VITAL-MINs: Health Supplement Intake Analysis

BIS 634 Final Presentation

Tom Shin

Yale SCHOOL OF PUBLIC HEALTH

The Question

Hypochondria → Got to take more vitamins! → Should I take them after meal or before meal? → May I take these simultaneously?

→ But wait, wouldn't it **damage my liver condition**?

The Data Source: FAIRness

A survey data aggregated by the program the **National Health and Nutrition Examination Survey (NHANES)** initiated by the **National Center for Health Statistics (NCHS)**, the part of the **Centers for Disease Control and Prevention (CDC)** and has the responsibility for producing vital and health statistics for the Nation.

Data User Agreement

Print

Warning Data Use Restrictions

Please Read Carefully Before Using NCHS Public Use Survey Data

The National Center for Health Statistics (NCHS), Centers for Disease Control and Prevention (CDC), conducts statistical and epidemiological activities under the authority granted by the Public Health Service Act (42 U.S.C. § 242k). NCHS survey data are protected by Federal confidentiality laws including Section 308(d) Public Health Service Act [42 U.S.C. 242m(d)] and the Confidential Information Protection and Statistical Efficiency Act or CIPSEA [Pub. L. No. 115-435, 132 Stat. 5529 § 302]. These confidentiality laws state the data collected by NCHS may be used only for statistical reporting and analysis. Any effort to determine the identity of individuals and establishments violates the assurances of confidentiality provided by federal law.

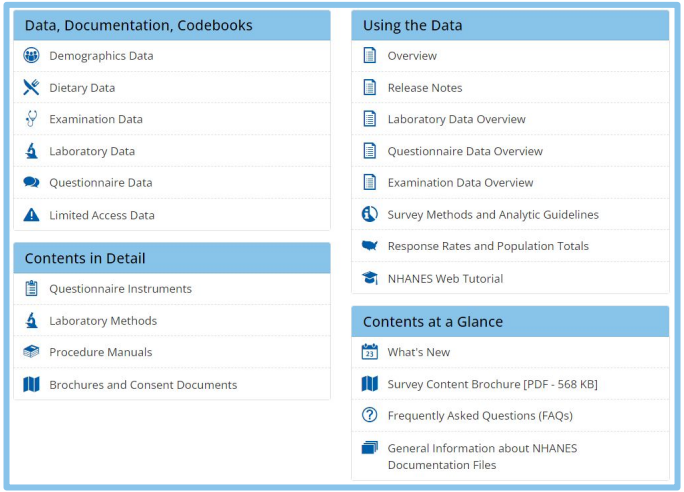
Terms and Conditions

NCHS does all it can to assure that the identity of individuals and establishments cannot be disclosed. All direct identifiers, as well as any characteristics that might lead to identification, are omitted from the dataset. Any intentional identification or disclosure of an individual or establishment violates the assurances of confidentiality given to the providers of the information. Therefore, users will:

1. Use the data in this dataset for statistical reporting and analysis only.
2. Make no attempt to learn the identity of any person or establishment included in these data.
3. Not link this dataset with individually identifiable data from other NCHS or non-NCHS datasets.
4. Not engage in any efforts to assess disclosure methodologies applied to protect individuals and establishments or any research on methods of re-identification of individuals and establishments.

By using these data you signify your agreement to comply with the above-stated statutorily based requirements.

Clearly stated license & policies



Data, Documentation, Codebooks

- Demographics Data
- Dietary Data
- Examination Data
- Laboratory Data
- Questionnaire Data
- Limited Access Data

Contents in Detail

- Questionnaire Instruments
- Laboratory Methods
- Procedure Manuals
- Brochures and Consent Documents

Using the Data

- Overview
- Release Notes
- Laboratory Data Overview
- Questionnaire Data Overview
- Examination Data Overview
- Survey Methods and Analytic Guidelines
- Response Rates and Population Totals
- NHANES Web Tutorial

Contents at a Glance

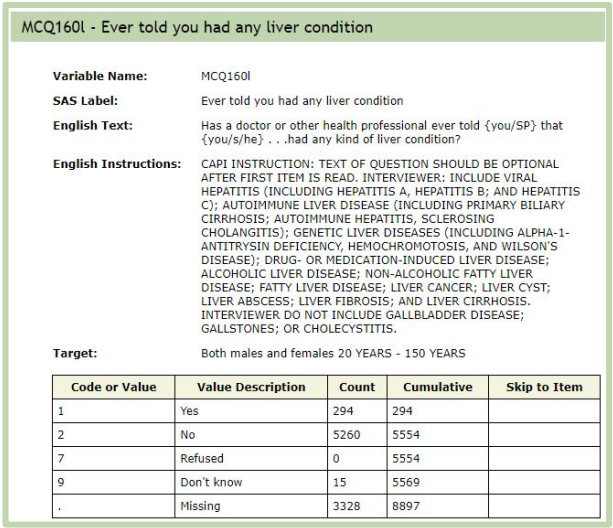
- What's New
- Survey Content Brochure [PDF - 568 KB]
- Frequently Asked Questions (FAQs)
- General Information about NHANES Documentation Files

Comprehensive usage information & guidelines

The Data Source: Acquisition

```python
import pandas as pd
food1 = pd.read_sas('/Users/tom/Desktop/DAIS_RETRY/foodday1', format='xport')
food2 = pd.read_sas('/Users/tom/Desktop/DAIS_RETRY/foodday2', format='xport')
body = pd.read_sas('/Users/tom/Desktop/DAIS_RETRY/body', format='xport')
demo = pd.read_sas('/Users/tom/Desktop/DAIS_RETRY/demo', format='xport')
glyco = pd.read_sas('/Users/tom/Desktop/DAIS_RETRY/glyco', format='xport')
medcond = pd.read_sas('/Users/tom/Desktop/DAIS_RETRY/medcond', format='xport')
trig = pd.read_sas('/Users/tom/Desktop/DAIS_RETRY/trig', format='xport')

food1.to_csv('/Users/tom/Desktop/DAIS_RETRY/foodday1.csv')
food2.to_csv('/Users/tom/Desktop/DAIS_RETRY/foodday2.csv')
body.to_csv('/Users/tom/Desktop/DAIS_RETRY/body.csv')
demo.to_csv('/Users/tom/Desktop/DAIS_RETRY/demo.csv')
glyco.to_csv('/Users/tom/Desktop/DAIS_RETRY/glyco.csv')
medcond.to_csv('/Users/tom/Desktop/DAIS_RETRY/medcond.csv')
trig.to_csv('/Users/tom/Desktop/DAIS_RETRY/trig.csv')
```

Code snippet for file conversion (.XPT → .CSV)



MCQ160l - Ever told you had any liver condition

Variable Name: MCQ160l

SAS Label: Ever told you had any liver condition

English Text: Has a doctor or other health professional ever told {you/SP} that {you/s/he} . . .had any kind of liver condition?

English Instructions: CAPI INSTRUCTION: TEXT OF QUESTION SHOULD BE OPTIONAL AFTER FIRST ITEM IS READ. INTERVIEWER: INCLUDE VIRAL HEPATITIS (INCLUDING HEPATITIS A, HEPATITIS B; AND HEPATITIS C); AUTOIMMUNE LIVER DISEASE (INCLUDING PRIMARY BILIARY CIRRHOSIS; AUTOIMMUNE HEPATITIS, SCLEROSING CHOLANGITIS); GENETIC LIVER DISEASES (INCLUDING ALPHA-1-ANTITRYSIN DEFICIENCY, HEMOCHROMOTOSIS, AND WILSON'S DISEASE); DRUG- OR MEDICATION-INDUCED LIVER DISEASE; ALCOHOLIC LIVER DISEASE; NON-ALCOHOLIC FATTY LIVER DISEASE; FATTY LIVER DISEASE; LIVER CANCER; LIVER CYST; LIVER ABSCESS; LIVER FIBROSIS; AND LIVER CIRRHOSIS. INTERVIEWER DO NOT INCLUDE GALLBLADDER DISEASE; GALLSTONES; OR CHOLECYSTITIS.

Target: Both males and females 20 YEARS - 150 YEARS

Code or Value	Value Description	Count	Cumulative	Skip to Item
1	Yes	294	294	
2	No	5260	5554	
7	Refused	0	5554	
9	Don't know	15	5569	
.	Missing	3328	8897	

Exhaustive, easily followable data inventory

The Dataset: Lab Variables

- activity_PAQ.csv
- alcohol_ALQ.csv
- bloodpcholesterol_BPQ.csv
- bloodpressure_BPXO.csv
- bodymeasure_BMX.csv
- cardiohealth_CDQ.csv
- childhood_ECQ.csv
- cigarette_SMQ.csv
- currenthealth_HSQ.csv
- demo_DEMO.csv
- diabetes_DIQ.csv
- dietnutrition_DBQ.csv
- foodday1_DR1IFF.csv
- foodday2_DR2IFF.csv
- hepatitis_HEQ.csv
- income_INQ.csv
- kidneyurol_KIQ.csv
- liver_LUX.csv
- medcond_MCQ.csv
- oral_OHXDEN.csv
- oralhealth_OHQ.csv
- osteo_OSQ.csv
- prescription_RXQRX.csv
- prescriptioninfo_RXQDRUG.csv
- supp30day_DSQIDS.csv
- suppblend_DSBI.csv
- suppday1_DS1IDS.csv
- suppday2_DS2IDS.csv
- suppinfo_DSPI.csv
- suppingred_DSII.csv
- weightpast_WHQ.csv
- weightyouth_WHQMEC.csv

