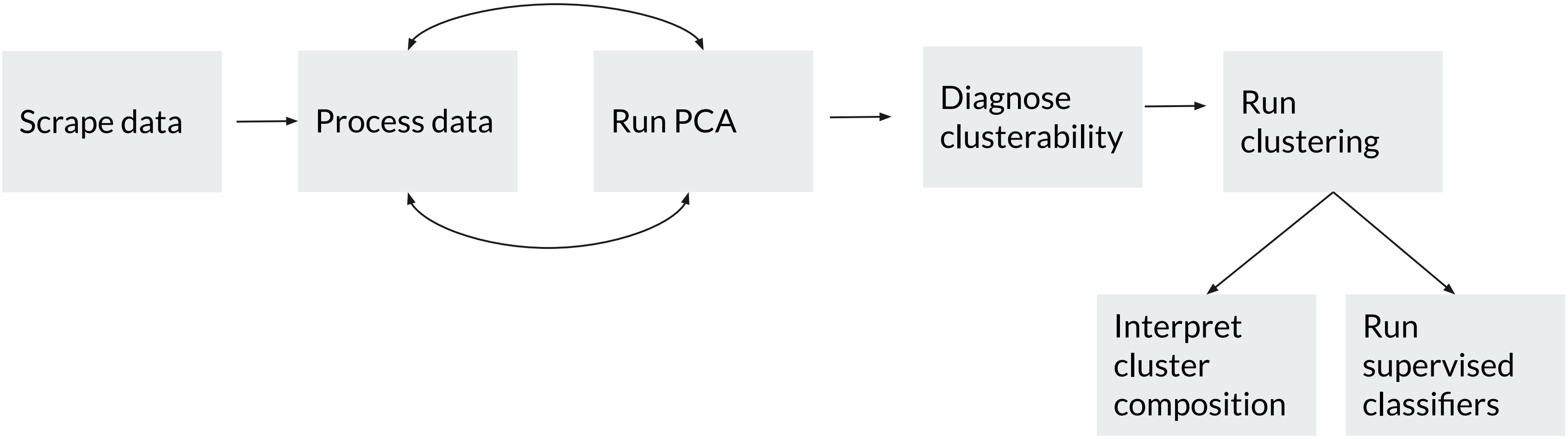
Dietary

Examination

Laboratory

Questionnaire

Methodology



Scrape data

Process data

Run PCA

Diagnose clusterability

Run clustering

Interpret cluster composition

Run supervised classifiers

Data processing and dimensionality reduction

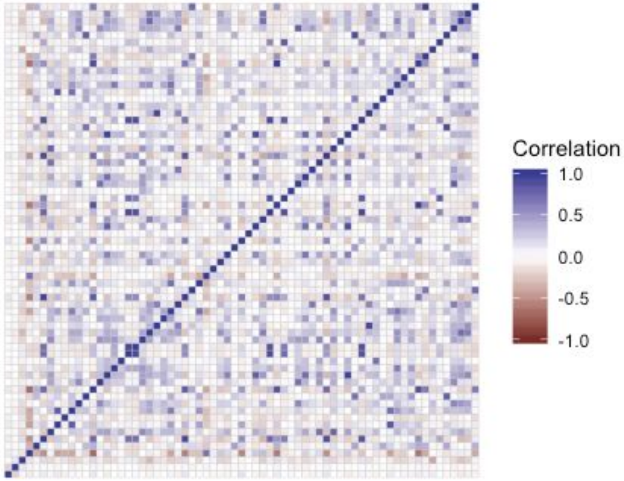
For our analysis we considered ~70 nutritional indicators

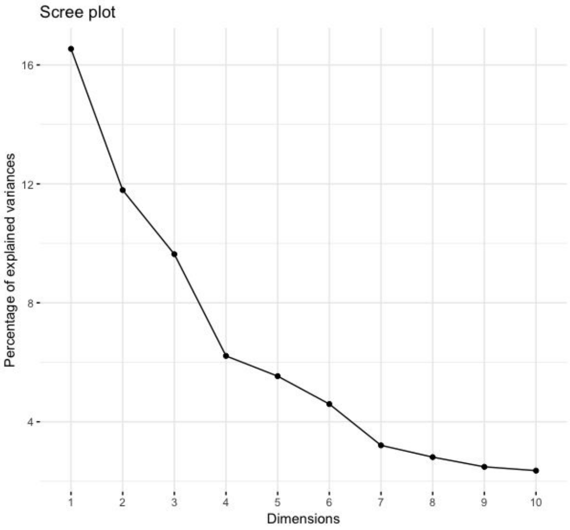
We limited to adults who completed both ‘Dietary Interview’ days. We then leveraged 2 questions to filter and ensure more accurate data:

