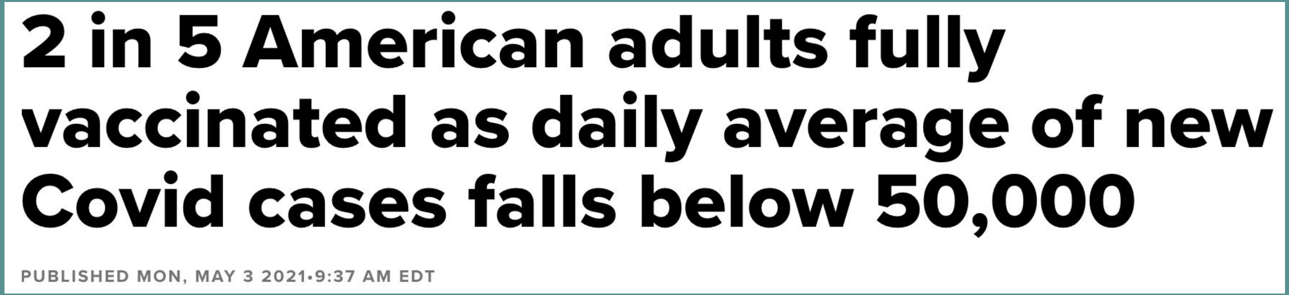
Comparison of Machine Learning Models to Predict COVID-19 Outcomes using Food Consumption Data

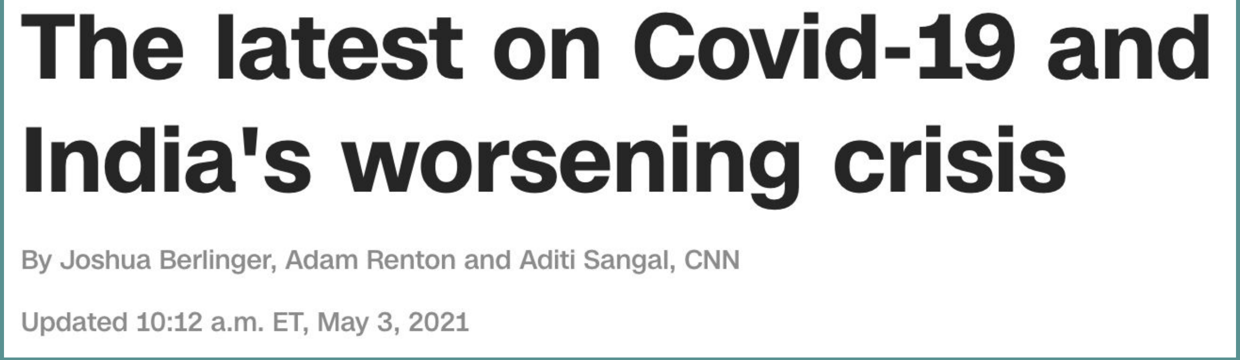
By: Alexis Payton, Eileen Yang, Zhitong Yu

May 4, 2021



2 in 5 American adults fully vaccinated as daily average of new Covid cases falls below 50,000