Risk Factor/Features	National Cholesterol Education Program, ATP-III	International Diabetes Federation	Joint statement of IDF, IAS, IASO
Abdominal obesity, (waist circumference)	>102 cm (males), >88 cm (females)	≥94 cm (males), ≥80 cm (females)(ethnic differences)	≥94 cm (males), ≥80 cm differences)
Lipoprotein level	TRG ≥150 mg/dL or treated for dyslipidemia	TRG ≥150 mg/dL or treated for dyslipidemia	TRG ≥150 mg/dL or treat
HDL level	HDL-Chol <40 mg/dL (males); <50 mg/dL (females)	HDL-Chol <40 mg/dL (males); <50 mg/dL (females)	HDL-Chol <40 mg/dL (ma (females)
Blood pressure	≥130/85 mmHg or treated for Htx	≥130/85 mmHg or treated for Htx	≥130/85 mmHg or treate
Fasting Glucose (FG)	≥110 mg/dL or treated for DM	≥100 mg/dL or treated for DM	≥100 mg/dL or treated fo
Note	3 of the above	Abdominal obesity + 2 of the above	3 of the above
Associated Risk	**Defining Factors**		
0	No abdominal adiposity and no other features of MetS		
1	Abdominal adiposity		
2	Abdominal adiposity +1 feature of MetS (i.e. atherogenic dyslipidemia, low HDL cholesterol and/or high TRG, hyperte hyperglycemia/glucose intolerance/diabetes)		
3	Abdominal adiposity + 2 features of MetS		
4	Abdominal adiposity + 3 features of MetS		

Definition of the metabolic syndrome, according to recent classifications

I. ldl_TRIGLY
LBXTR - Triglyceride (mg/dL)
LBDLDLN - LDL-Cholesterol, NIH equation 2 (mg/dL)

II. bodymeasure_BMX
BMXWAIST - Waist Circumference (cm)
BMIWT - Weight Comment
BMXBMI - Body Mass Index (kg/m**2)

III. Plasma Fasting Glucose (P_GLU)
LBXGLU - Fasting Glucose (mg/dL)

IV. Glycohemoglobin (P_GHB)
LBXGH - Glycohemoglobin (%)

The Dataset: New-trition Variables

NIH National Institutes of Health
Office of Dietary Supplements

Life Stage Group	Vitamin A (µg/d)[a]	Vitamin C (mg/d)	Vitamin D (µg/d)[b,c]	Vitamin E (mg/d)[d]	Vitamin K (µg/d)	Thiamin (mg/d)
Infants						
0–6 mo	400*	40*	10*	4*	2.0*	0.2*
6–12 mo	500*	50*	10*	5*	2.5*	0.3*
Children						
1–3 y	**300**	**15**	**15**	**6**	30*	**0.5**
4–8 y	**400**	**25**	**15**	**7**	55*	**0.6**
Males						
9–13 y	**600**	**45**	**15**	**11**	60*	**0.9**
14–18 y	**900**	**75**	**15**	**15**	75*	**1.2**
19–30 y	**900**	**90**	**15**	**15**	120*	**1.2**
31–50 y	**900**	**90**	**15**	**15**	120*	**1.2**
51–70 y	**900**	**90**	**15**	**15**	120*	**1.2**
> 70 y	**900**	**90**	**20**	**15**	120*	**1.2**

Recommended Dietary Allowances and Adequate Intakes, Vitamins

Life Stage Group	Vitamin A (µg/d)[a]	Vitamin C (mg/d)	Vitamin D (µg/d)	Vitamin E (mg/d)[b,c]	Vitamin K	Thiamin
Infants						
0–6 mo	600	ND[e]	25	ND	ND	ND
6–12 mo	600	ND	38	ND	ND	ND
Children						
1–3 y	600	400	63	200	ND	ND
4–8 y	900	650	75	300	ND	ND
Males						
9–13 y	1,700	1,200	100	600	ND	ND
14–18 y	2,800	1,800	100	800	ND	ND
19–30 y	3,000	2,000	100	1,000	ND	ND
31–50 y	3,000	2,000	100	1,000	ND	ND
51–70 y	3,000	2,000	100	1,000	ND	ND
> 70 y	3,000	2,000	100	1,000	ND	ND

Tolerable Upper Intake Levels, Vitamins

The Dataset: New Nutrition Variables

Life Stage Group	Vitamin A (µg/d)[a]	Vitamin C (mg/d)	Vitamin D (µg/d)[b,c]	Vitamin E (mg/d)[d]	Vitamin K (µg/d)	Thiamin (mg/d)
Infants						
0–6 mo	400*	40*	10*	4*	2.0*	0.2*
6–12 mo	500*	50*	10*	5*	2.5*	0.3*
Children						
1–3 y	**300**	**15**	**15**	**6**	30*	**0.5**
4–8 y	**400**	**25**	**15**	**7**	55*	**0.6**
Males						
9–13 y	**600**	**45**	**15**	**11**	60*	**0.9**
14–18 y	**900**	**75**	**15**	**15**	75*	**1.2**
19–30 y	**900**	**90**	**15**	**15**	120*	**1.2**
31–50 y	**900**	**90**	**15**	**15**	120*	**1.2**
51–70 y	**900**	**90**	**15**	**15**	120*	**1.2**
> 70 y	**900**	**90**	**20**	**15**	120*	**1.2**

Recommended Dietary Allowances and Adequate Intakes, Vitamins

Life Stage Group	Vitamin A (µg/d)[a]	Vitamin C (mg/d)	Vitamin D (µg/d)	Vitamin E (mg/d)[b,c]	Vitamin K	Thiamin
Infants						
0–6 mo	600	ND[c]	25	ND	ND	ND
6–12 mo	600	ND	38	ND	ND	ND
Children						
1–3 y	600	400	63	200	ND	ND
4–8 y	900	650	75	300	ND	ND
Males						
9–13 y	1,700	1,200	100	600	ND	ND
14–18 y	2,800	1,800	100	800	ND	ND
19–30 y	3,000	2,000	100	1,000	ND	ND
31–50 y	3,000	2,000	100	1,000	ND	ND
51–70 y	3,000	2,000	100	1,000	ND	ND
> 70 y	3,000	2,000	100	1,000	ND	ND

Tolerable Upper Intake Levels, Vitamins

SEQN	RIAGENDR	RIDAGEYR	RIDEXPRG
109263	1	2	*NA*
109264	2	13	*NA*
109265	1	2	*NA*
109266	2	29	2
109269	1	2	*NA*
109270	2	11	*NA*

```
create_grp <- function(gender, age, pregnancy) {
  if (is.na(pregnancy) || pregnancy != 1) {
    pregnancy <- 2
  }
  if (age >= 1 & age <= 3) {
    return("C1_3")
  }

  if (gender == 1) {
    if (age >= 4 & age <= 8) {
      return("C4_8")
    } else if (age >= 9 & age <= 13) {
      return("M9_13")
```

SEQN	GRP	RIAGENDR	RIDAGEYR	RIDEXPRG
109263	C1_3	1	2	*NA*
109264	F9_13	2	13	*NA*
109265	C1_3	1	2	*NA*
109266	F19_30	2	29	2
109269	C1_3	1	2	*NA*
109270	F9_13	2	11	*NA*

The Dataset: New Nutrition Variables

Life Stage Group	Riboflavin (mg/d)	Niacin (mg/d)[e]
Infants		
0–6 mo	0.3*	2*
6–12 mo	0.4*	4*
Children		
1–3 y	**0.5**	**6**
4–8 y	**0.6**	**8**
Males		
9–13 y	**0.9**	**12**
14–18 y	**1.3**	**16**
19–30 y	**1.3**	**16**
31–50 y	**1.3**	**16**
51–70 y	**1.3**	**16**
> 70 y	**1.3**	**16**

Recommended Dietary Allowances and Adequate Intakes, Vitamins

Life Stage Group	Riboflavin	Niacin (mg/d)[c]	Vitamin B_6 (mg/d)
Infants			
0–6 mo	ND	ND	ND
6–12 mo	ND	ND	ND
Children			
1–3 y	ND	10	30
4–8 y	ND	15	40
Males			
9–13 y	ND	20	60
14–18 y	ND	30	80
19–30 y	ND	35	100
31–50 y	ND	35	100
51–70 y	ND	35	100
> 70 y	ND	35	100

Tolerable Upper Intake Levels, Vitamins

The Dataset: New Nutrition Variables

Life Stage Group	Riboflavin (mg/d)	Niacin (mg/d)[e]
Infants		
0–6 mo	0.3*	2*
6–12 mo	0.4*	4*
Children		
1–3 y	0.5	6
4–8 y	0.6	8
Males		
9–13 y	0.9	12
14–18 y	1.3	16
19–30 y	1.3	16
31–50 y	1.3	16
51–70 y	1.3	16
> 70 y	1.3	16

Recommended Dietary Allowances and Adequate Intakes, Vitamins

Life Stage Group	Riboflavin	Niacin (mg/d)[e]	Vitamin B_6 (mg/d)
Infants			
0–6 mo	ND	ND	ND
6–12 mo	ND	ND	ND
Children			
1–3 y	ND	10	30
4–8 y	ND	15	40
Males			
9–13 y	ND	20	60
14–18 y	ND	30	80
19–30 y	ND	35	100
31–50 y	ND	35	100
51–70 y	ND	35	100
> 70 y	ND	35	100