- Whether this was an usual amount of food for the individual to consume
- Recall reliability status, as assessed by the CDC

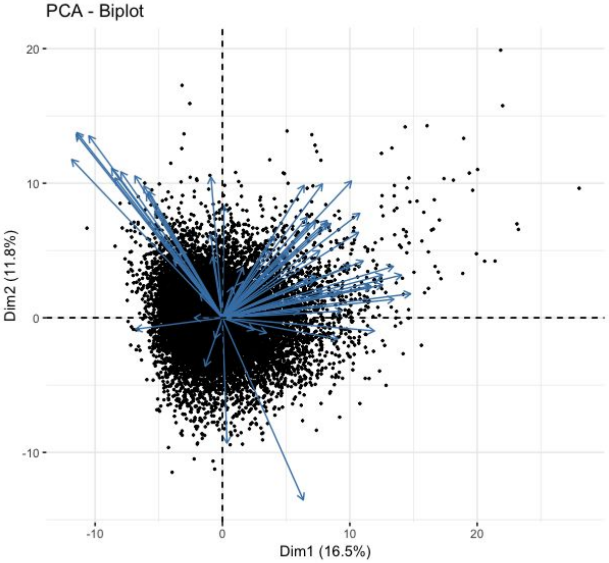
We then summed each nutrient across the two days and scaled by total caloric intake. Our final dataset included 13,623 individuals and 67 nutritional features from 2003 - 2016.

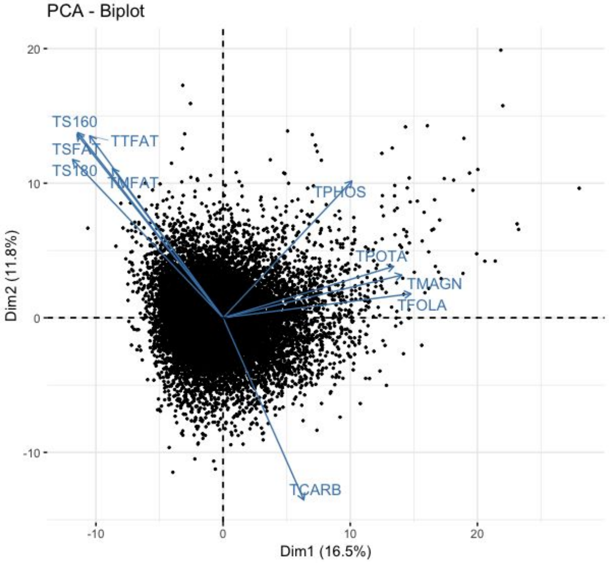
Given highly correlated variables, we used PCA to reduce our feature space to 7 dimensions





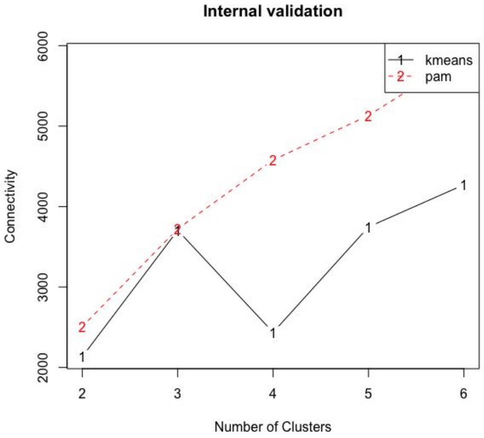
The principal components aligned with our preconceptions