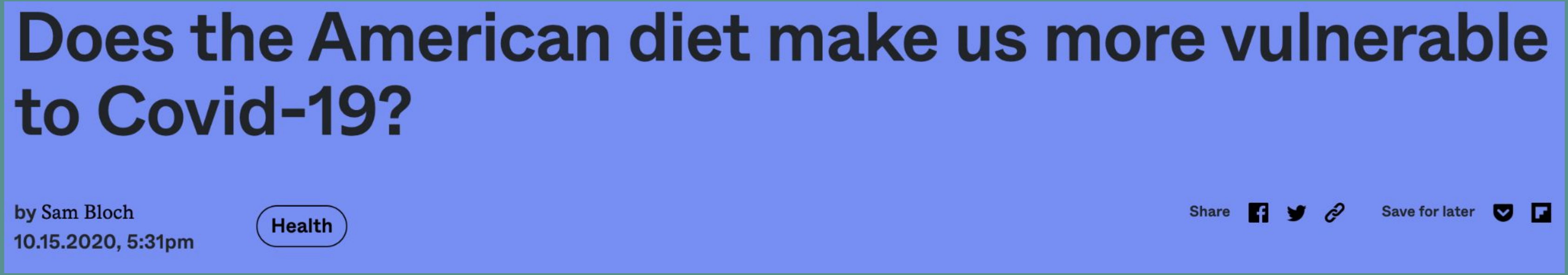
PUBLISHED MON, MAY 3 2021•9:37 AM EDT



The latest on Covid-19 and India's worsening crisis

By Joshua Berlinger, Adam Renton and Aditi Sangal, CNN

Updated 10:12 a.m. ET, May 3, 2021



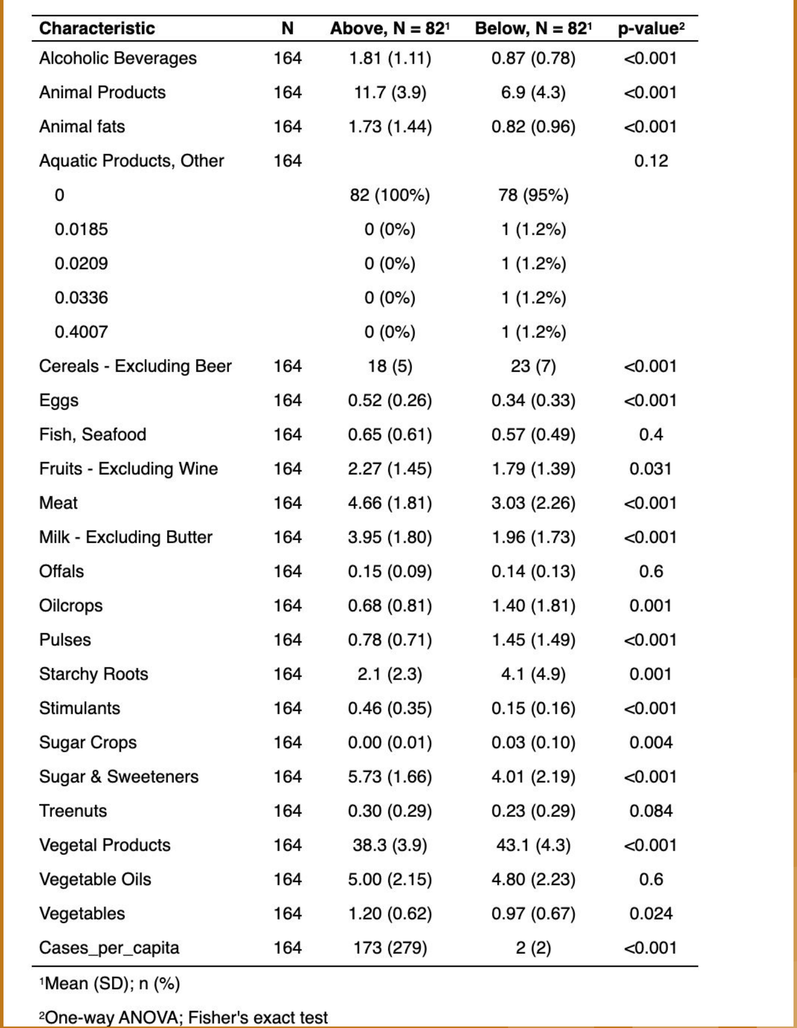
Does the American diet make us more vulnerable to Covid-19?

by Sam Bloch
10.15.2020, 5:31pm

Dataset Information & Preprocessing

- Kaggle dataset with information on the **average food intake breakdowns by food category** and **COVID-19 cases and deaths** of **170 countries** for the week of February 6, 2021
  - COVID-19 data - JHU COVID dashboard
  - Food intake data - Food and Agriculture Organization of the United Nations
- **Predictors:** Percentages for 22 different food categories (Animal Products, Animal Fats, Cereals, Stimulants, Vegetal Products, etc.)
  - *Spices, miscellaneous removed due to ambiguity
- **Outcomes:** Cases per capita, Deaths per capita, Cases vs. Median, Deaths vs. Median
  - Calculated using case counts and population size numbers provided in original dataset
- **Preprocessing:** Spearman’s Correlation Coefficients showed Vegetal Products and Animal Products had high correlations with other variables
  - Removed from dataset before fitting/running most models