Tolerable Upper Intake Levels, Vitamins

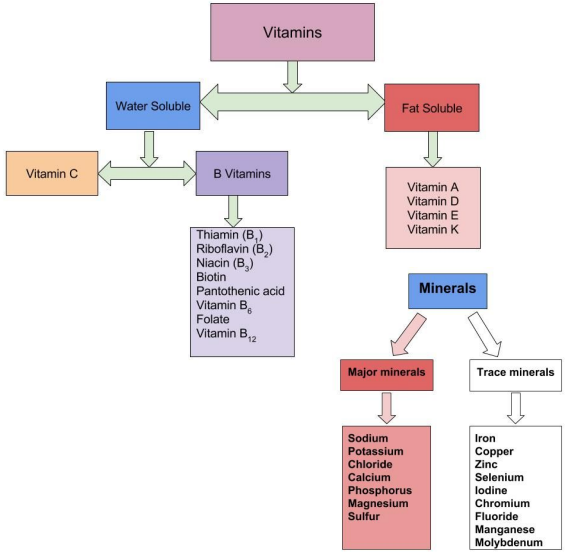
SEQN	VITAMIN A	VITAMIN B1	VITAMIN B2	VITAMIN B3
109263	304.5	1.70	1.01	15.44
109264	1122.5	2.20	1.88	36.15
109265	633.0	1.19	1.55	16.45
109266	420.0	1.29	1.65	12.53
109269	280.0	0.58	0.81	5.50
109270	648.0	1.75	2.34	19.31

SEQN	VITAMIN A	VITAMIN B1	VITAMIN B2	VITAMIN B3
109263	4.5	1.20	0.51	9.44
109264	522.5	1.30	0.98	24.15
109265	333.0	0.69	1.05	10.45
109266	-280.0	0.19	0.55	-1.47
109269	-20.0	0.08	0.31	-0.50
109270	48.0	0.85	1.44	7.31

→ -1

→ +1

The Dataset: New Nutrition Variables



The Dataset: Data Preprocessing

```
2.0    11777
1.0      642
9.0       30
Name: MCQ160L, dtype: int64
```

MCQ160l - Ever told you had any liver condition

Variable Name: MCQ160l

SAS Label: Ever told you had any liver condition

English Text: Has a doctor or other health professional ever told {you/SP} that {you/s/he} . . .had any kind of liver condition?

Target: Both males and females 20 YEARS - 150 YEARS

Code or Value	Value Description	Count	Cumulative	Skip to Item
1	Yes	294	294	
2	No	5260	5554	
7	Refused	0	5554	
9	Don't know	15	5569	
.	Missing	3328	8897	

```
df = df[df['MCQ160L'] != 9]
df['MCQ160L'] = df['MCQ160L'].replace({1: 1, 2: 0})
df = df.rename(columns={'MCQ160L': 'target'})
```

The Dataset: Missing Data

```
2.0    11777
1.0      642
9.0       30
Name: MCQ160L, dtype: int64
```

MCQ160l - Ever told you had any liver condition

Variable Nam
SAS Label:
English Text:
Target:

{you/SP} that

Code or Va	Skip to Item
1	
2	
7	
9	
.	

	Total	Percent
BMXWAIST	7864	63.32
LBDLDLN	6676	53.76
LBXTR	6661	53.64
LBXGLU	6554	52.77
LBXGH	532	4.28
BMXBMI	145	1.17
BMXWT	122	0.98
major_minerals_sum	0	0.00
target	0	0.00

```
df = df[df['MCQ160L'] != 9]
df['MCQ160L'] = df['MCQ160L'].replace({1: 1, 2: 0})
df = df.rename(columns={'MCQ160L': 'target'})
```

The Dataset: Final Dataset

The dataset, composed of 5569 data, is the health and nutritional status of adults in the U.S.

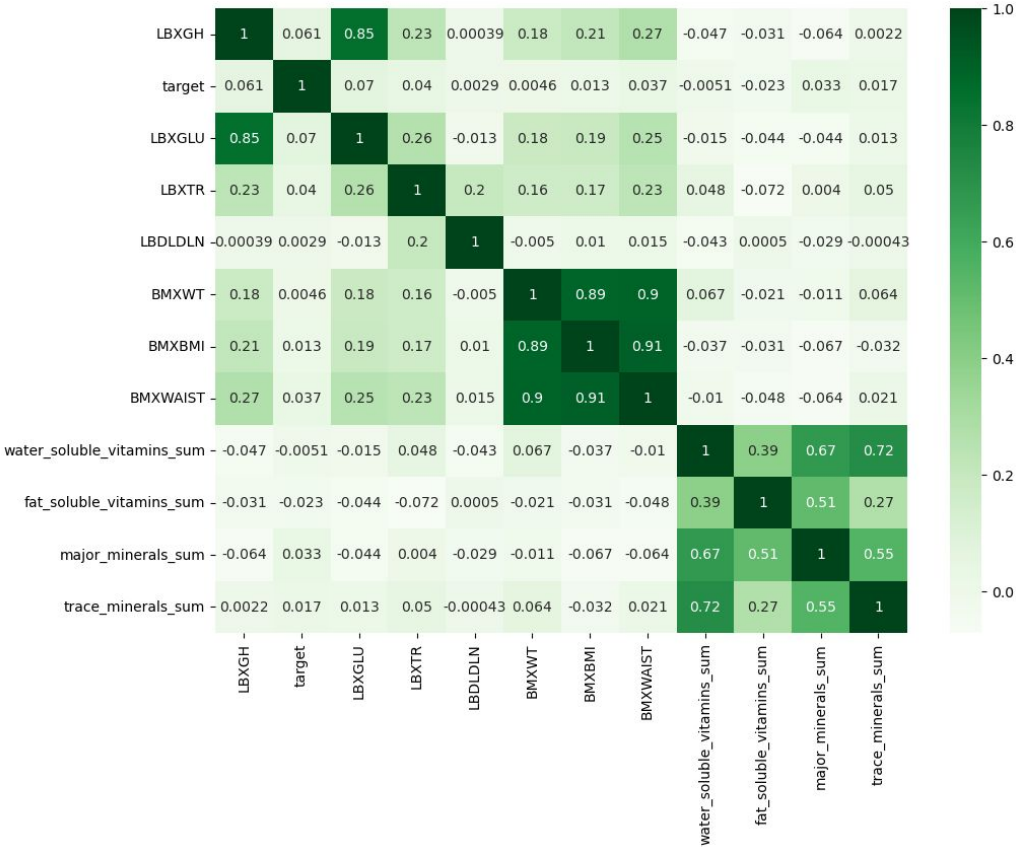
	LBXGH	target	LBXGLU	LBXTR	LBDLDLN	BMXWT	BMXBMI	BMXWAIST	water_soluble_vitamins_sum	fat_soluble_vitamins_sum	major_minerals_sum	trace_minerals_sum
0	6.2	0	122	58	111	53.5	23.7	88.2	-4	-2	-4	-4
1	5.7	0	107	48	158	62.1	21.3	86.6	3	-2	3	0
2	5.1	0	91	102	142	74.4	24.5	86.2	2	-2	-1	0
3	5.2	0	104	79	61	85.1	35.9	113.2	3	-3	2	0
4	5.8	0	101	54	96	56.8	23.8	89.7	2	-4	-1	-1

```
RangeIndex: 5569 entries, 0 to 5568
Data columns (total 12 columns):
 #   Column                     Non-Null Count  Dtype
---  ------                     --------------  -----
 0   LBXGH                      5569 non-null   float64
 1   target                     5569 non-null   int64
 2   LBXGLU                     5569 non-null   int64
 3   LBXTR                      5569 non-null   int64
 4   LBDLDLN                    5569 non-null   int64
 5   BMXWT                      5569 non-null   float64
 6   BMXBMI                     5569 non-null   float64
 7   BMXWAIST                   5569 non-null   float64
 8   water_soluble_vitamins_sum 5569 non-null   int64
 9   fat_soluble_vitamins_sum   5569 non-null   int64
 10  major_minerals_sum         5569 non-null   int64
 11  trace_minerals_sum         5569 non-null   int64
dtypes: float64(4), int64(8)
```