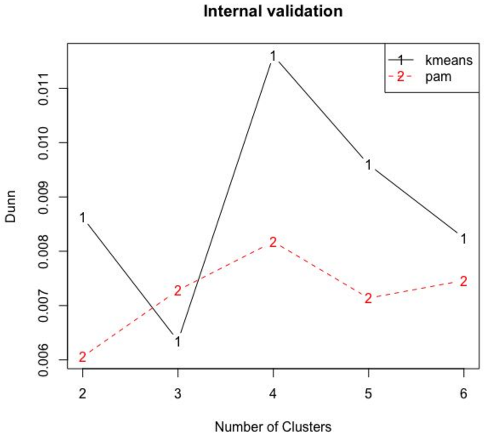


Clustering analysis

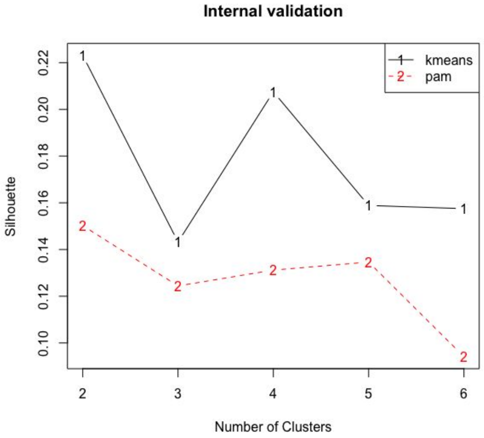
k-means with 2 clusters performs strongest on 2 out of 3 validation measures



Internal validation

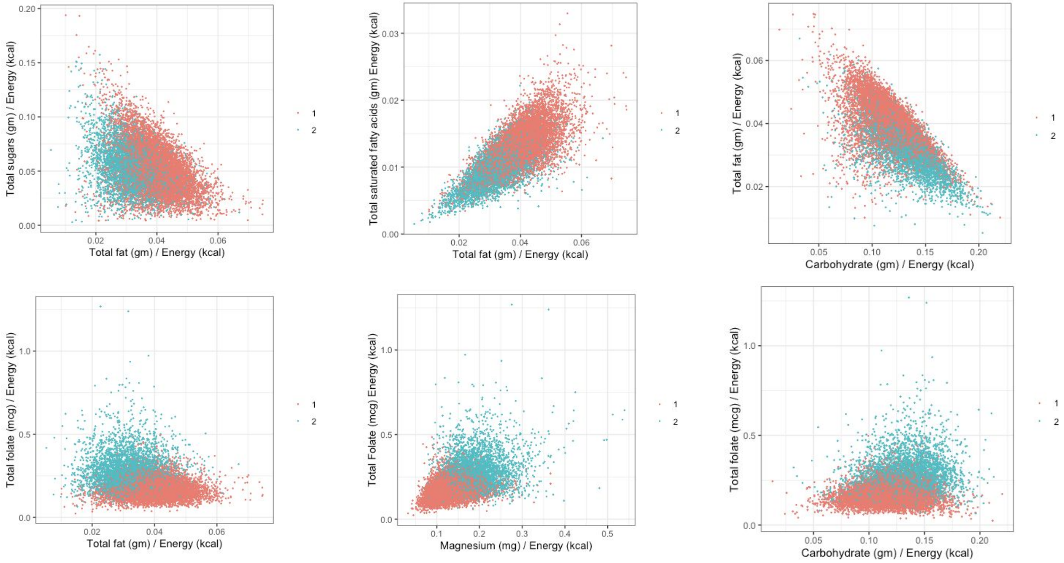


Internal validation

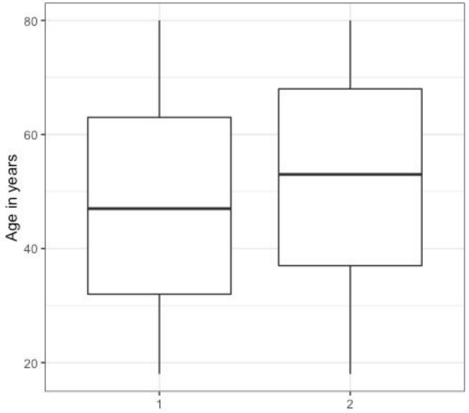


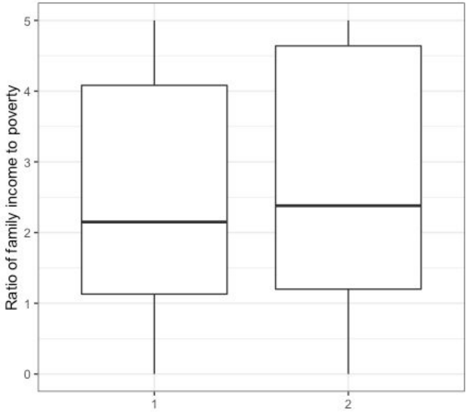
Internal validation

The clusters roughly correspond to 'healthy' and 'unhealthy' nutritional profiles