Summary Statistics

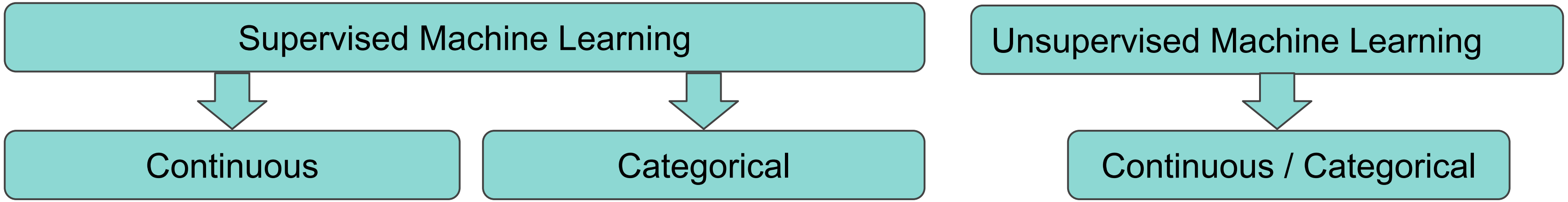


Characteristic	N	Above, N = 82[1]	Below, N = 82[1]	p-value[2]
Alcoholic Beverages	164	1.81 (1.11)	0.87 (0.78)	<0.001
Animal Products	164	11.7 (3.9)	6.9 (4.3)	<0.001
Animal fats	164	1.73 (1.44)	0.82 (0.96)	<0.001
Aquatic Products, Other	164			0.12
0		82 (100%)	78 (95%)	
0.0185		0 (0%)	1 (1.2%)	
0.0209		0 (0%)	1 (1.2%)	
0.0336		0 (0%)	1 (1.2%)	
0.4007		0 (0%)	1 (1.2%)	
Cereals - Excluding Beer	164	18 (5)	23 (7)	<0.001
Eggs	164	0.52 (0.26)	0.34 (0.33)	<0.001
Fish, Seafood	164	0.65 (0.61)	0.57 (0.49)	0.4
Fruits - Excluding Wine	164	2.27 (1.45)	1.79 (1.39)	0.031
Meat	164	4.66 (1.81)	3.03 (2.26)	<0.001
Milk - Excluding Butter	164	3.95 (1.80)	1.96 (1.73)	<0.001
Offals	164	0.15 (0.09)	0.14 (0.13)	0.6
Oilcrops	164	0.68 (0.81)	1.40 (1.81)	0.001
Pulses	164	0.78 (0.71)	1.45 (1.49)	<0.001
Starchy Roots	164	2.1 (2.3)	4.1 (4.9)	0.001
Stimulants	164	0.46 (0.35)	0.15 (0.16)	<0.001
Sugar Crops	164	0.00 (0.01)	0.03 (0.10)	0.004
Sugar & Sweeteners	164	5.73 (1.66)	4.01 (2.19)	<0.001
Treenuts	164	0.30 (0.29)	0.23 (0.29)	0.084
Vegetal Products	164	38.3 (3.9)	43.1 (4.3)	<0.001
Vegetable Oils	164	5.00 (2.15)	4.80 (2.23)	0.6
Vegetables	164	1.20 (0.62)	0.97 (0.67)	0.024
Cases_per_capita	164	173 (279)	2 (2)	<0.001

[1] Mean (SD); n (%)

[2] One-way ANOVA; Fisher's exact test

Methods



Supervised Machine Learning

Unsupervised Machine Learning

Continuous

Categorical

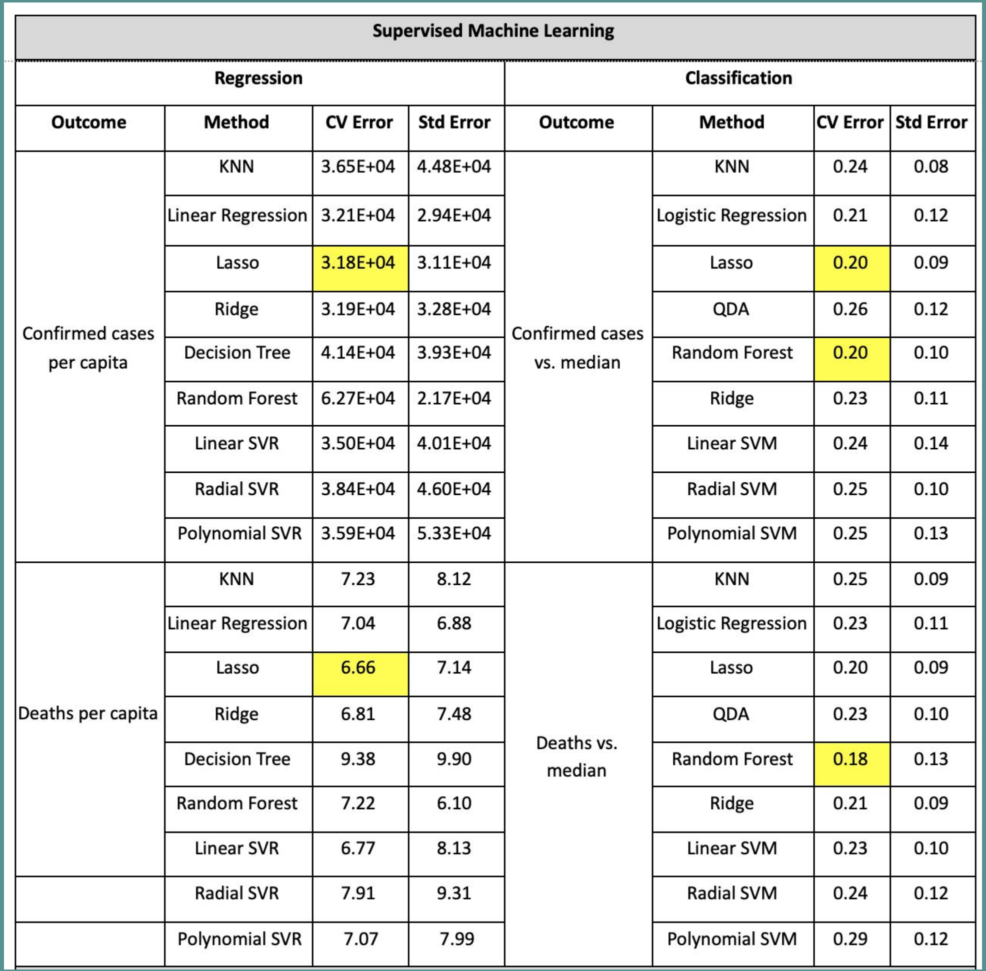
Continuous / Categorical

- KNN
- Linear regression
- Lasso
- Ridge
- Decision tree
- Random forest
- SVM
    - Linear
    - Radial
    - polynomial

- KNN
- Logistic regression
- Lasso
- Ridge
- QDA
- Random forest
- SVM
    - Linear
    - Radial
    - polynomial