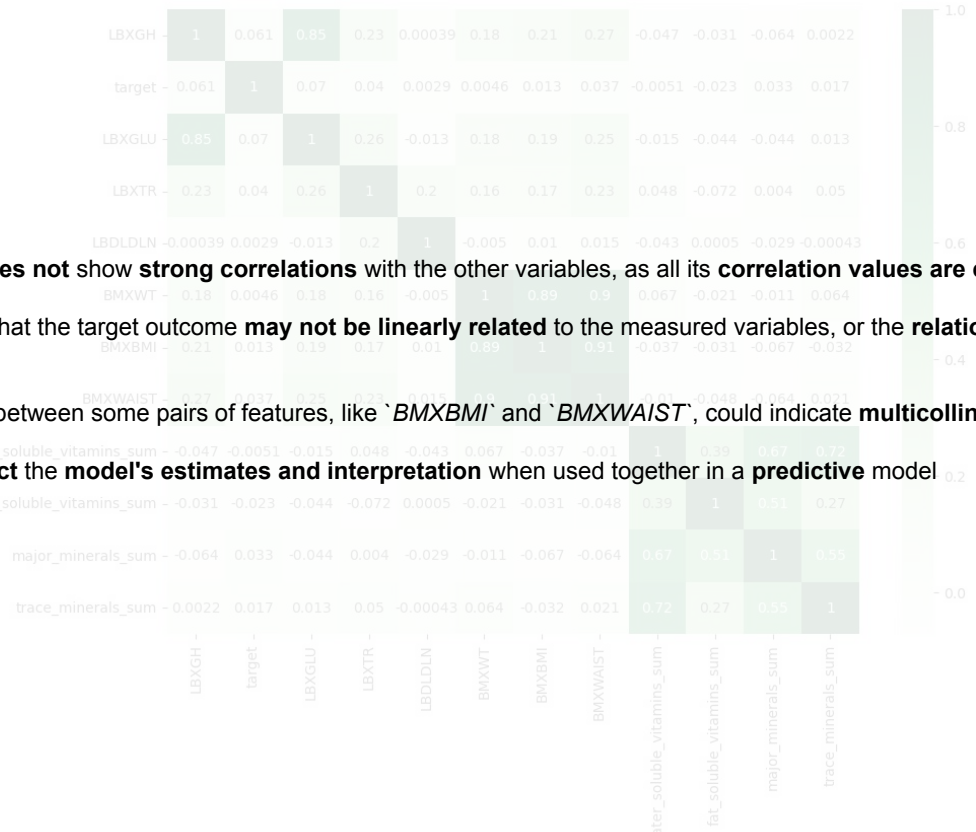
The Statistics: Overview

	LBXGH	target	LBXGLU	LBXTR	LBDLDLN	BMXWT	BMXBMI	BMXWAIST	water_soluble_vitamins_sum	fat_soluble_vitamins_sum	major_minerals_sum	trace_minerals_sum
count	5569.000000	5569.000000	5569.000000	5569.00000	5569.000000	5569.000000	5569.000000	5569.000000	5569.000000	5569.000000	5569.000000	5569.000000
mean	5.858880	0.052074	113.376369	108.63207	111.462561	83.493284	29.875795	100.892656	0.055665	-2.773029	-0.672473	-0.937870
std	1.122368	0.222196	37.076646	71.31120	36.493205	22.332643	7.310406	17.189165	3.210266	1.350494	2.133333	1.134499
min	2.800000	0.000000	47.000000	10.00000	14.000000	39.600000	15.400000	63.200000	-7.000000	-4.000000	-4.000000	-4.000000
25%	5.300000	0.000000	96.000000	61.00000	86.000000	67.700000	24.700000	88.700000	-2.000000	-4.000000	-2.000000	-1.000000
50%	5.600000	0.000000	104.000000	91.00000	108.000000	80.200000	28.700000	99.400000	1.000000	-3.000000	-1.000000	-1.000000
75%	6.000000	0.000000	115.000000	134.00000	133.000000	95.700000	33.700000	111.600000	3.000000	-2.000000	1.000000	0.000000
max	14.900000	1.000000	451.000000	780.00000	359.000000	210.800000	82.000000	178.000000	5.000000	2.000000	4.000000	1.000000

The Statistics: Correlation Plot



The Statistics: Correlation Plot



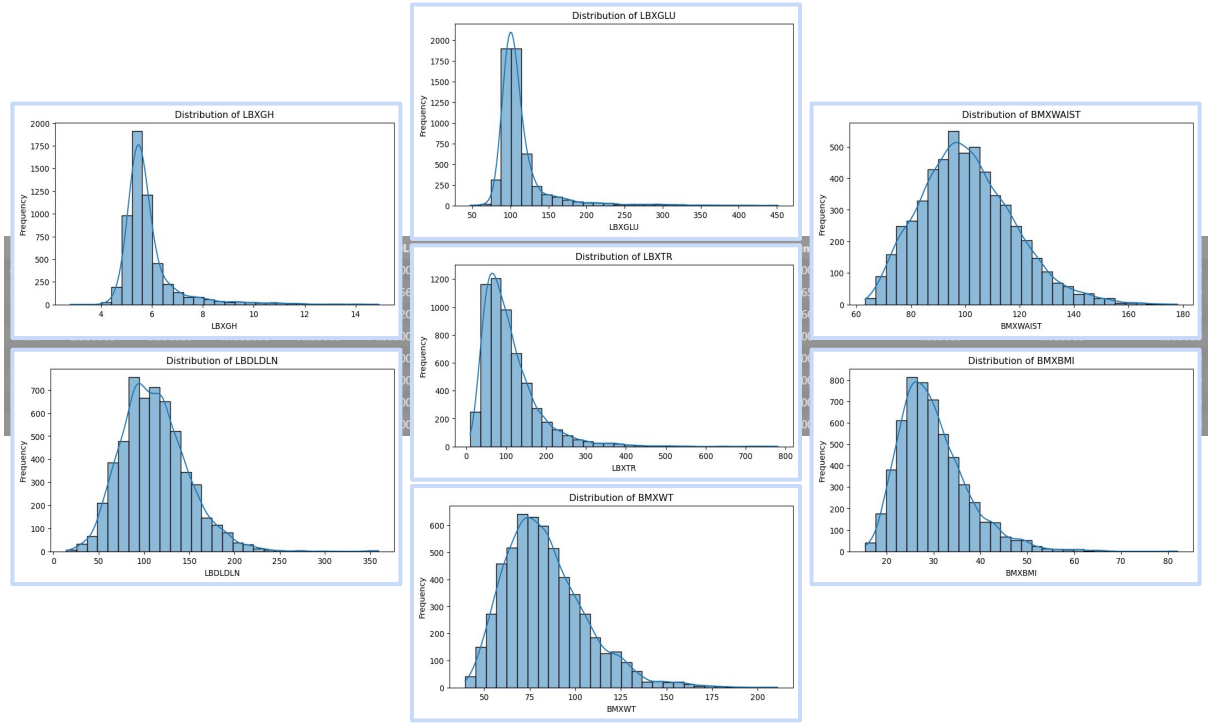
The **target** variable **does not** show **strong correlations** with the other variables, as all its **correlation values are close to zero**

→ Suggests that the target outcome **may not be linearly related** to the measured variables, or the **relationship is weak**

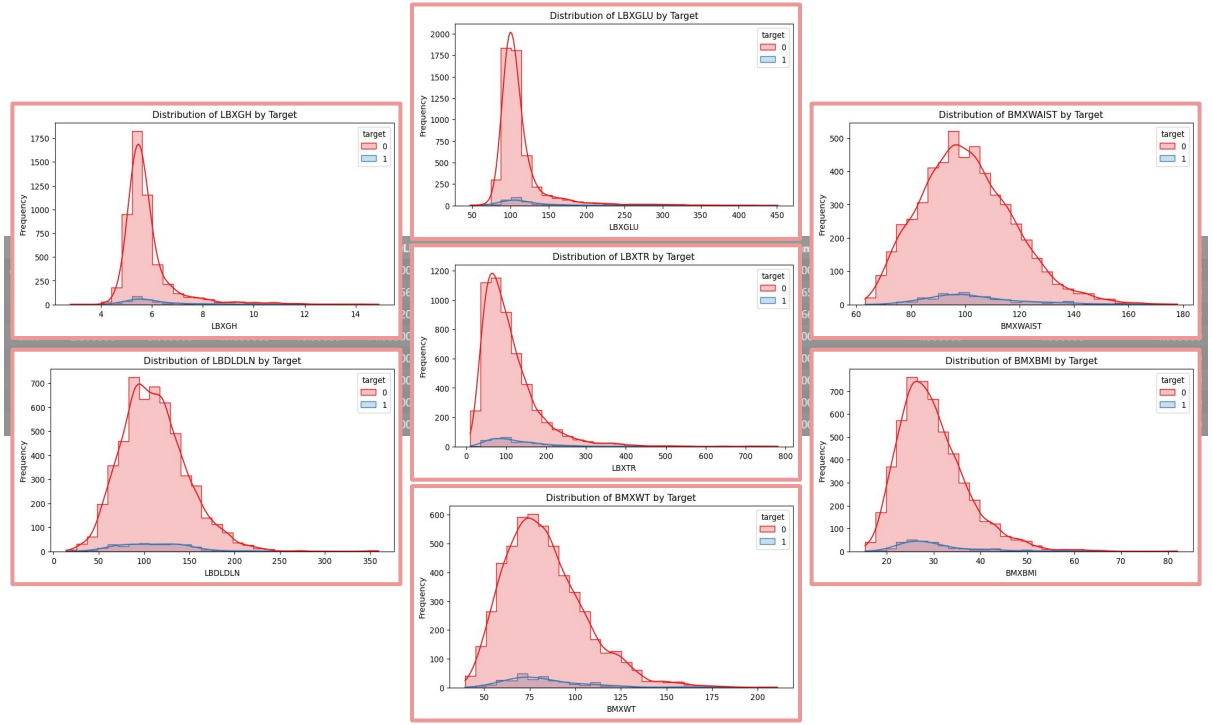
The **high correlation** between some pairs of features, like `*BMXBMI*` and `*BMXWAIST*`, could indicate **multicollinearity**

→ Could **affect** the **model's estimates and interpretation** when used together in a **predictive** model