Some demographic differences emerge between clusters



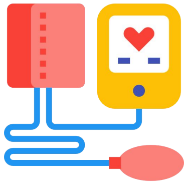


Cluster 1 is more white (52% vs. 46%), more black (19% vs. 12%) and less Hispanic (21% vs. 25%). Cluster 1 is also more male (53% vs. 44%).

Supervised classifiers

Overview of supervised approach (I/II)

- Research suggests social determinants of health associated with conditions such as obesity, high blood pressure - but what you eat is important, too!
- We take our components and clusters (separately) to predict, respectively, is (1) whether a person is obese and (2) whether a person has had high blood pressure
- We recognize these are not “perfect” outcomes/assessments, but our aim is to see what features are most important in this classification - demographic? nutritional?



Overview of supervised approach (II/II)

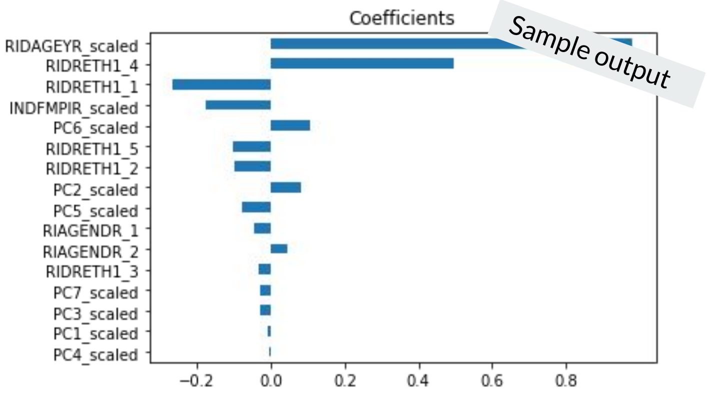
Literature suggests random forest classifiers and support vector machines (SVMs) most suited for this task, thus we ran:

- Random forest models leveraging demographic variables + cluster assignments
- Random forest models leveraging demographic variables + PCA factors
- SVM models leveraging demographic variables + cluster assignments
- SVM models leveraging demographic variables + PCA factors

Initial takeaways

- Demographic variables tend to rise to the top looking at feature importance (RF) + coefficients (SVM)
- ...however we can't isolate the effect controlling for other variables
    - In other words, someone who is poor probably also consumes a poor diet - what is the relative contribution of the diet?
- We plan to run regression analysis to hone in on this

SVM predicting high blood pressure (accuracy = .732)



Sample output

Caveats and thoughts on future work

- Asking individuals what they have consumed may not be most accurate gauge of intake
- Inherent limitations to a dataset that tracks participants over only two days
- However, with the right data, leveraging nutritional clusters could be a useful input in supervised settings

Appendix

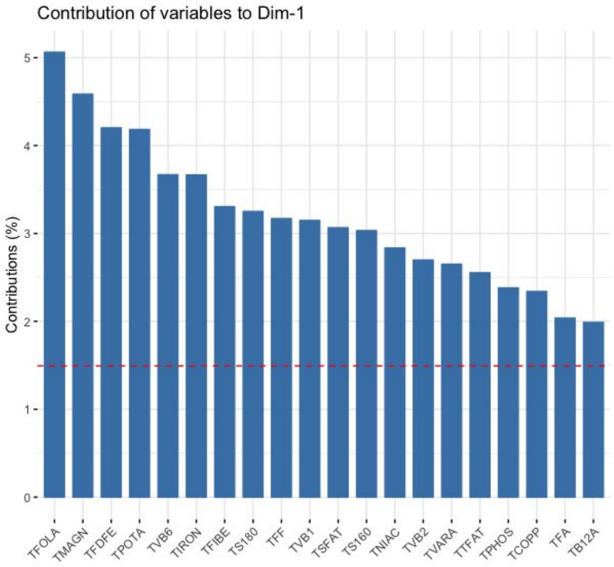
Full list of nutrients included in clustering