- PCA + KNN
- K-means

Supervised ML Results Overview



Supervised Machine Learning							
Regression				Classification			
Outcome	Method	CV Error	Std Error	Outcome	Method	CV Error	Std Error
Confirmed cases per capita	KNN	3.65E+04	4.48E+04	Confirmed cases vs. median	KNN	0.24	0.08
	Linear Regression	3.21E+04	2.94E+04		Logistic Regression	0.21	0.12
	Lasso	3.18E+04	3.11E+04		Lasso	0.20	0.09
	Ridge	3.19E+04	3.28E+04		QDA	0.26	0.12
	Decision Tree	4.14E+04	3.93E+04		Random Forest	0.20	0.10
	Random Forest	6.27E+04	2.17E+04		Ridge	0.23	0.11
	Linear SVR	3.50E+04	4.01E+04		Linear SVM	0.24	0.14
	Radial SVR	3.84E+04	4.60E+04		Radial SVM	0.25	0.10
	Polynomial SVR	3.59E+04	5.33E+04		Polynomial SVM	0.25	0.13
Deaths per capita	KNN	7.23	8.12	Deaths vs. median	KNN	0.25	0.09
	Linear Regression	7.04	6.88		Logistic Regression	0.23	0.11
	Lasso	6.66	7.14		Lasso	0.20	0.09
	Ridge	6.81	7.48		QDA	0.23	0.10
	Decision Tree	9.38	9.90		Random Forest	0.18	0.13
	Random Forest	7.22	6.10		Ridge	0.21	0.09
	Linear SVR	6.77	8.13		Linear SVM	0.23	0.10
	Radial SVR	7.91	9.31		Radial SVM	0.24	0.12
	Polynomial SVR	7.07	7.99		Polynomial SVM	0.29	0.12

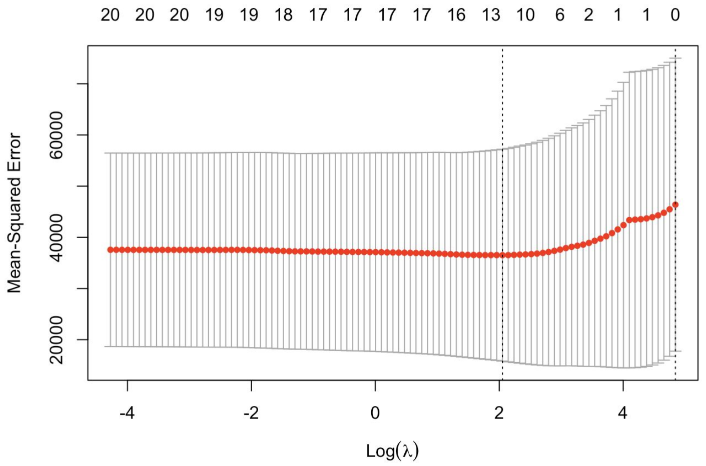
Results - Supervised Continuous

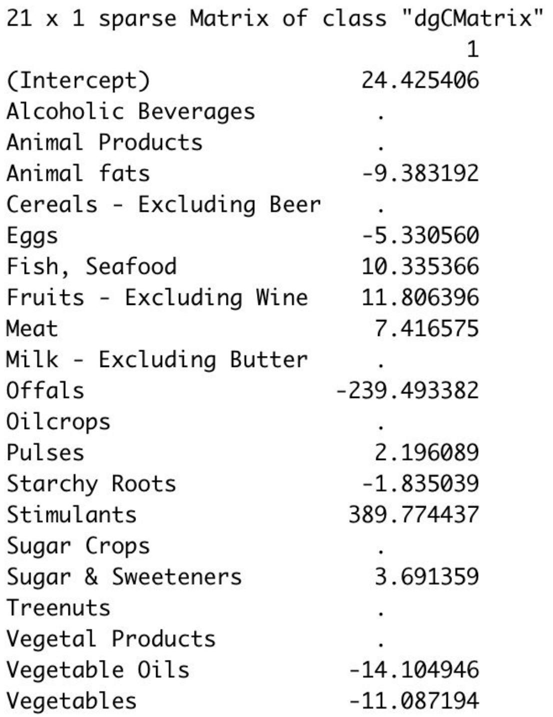
Top 3 Models for Cases per Capita outcome:

- Lasso Penalized Regression
- Ridge Penalized Regression
- Linear Regression

Full Lasso Model Results:

- 12 variables kept in model
- Stimulants, Offals were most important predictors





```
21 x 1 sparse Matrix of class "dgCMatrix"
                                    1
(Intercept)                 24.425406
Alcoholic Beverages                 .
Animal Products                     .
Animal fats                 -9.383192
Cereals - Excluding Beer            .
Eggs                        -5.330560
Fish, Seafood               10.335366
Fruits - Excluding Wine     11.806396
Meat                         7.416575
Milk - Excluding Butter             .
Offals                    -239.493382
Oilcrops                            .
Pulses                       2.196089
Starchy Roots               -1.835039
Stimulants                 389.774437
Sugar Crops                         .
Sugar & Sweeteners           3.691359
Treenuts                            .
Vegetal Products                    .
Vegetable Oils             -14.104946
Vegetables                 -11.087194
```