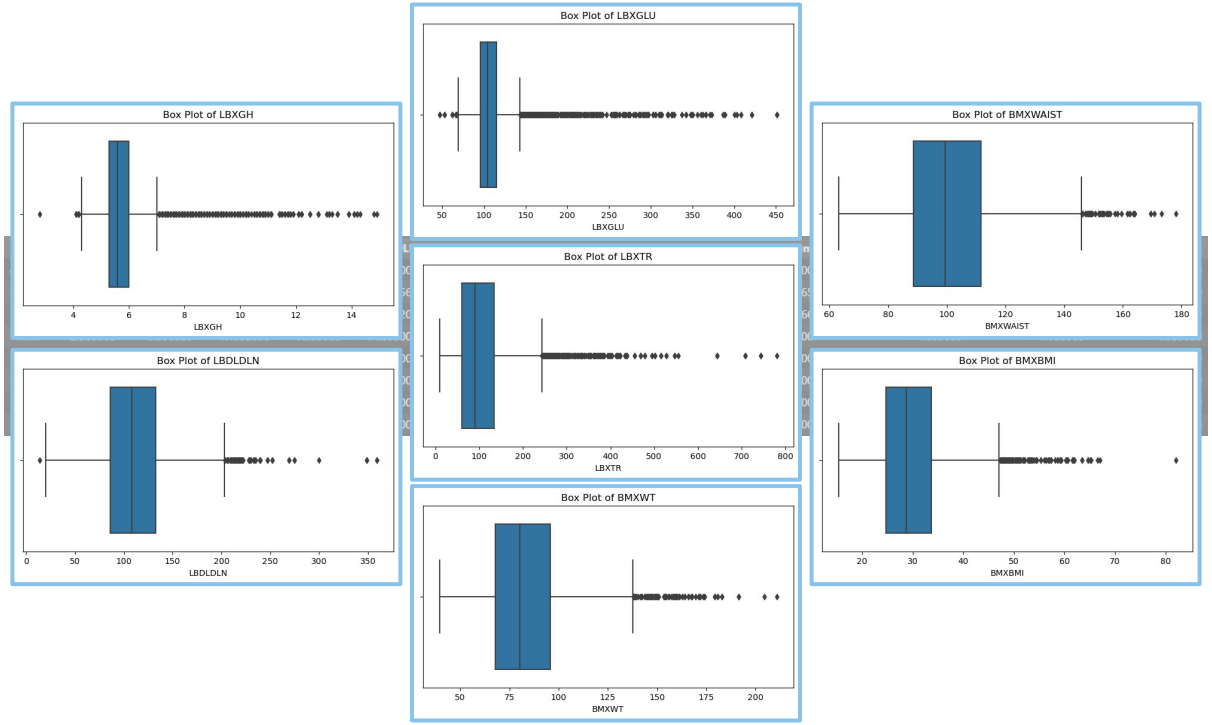
The Statistics: Lab Variables



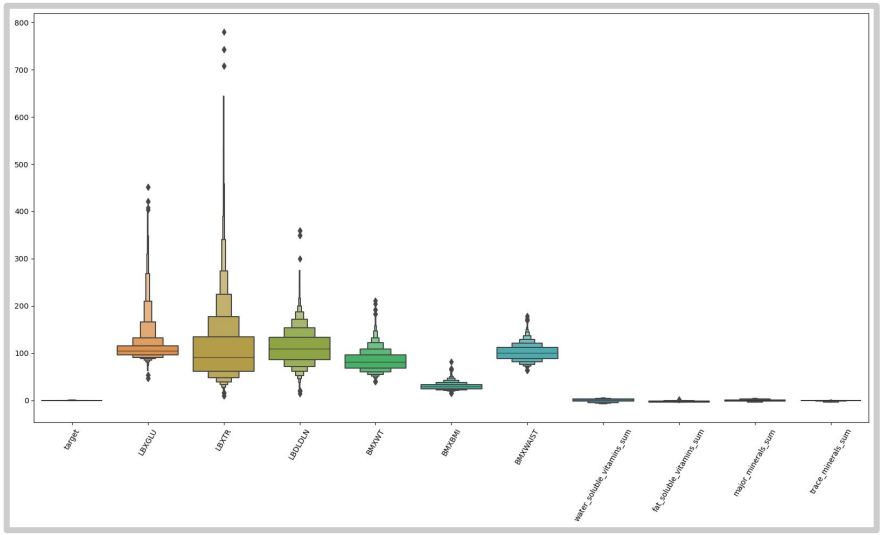
The Statistics: Lab Variables

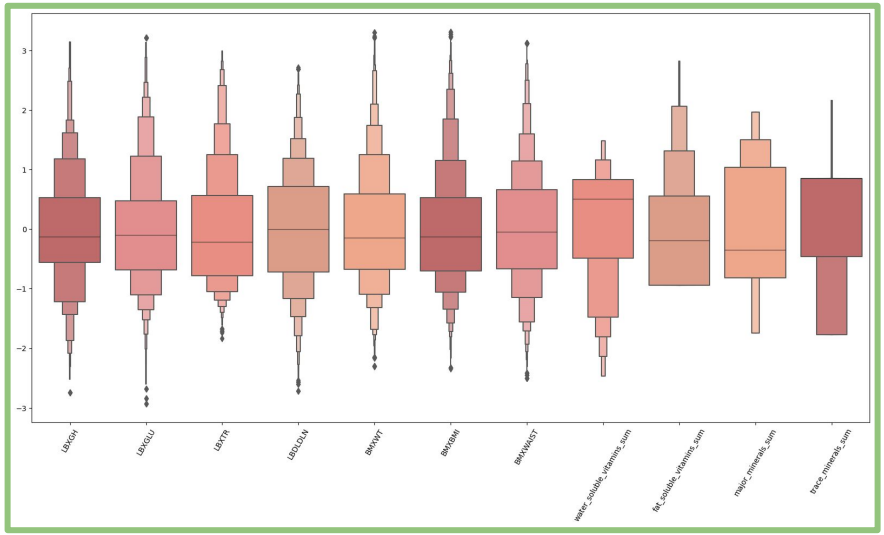


The Statistics: Lab Variables

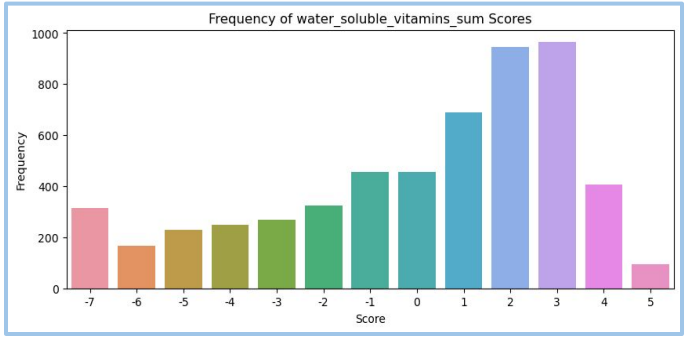


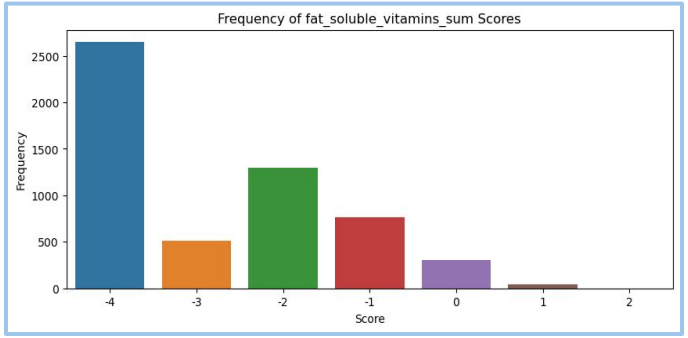
The Statistics: Lab Variables Outlier Removed & Scaled

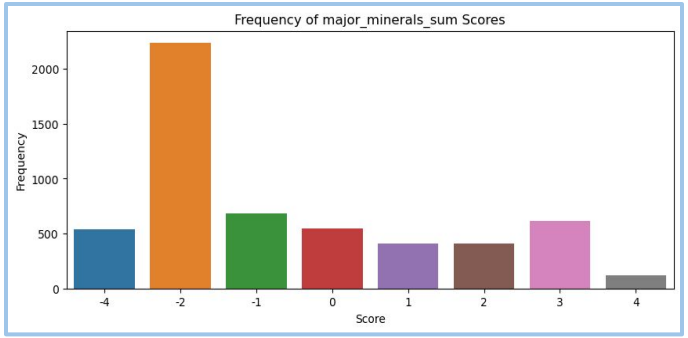


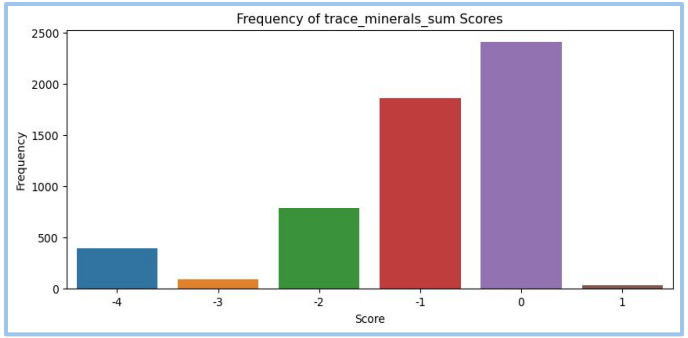


The Statistics: Nutrition Score Variables









The Analysis: ML Predictive Model Selection

Support Vector Classification

Interpretability: Low; it is hard to interpret the model, especially with non-linear kernels.
Flexibility: Very flexible, especially with non-linear kernels.
Training Speed: Between *O(n^2)* and *O(n^3)* where *n* is the number of samples.
Prediction Complexity: *O(n_support_vectors * n_features)*
Feature Scaling: Requires feature scaling for optimal performance

Logistic Regression

Interpretability: High; the contribution of each feature is quantifiable.
Flexibility: Less flexible than SVC, typically used for linearly separable data.
Training Speed: *O(n * m) to O(n * m^2)*, *n* = number of samples *m* = number of features
Prediction Complexity: *O(m)*
Feature Scaling: Benefits from feature scaling

Decision Tree

Interpretability: High; easy to understand and interpret, often represented visually
Flexibility: Very flexible, can model non-linear relationships well
Training Speed: *O(n * m * log(n))*, *n* = number of samples, and *m* = number of features
Prediction Complexity: *O(log(n))* if the tree is balanced.
Feature Scaling: This does not require feature scaling.