Alpha-carotene (mcg)
Alcohol (gm)
Added alpha-tocopherol (Vitamin E) (mg)
Vitamin E as alpha-tocopherol (mg)
Added vitamin B12 (mcg)
Beta-carotene (mcg)
Caffeine (mg)
Calcium (mg)
Carbohydrate (gm)
Total choline (mg)
Cholesterol (mg)
Copper (mg)
Beta-cryptoxanthin (mcg)
Folic acid (mcg)
Folate as dietary folate equivalents (mcg)
Food folate (mcg)
Dietary fiber (gm)
Total Folate (mcg)
Iron (mg)
Energy (kcal)
Lycopene (mcg
Lutein + zeaxanthin (mcg)
MFA 16:1 (Hexadecenoic) (gm))

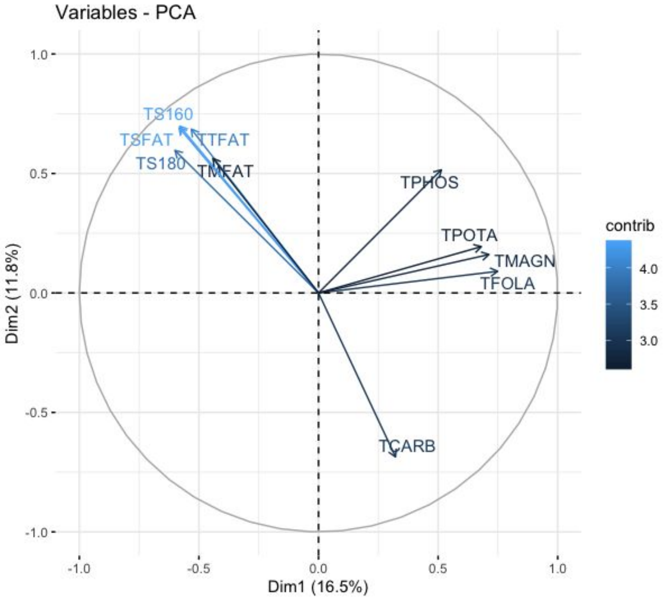
MFA 18:1 (Octadecenoic) (gm)
MFA 20:1 (Eicosenoic) (gm)
MFA 22:1 (Docosenoic) (gm)
Magnesium (mg)
Total monounsaturated fatty acids (gm)
Moisture (gm)
Niacin (mg)
PFA 18:2 (Octadecadienoic) (gm)
PFA 18:3 (Octadecatrienoic) (gm)
PFA 18:4 (Octadecatetraenoic) (gm
Niacin (mg))
PFA 18:2 (Octadecadienoic) (gm)
PFA 18:3 (Octadecatrienoic) (gm)
PFA 18:4 (Octadecatetraenoic) (gm)
PFA 20:4 (Eicosatetraenoic) (gm)
PFA 20:5 (Eicosapentaenoic) (gm)
PFA 22:5 (Docosapentaenoic) (gm)
PFA 22:6 (Docosahexaenoic) (gm)
Total polyunsaturated fatty acids (gm
Phosphorus (mg)
Potassium (mg)
Protein (gm)
Retinol (mcg))

SFA 4:0 (Butanoic) (gm)
SFA 6:0 (Hexanoic) (gm)
SFA 8:0 (Octanoic) (gm)
SFA 10:0 (Decanoic) (gm)
SFA 12:0 (Dodecanoic) (gm)
SFA 14:0 (Tetradecanoic) (gm)
SFA 16:0 (Hexadecanoic) (gm)
SFA 18:0 (Octadecanoic) (gm)
Selenium (mcg)
Total saturated fatty acids (gm)
Sodium (mg)
Total sugars (gm)
Total fat (gm)
Theobromine (mg)
Vitamin A as retinol activity equivalents (mcg)
Thiamin (Vitamin B1) (mg)
Vitamin B12 (mcg)
Riboflavin (Vitamin B2) (mg)
Vitamin B6 (mg)
Vitamin C (mg)
Vitamin D (D2 + D3) (mcg)
Vitamin K (mcg)
Zinc (mg)