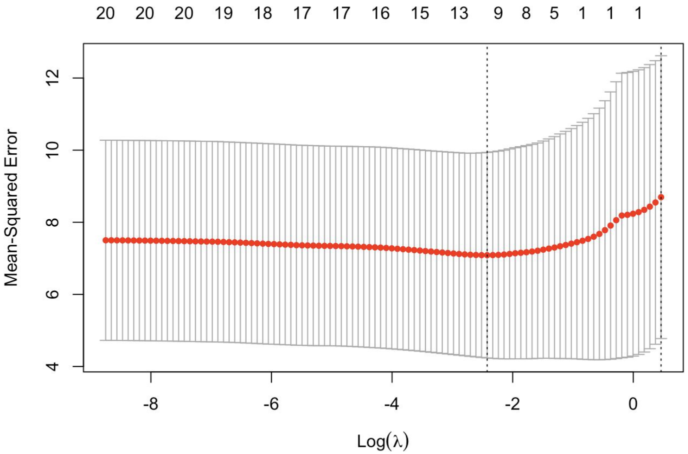
Results - Supervised Continuous

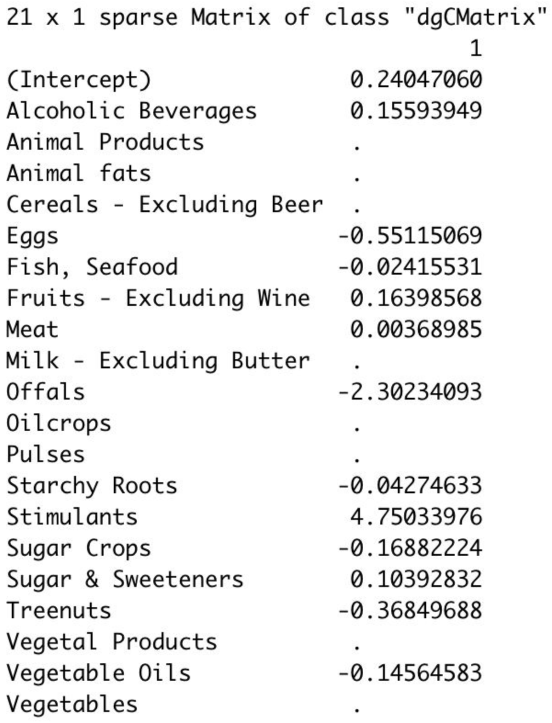
Top 3 Models for Deaths per Capita outcome:

- Lasso Penalized Regression
- Ridge Penalized Regression
- Linear SVR

Full Lasso Model Results:

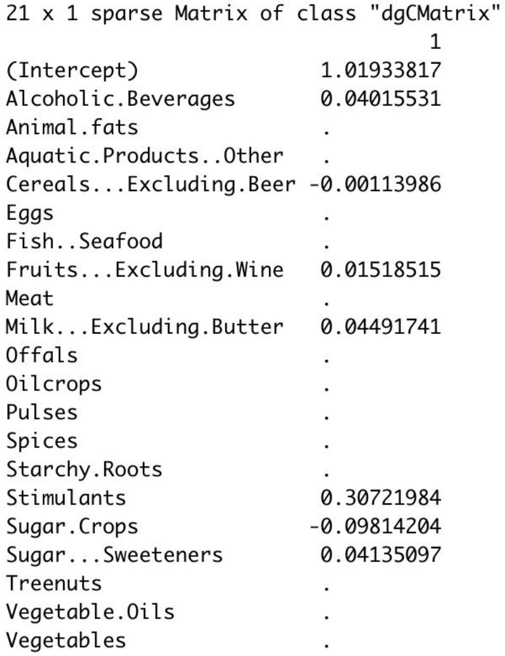
- 12 variables kept in model
- Stimulants, Offals were most important predictors




```
21 x 1 sparse Matrix of class "dgCMatrix"
                                 1
(Intercept)             0.24047060
Alcoholic Beverages     0.15593949
Animal Products         .
Animal fats             .
Cereals - Excluding Beer .
Eggs                   -0.55115069
Fish, Seafood          -0.02415531
Fruits - Excluding Wine  0.16398568
Meat                    0.00368985
Milk - Excluding Butter  .
Offals                 -2.30234093
Oilcrops                .
Pulses                  .
Starchy Roots          -0.04274633
Stimulants              4.75033976
Sugar Crops            -0.16882224
Sugar & Sweeteners      0.10392832
Treenuts               -0.36849688
Vegetal Products        .
Vegetable Oils         -0.14564583
Vegetables              .
```