Random Forest

Interpretability: Medium; the ensemble nature makes it less interpretable.
Flexibility: High; capable of capturing complex, non-linear relationships
Training Speed: *O(n * m * log(m) * t)*, training can be parallelized over multiple CPUs.
Prediction Complexity: *O(t * log(m))* for a single prediction, assuming balanced trees.
Feature Scaling: Not required

The Analysis: ML Predictive Model Selection

Support Vector Classification

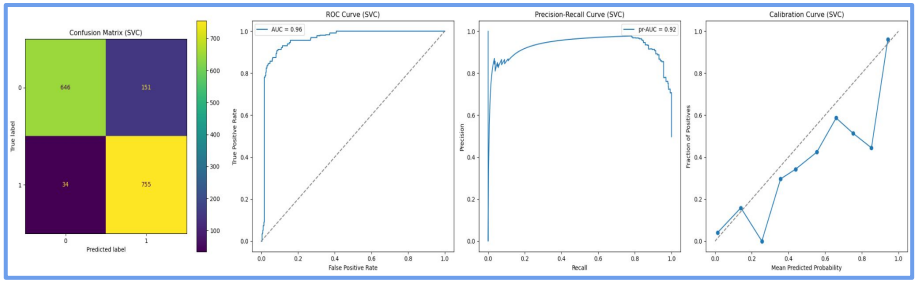
```
Cross-Validation Accuracy Scores for SVC: [0.86997636 0.86682427
Average CV Accuracy for SVC: 0.8830068137806911
SVC Accuracy: 0.8833543505674654
              precision    recall  f1-score   support

           0       0.95      0.81      0.87       797
           1       0.83      0.96      0.89       789

    accuracy                           0.88      1586
   macro avg       0.89      0.88      0.88      1586
weighted avg       0.89      0.88      0.88      1586
```

Random Forest

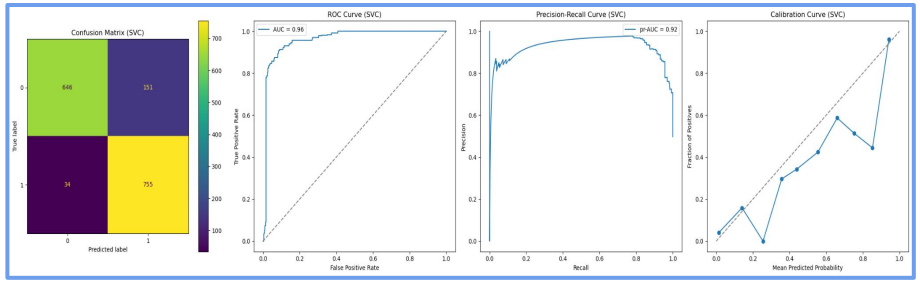
```
Cross-Validation Accuracy Scores for Decision Tree: [0.98423956
Average CV Accuracy for Decision Tree: 0.9831286216077142
Decision Tree Accuracy: 0.9848675914249685
              precision    recall  f1-score   support

           0       1.00      0.97      0.98       797
           1       0.97      1.00      0.99       789

    accuracy                           0.98      1586
   macro avg       0.99      0.98      0.98      1586
weighted avg       0.99      0.98      0.98      1586
```

The Analysis: Results

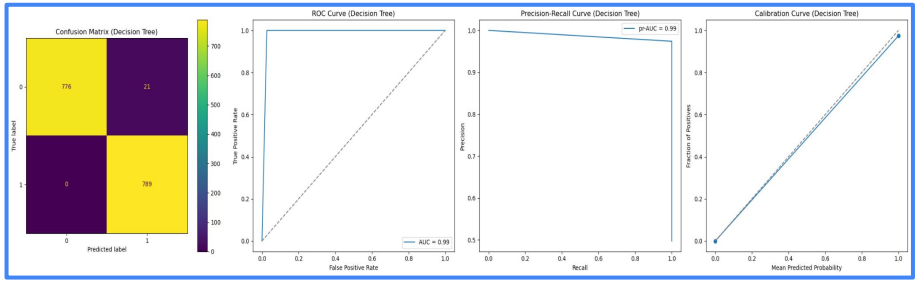
Support Vector Classification



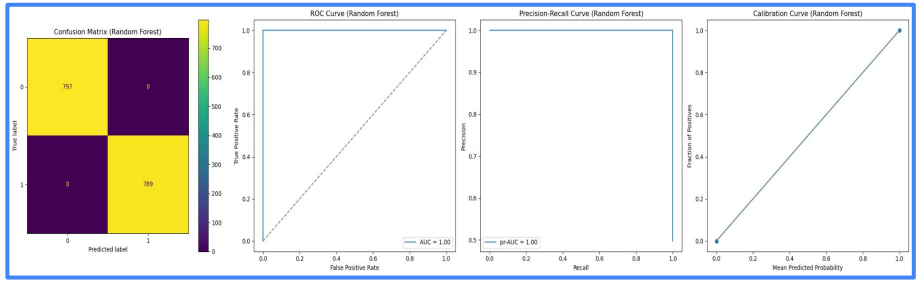
Logistic Regression



Decision Tree

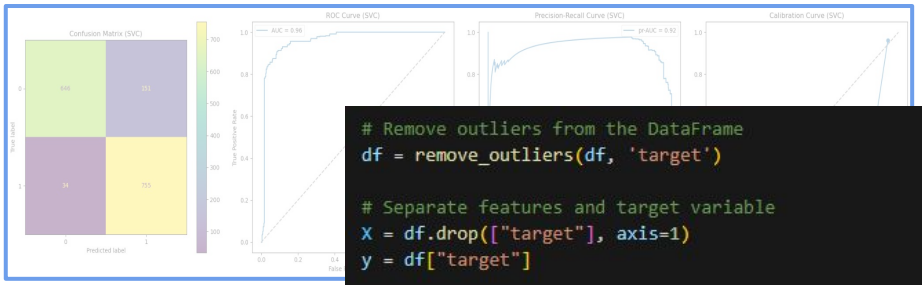


Random Forest

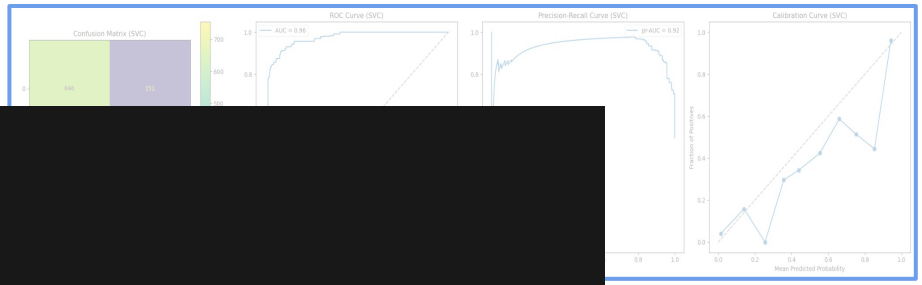


The Analysis: Results

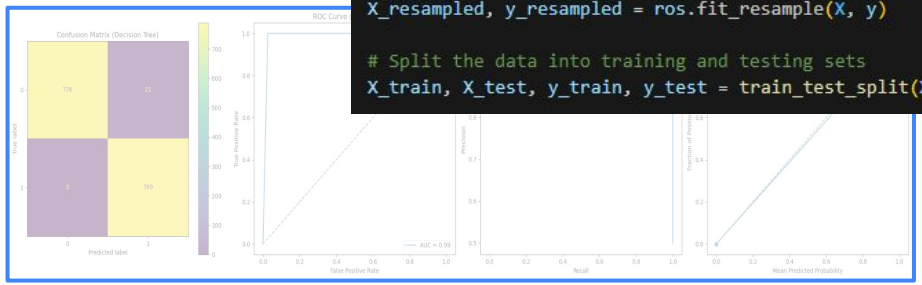
Support Vector Classification

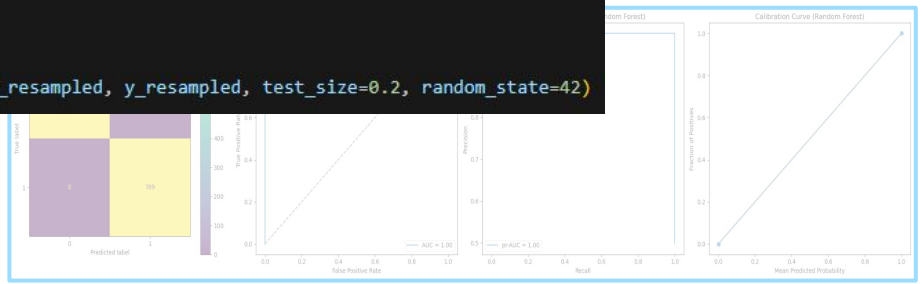


Logistic Regression



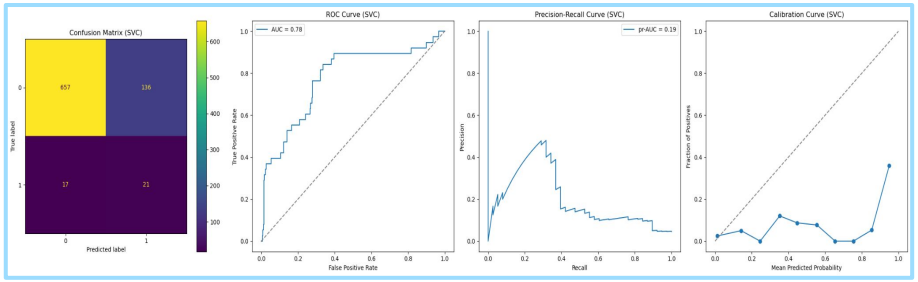
```
# Remove outliers from the DataFrame
df = remove_outliers(df, 'target')

# Separate features and target variable
X = df.drop(["target"], axis=1)
y = df["target"]

# Create an instance of the RandomOverSampler class
ros = RandomOverSampler(random_state=0)

# Resample the data to balance the classes
X_resampled, y_resampled = ros.fit_resample(X, y)

# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X_resampled, y_resampled, test_size=0.2, random_state=42)
```