Relative contribution of variables to first component



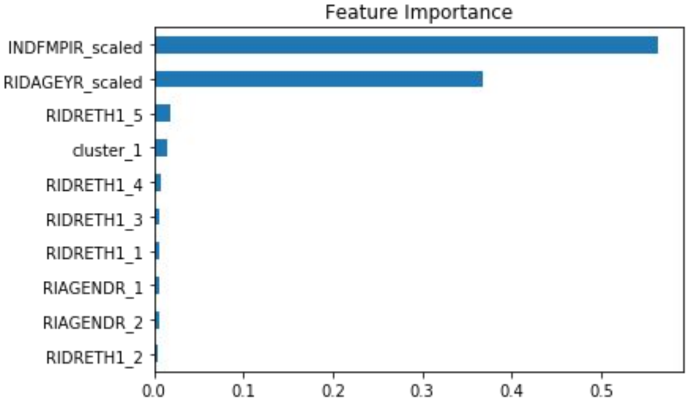
Biplot with top 10 contributing features



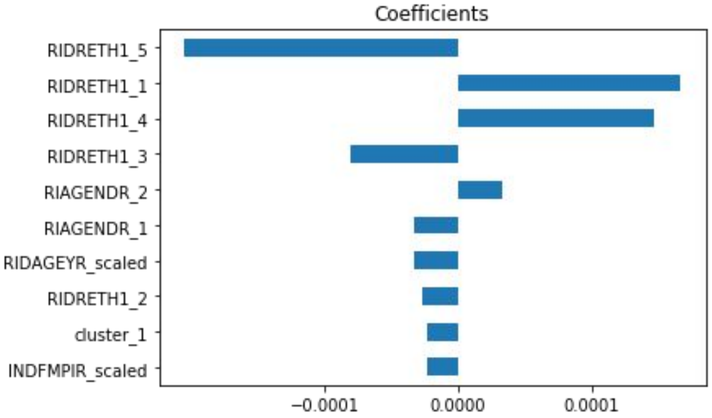
Results for obesity (I/II)

Leveraging cluster as a feature

RF (accuracy = .633)



SVM (accuracy = .634)

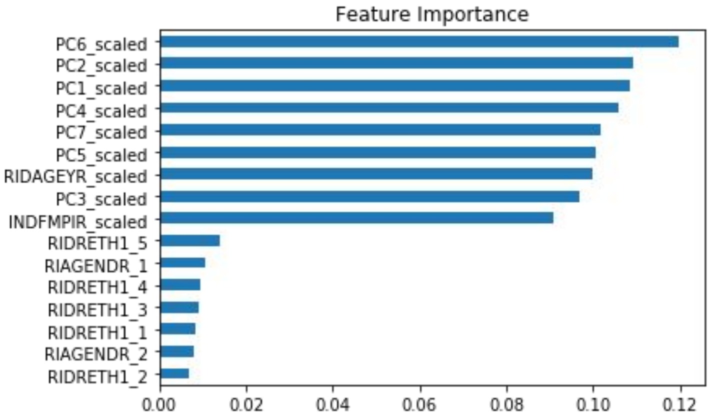


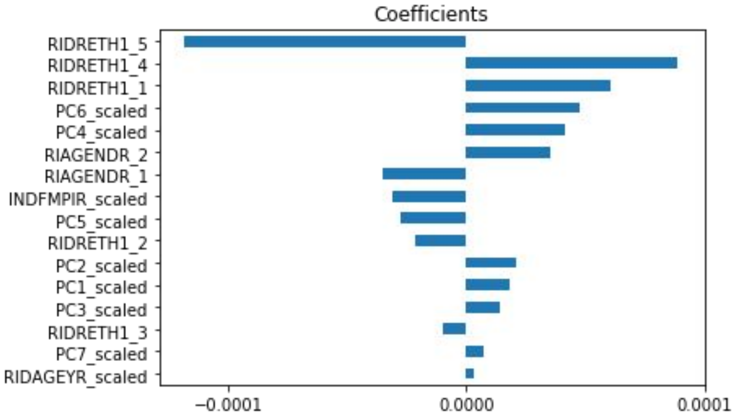
Results for obesity (II/II)

Leveraging components as features

RF (accuracy = .639

SVM (accuracy = .635)