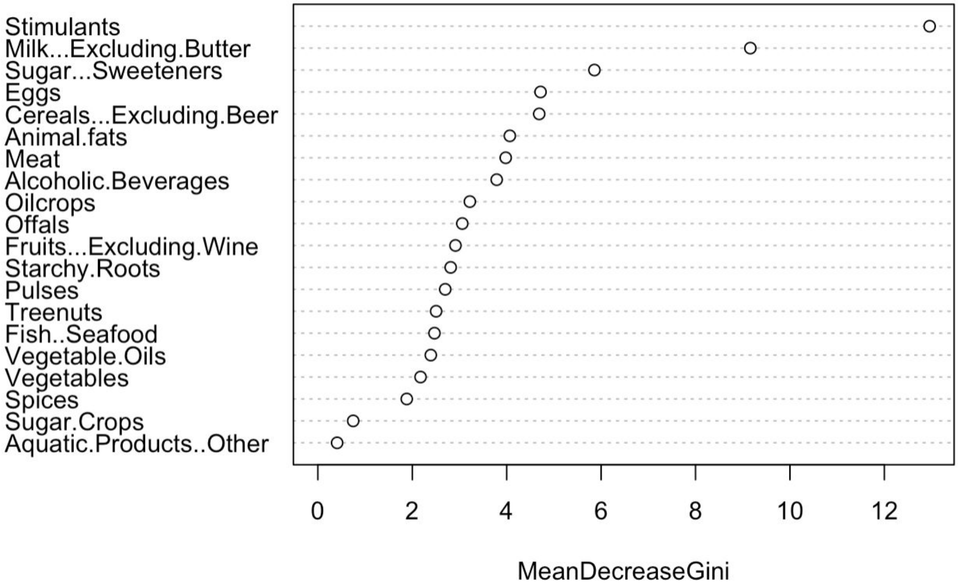
Results - Supervised Classification

Lasso - Predicting Cases vs Median



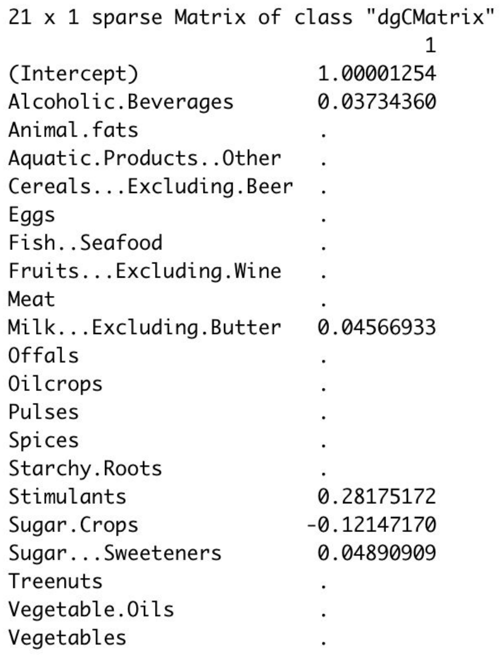
```
21 x 1 sparse Matrix of class "dgCMatrix"
                                   1
(Intercept)                1.01933817
Alcoholic.Beverages        0.04015531
Animal.fats                .
Aquatic.Products..Other    .
Cereals...Excluding.Beer -0.00113986
Eggs                       .
Fish..Seafood              .
Fruits...Excluding.Wine    0.01518515
Meat                       .
Milk...Excluding.Butter    0.04491741
Offals                     .
Oilcrops                   .
Pulses                     .
Spices                     .
Starchy.Roots              .
Stimulants                 0.30721984
Sugar.Crops               -0.09814204
Sugar...Sweeteners         0.04135097
Treenuts                   .
Vegetable.Oils             .
Vegetables                 .
```