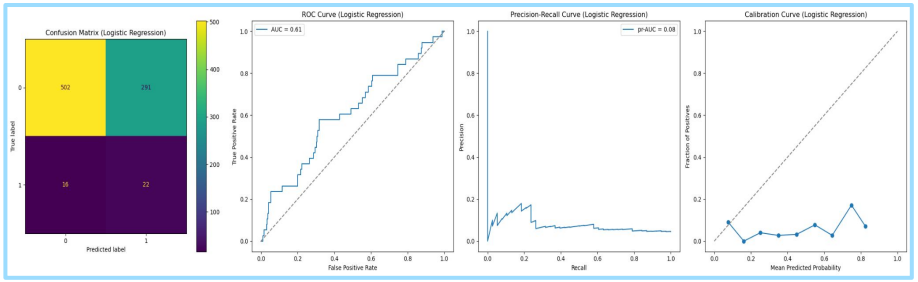


The Analysis: Results After Re-ordering

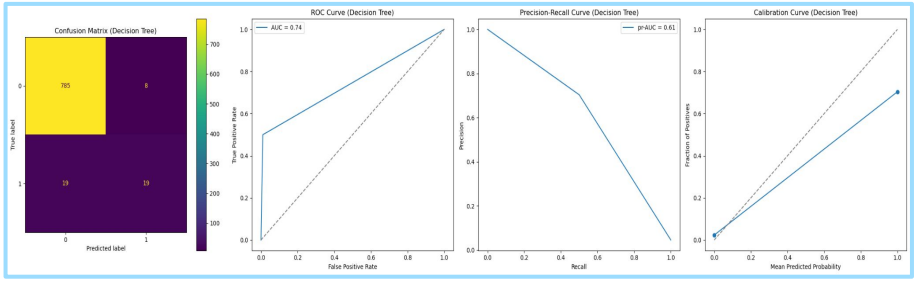
Support Vector Classification



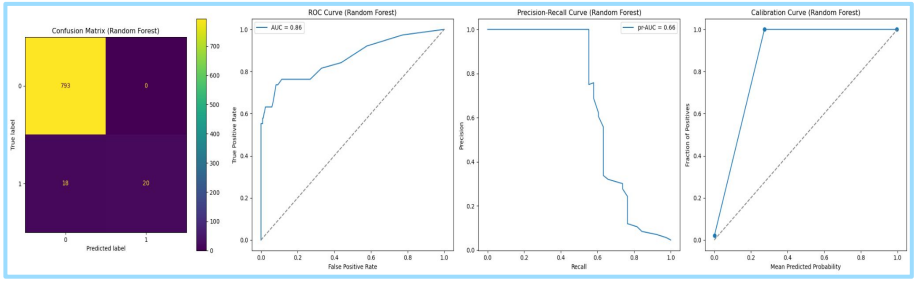
Logistic Regression



Decision Tree

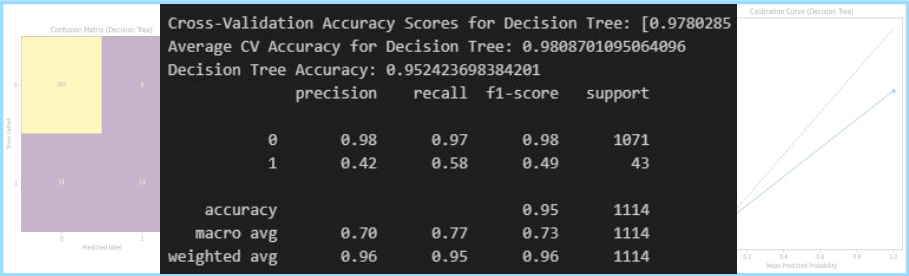


Random Forest



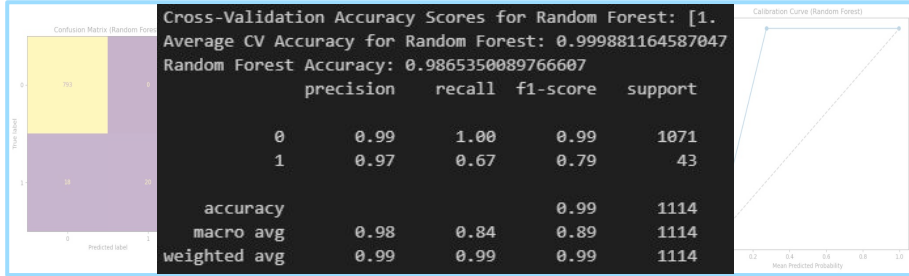
The Analysis: Results After Re-ordering

Decision Tree



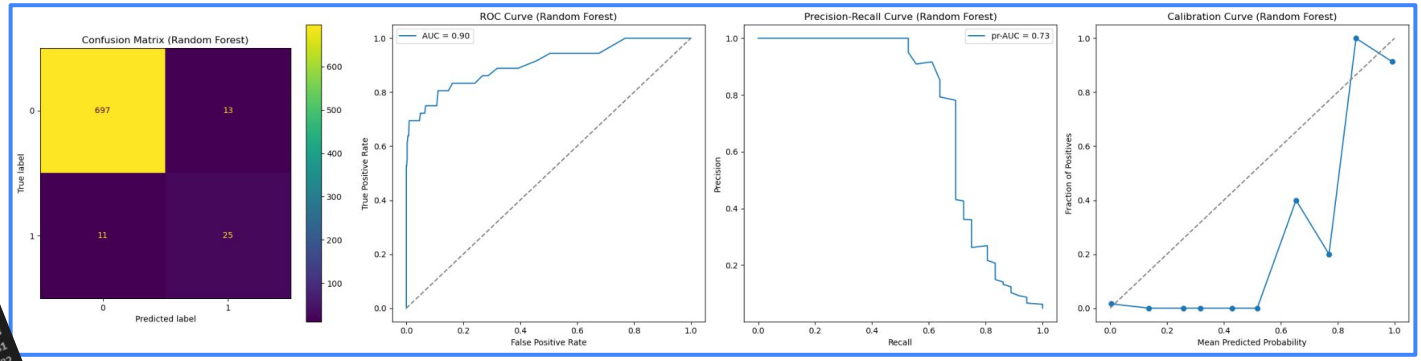

```
Cross-Validation Accuracy Scores for Decision Tree: [0.9780285
Average CV Accuracy for Decision Tree: 0.9808701095064096
Decision Tree Accuracy: 0.952423698384201
              precision    recall  f1-score   support

           0       0.98      0.97      0.98      1071
           1       0.42      0.58      0.49        43

    accuracy                           0.95      1114
   macro avg       0.70      0.77      0.73      1114
weighted avg       0.96      0.95      0.96      1114
```

Random Forest



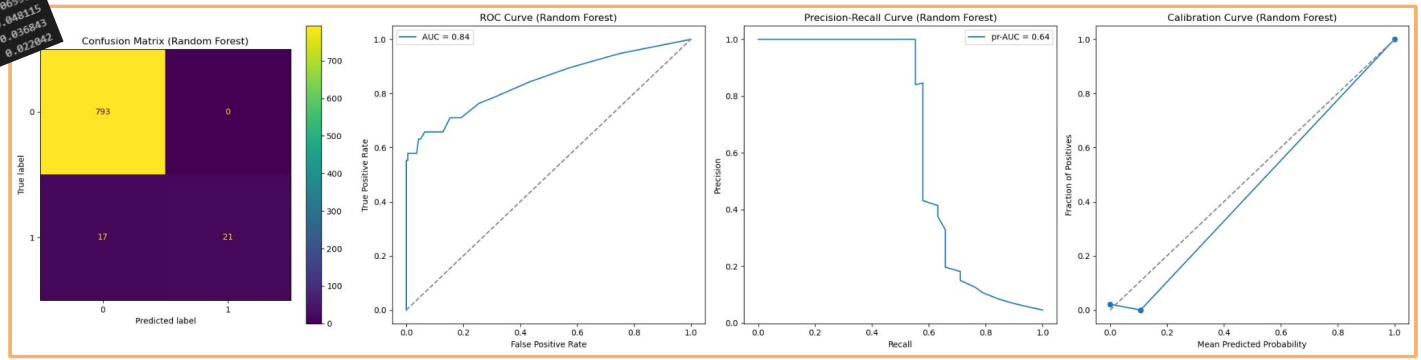

```
Cross-Validation Accuracy Scores for Random Forest: [1.
Average CV Accuracy for Random Forest: 0.999881164587047
Random Forest Accuracy: 0.9865350089766607
              precision    recall  f1-score   support

           0       0.99      1.00      0.99      1071
           1       0.97      0.67      0.79        43

    accuracy                           0.99      1114
   macro avg       0.98      0.84      0.89      1114
weighted avg       0.99      0.99      0.99      1114
```