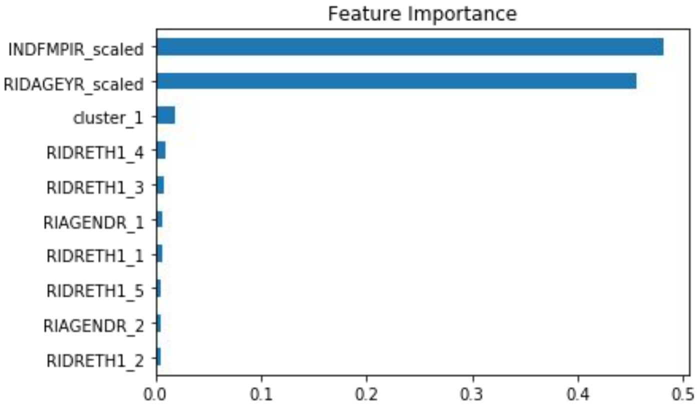




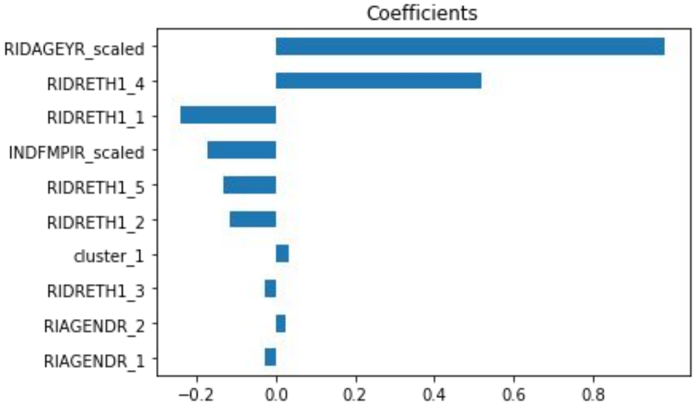
Results for blood pressure (I/II)

Leveraging cluster as a feature

RF (accuracy = .691)



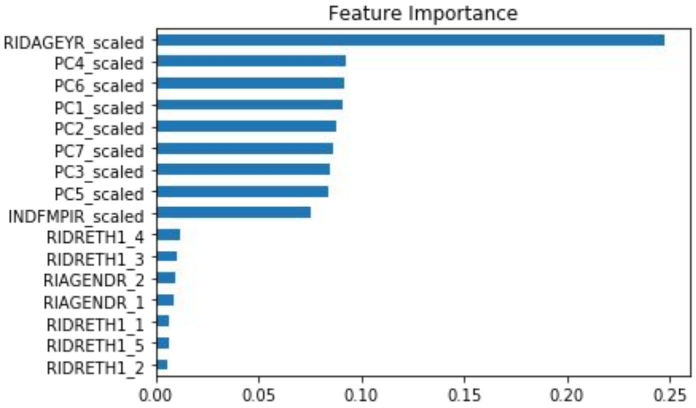
SVM (accuracy = .729)



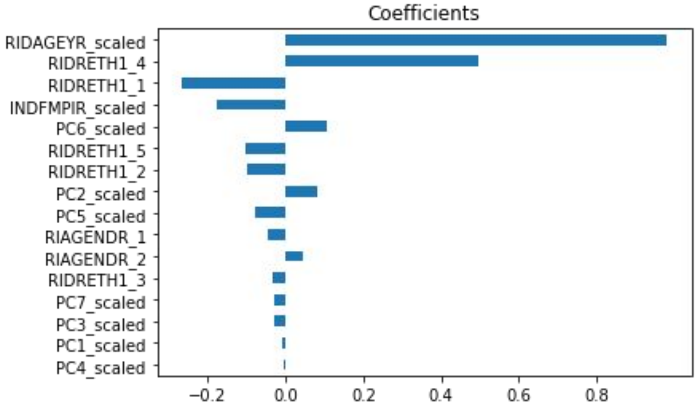
Results for blood pressure (II/II)

Leveraging components as features

RF (accuracy = .711)



SVM (accuracy = .732)


Coefficients
RIDAGEYR_scaled
RIDRETH1_4
RIDRETH1_1
INDFMPIR_scaled
PC6_scaled
RIDRETH1_5
RIDRETH1_2
PC2_scaled
PC5_scaled
RIAGENDR_1
RIAGENDR_2
RIDRETH1_3
PC7_scaled
PC3_scaled
PC1_scaled
PC4_scaled