Random forest - Variable Importance Ranking



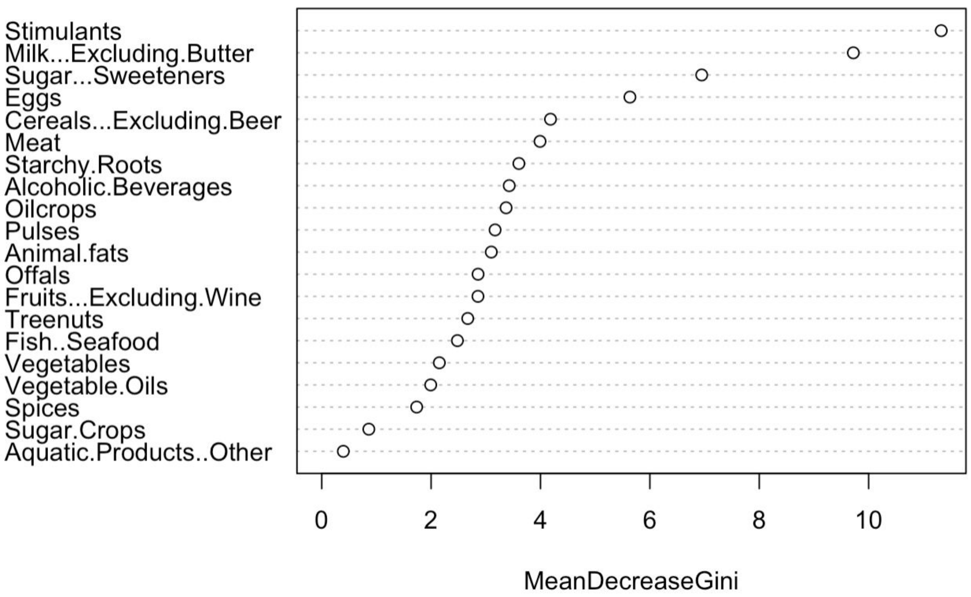
Results - Supervised Classification

Lasso - Predicting Deaths vs Median



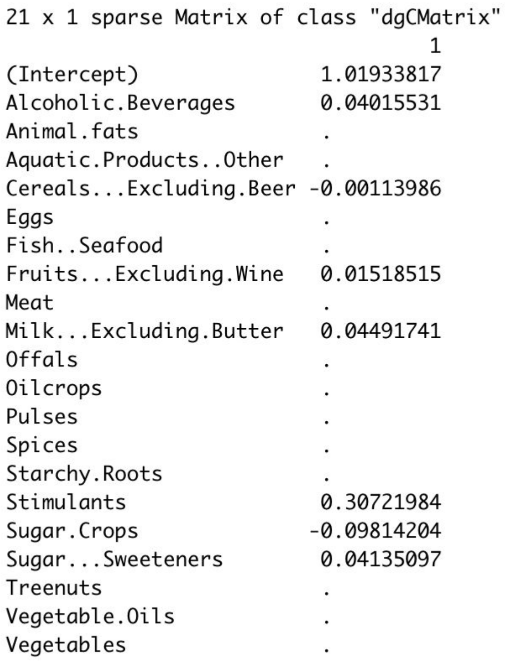
```
21 x 1 sparse Matrix of class "dgCMatrix"
                                  1
(Intercept)              1.00001254
Alcoholic.Beverages      0.03734360
Animal.fats              .
Aquatic.Products..Other  .
Cereals...Excluding.Beer .
Eggs                     .
Fish..Seafood            .
Fruits...Excluding.Wine  .
Meat                     .
Milk...Excluding.Butter  0.04566933
Offals                   .
Oilcrops                 .
Pulses                   .
Spices                   .
Starchy.Roots            .
Stimulants               0.28175172
Sugar.Crops             -0.12147170
Sugar...Sweeteners       0.04890909
Treenuts                 .
Vegetable.Oils           .
Vegetables               .
```

Random forest - Variable Importance Ranking



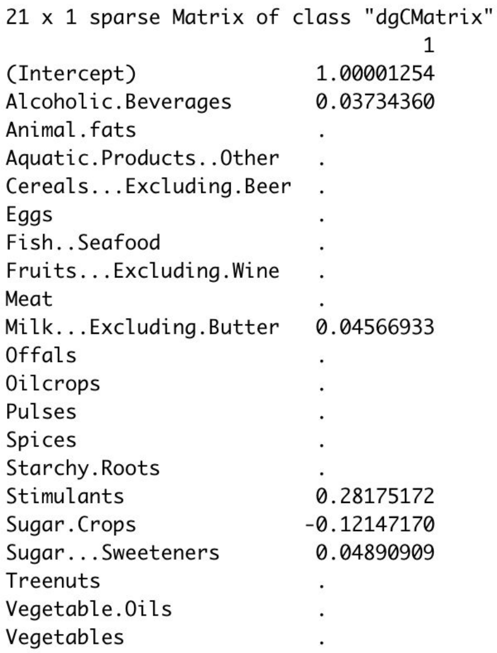
Results - Supervised Classification

Lasso - Predicting Cases vs Median



```
21 x 1 sparse Matrix of class "dgCMatrix"
                                     1
(Intercept)                 1.01933817
Alcoholic.Beverages         0.04015531
Animal.fats                 .
Aquatic.Products..Other     .
Cereals...Excluding.Beer   -0.00113986
Eggs                        .
Fish..Seafood               .
Fruits...Excluding.Wine     0.01518515
Meat                        .
Milk...Excluding.Butter     0.04491741
Offals                      .
Oilcrops                    .
Pulses                      .
Spices                      .
Starchy.Roots               .
Stimulants                  0.30721984
Sugar.Crops                -0.09814204
Sugar...Sweeteners          0.04135097
Treenuts                    .
Vegetable.Oils              .
Vegetables                  .
```

Lasso - Predicting Deaths vs Median

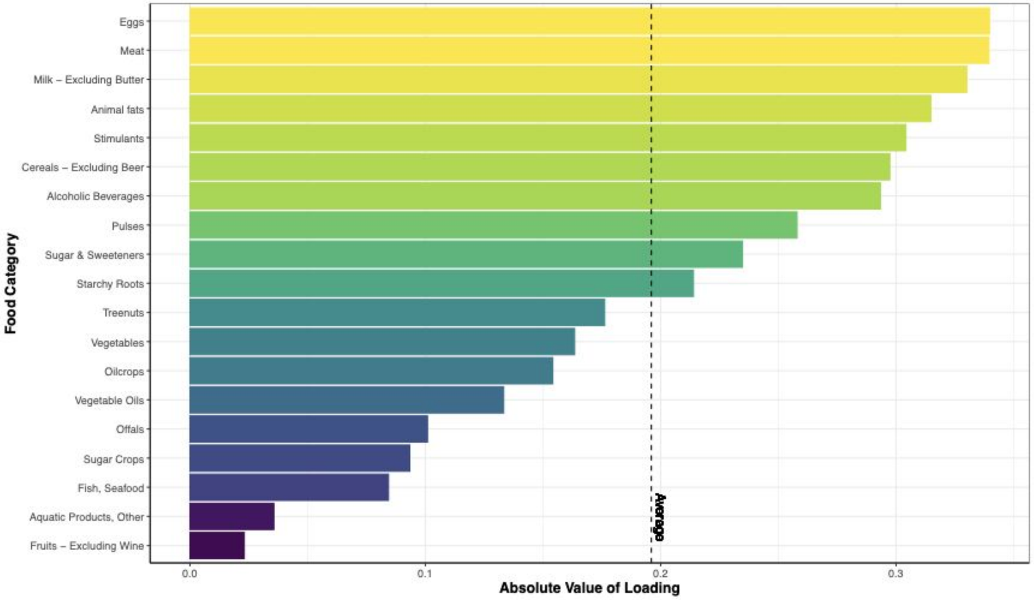


```
21 x 1 sparse Matrix of class "dgCMatrix"
                                     1
(Intercept)                 1.00001254
Alcoholic.Beverages         0.03734360
Animal.fats                 .
Aquatic.Products..Other     .
Cereals...Excluding.Beer    .
Eggs                        .
Fish..Seafood               .
Fruits...Excluding.Wine     .
Meat                        .
Milk...Excluding.Butter     0.04566933
Offals                      .
Oilcrops                    .
Pulses                      .
Spices                      .
Starchy.Roots               .
Stimulants                  0.28175172
Sugar.Crops                -0.12147170
Sugar...Sweeteners          0.04890909
Treenuts                    .
Vegetable.Oils              .
Vegetables                  .
```