The Analysis: Model



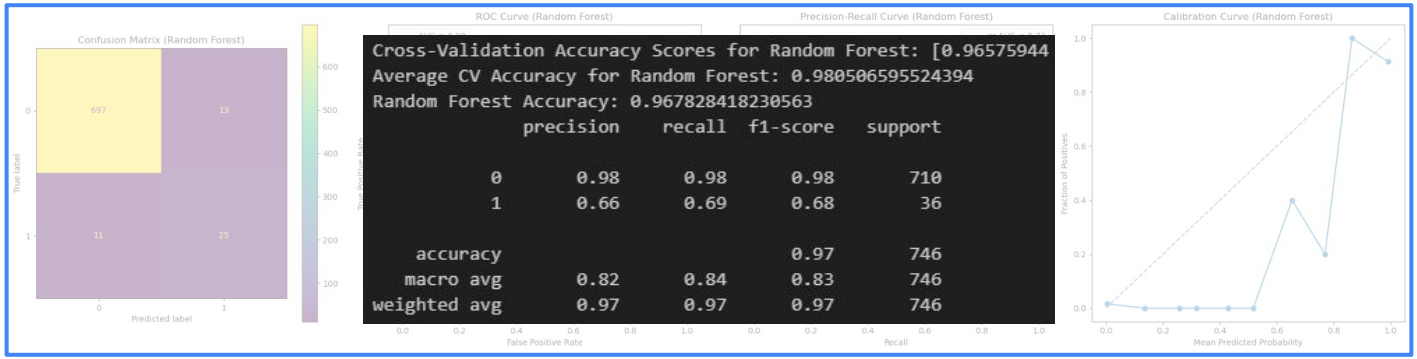
Random Forest ***with*** SMOTE

feature	importance
BMXWAIST	0.142999
BMXWT	0.133858
LBXTR	0.128441
BMXBMI	0.118782
LBDLDLN	0.113361
LBXGLU	0.104969
LBXGH	0.084685
water_soluble_vitamins_sum	0.065900
major_minerals_sum	0.048115
fat_soluble_vitamins_sum	0.036843
trace_minerals_sum	0.022042



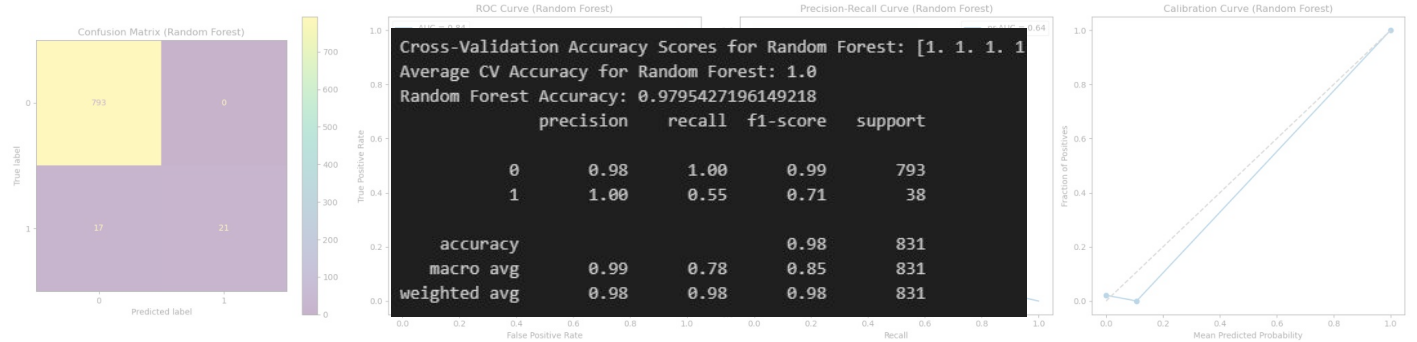
Random Forest ***with*** RandomOverSampler()
without `trace_minerals_sum`

The Analysis: Model




```
Cross-Validation Accuracy Scores for Random Forest: [0.96575944
Average CV Accuracy for Random Forest: 0.980506595524394
Random Forest Accuracy: 0.967828418230563
              precision    recall  f1-score   support

           0       0.98      0.98      0.98       710
           1       0.66      0.69      0.68        36

    accuracy                           0.97       746
   macro avg       0.82      0.84      0.83       746
weighted avg       0.97      0.97      0.97       746
```

Random Forest ***with*** SMOTE

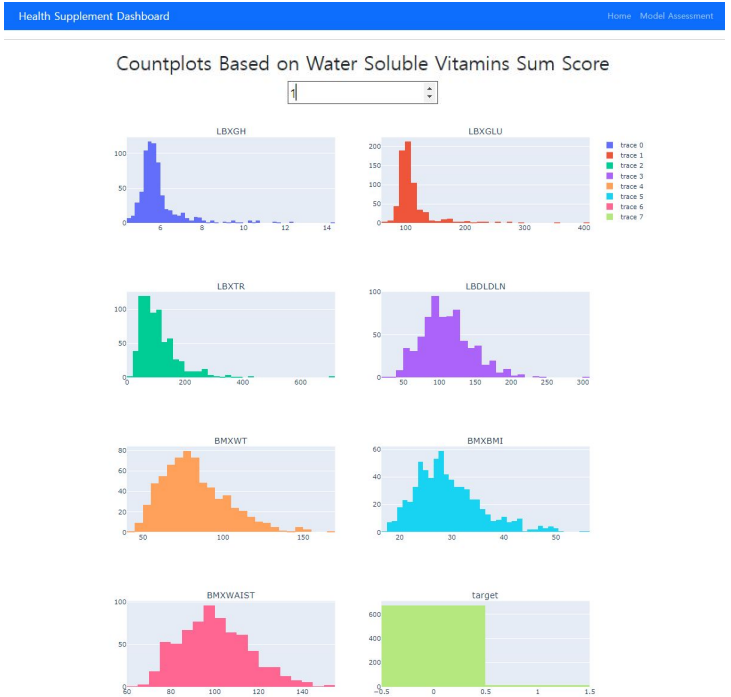



```
Cross-Validation Accuracy Scores for Random Forest: [1. 1. 1. 1
Average CV Accuracy for Random Forest: 1.0
Random Forest Accuracy: 0.9795427196149218
              precision    recall  f1-score   support

           0       0.98      1.00      0.99       793
           1       1.00      0.55      0.71        38

    accuracy                           0.98       831
   macro avg       0.99      0.78      0.85       831
weighted avg       0.98      0.98      0.98       831
```

Random Forest ***with*** RandomOverSampler()
without `trace_minerals_sum`

The API


Health Supplement Dashboard
Home
Model Assessment
Countplots Based on Water Soluble Vitamins Sum Score
-5
LBXGH
LBXGLU
LBXTR
LBDLDLN
BMXWT
BMXBMI
BMXWAIST
target
trace 0
trace 1
trace 2
trace 3
trace 4
trace 5
trace 6
trace 7


Health Supplement Dashboard
Home
Model Assessment
Countplots Based on Water Soluble Vitamins Sum Score
1
LBXGH
LBXGLU
LBXTR
LBDLDLN
BMXWT
BMXBMI
BMXWAIST
target
trace 0
trace 1
trace 2
trace 3
trace 4
trace 5
trace 6
trace 7

The API


Model Assessment Plots
Model Assessment Plots
SVC - ROC
SVC - Precision-Recall
Logistic Regression - ROC
Logistic Regression - Precision-Recall
Decision Tree - ROC
Decision Tree - Precision-Recall
Random Forest - ROC
Random Forest - Precision-Recall

The Unexpected & Difficulties

- **More sample**
- Stronger examination and analysis of variables are necessary (network analyses, interaction terms, and more)
- Time complexity while loading up the web API
- Failure to implement multiple user interactable functionality
- Maybe user inputting their labs in the near future?

Thank you! Questions?

References

1. National Institutes of Health. (n.d.). DSLD API Guide. Retrieved from https://dsld.od.nih.gov/api-guide
2. National Institutes of Health. (n.d.). DSID Conversions. Retrieved from https://dsid.od.nih.gov/Conversions.php
3. LibreTexts. (n.d.). Minerals and Vitamins: a closer look. In CHE 301 Biochemistry. Brevard College. Retrieved from https://chem.libretexts.org/Courses/Brevard_College/CHE_301_Biochemistry/07%3A_Nutrition/7.02%3A_Minerals_and_Vitamins-_a_closer_look
4. Schmieder, R., Edwards, R., Tischler, G., & He, J. (2018). Quality and quantity of data recovered from massively parallel sequencing: Examples in Asparagales and Poaceae. Scientific Reports, 8(1), 316. https://doi.org/10.1038/s41598-018-20166-x
5. Office of Dietary Supplements - National Institutes of Health. (n.d.). Nutrient Recommendations. Retrieved from https://ods.od.nih.gov/HealthInformation/nutrientrecommendations.aspx