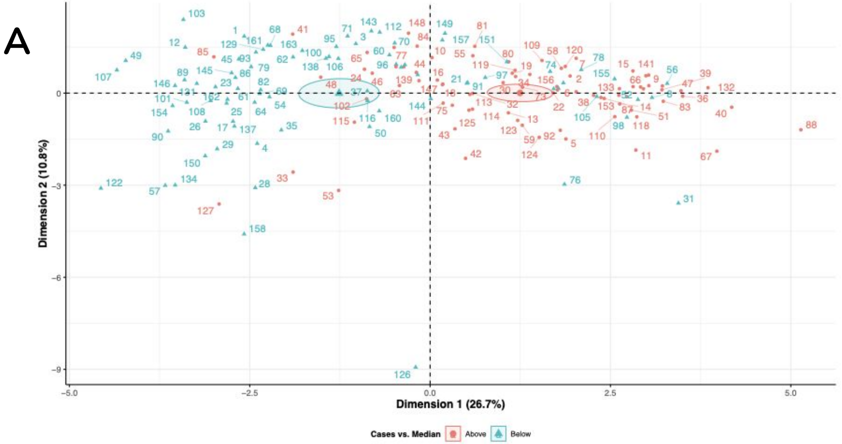
Results - Unsupervised

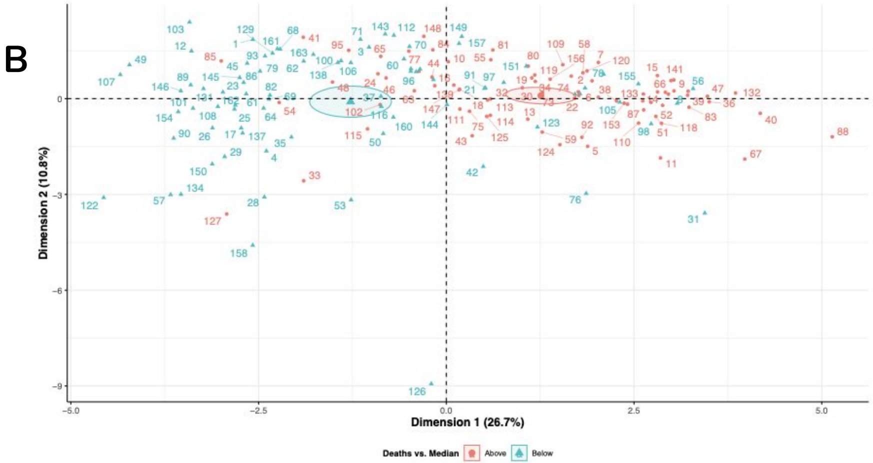
- First 3 eigenvectors captured 45.8% of the variance

- Top contributors to principal components
    1. Eggs
    2. Meat
    3. Milk
    4. Animal Fats
    5. Stimulants (ie. Coffee & Tea)



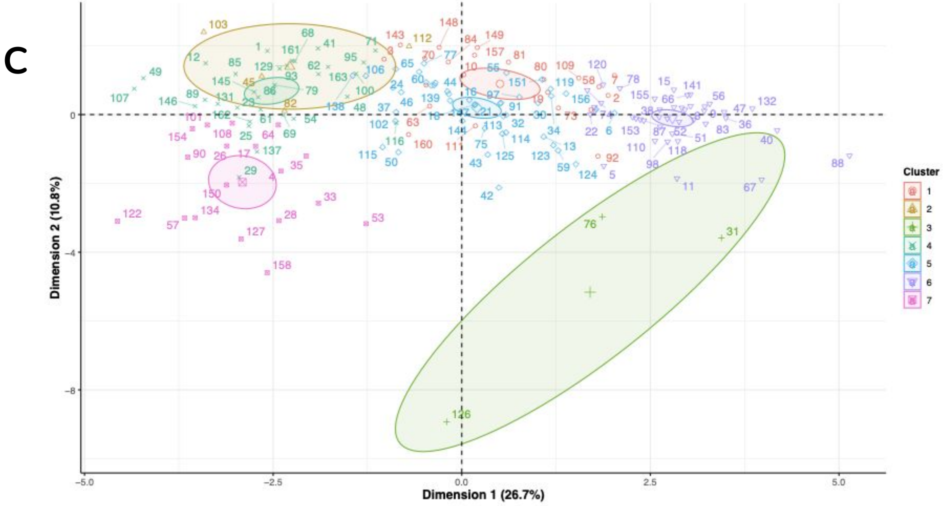
Results - Unsupervised





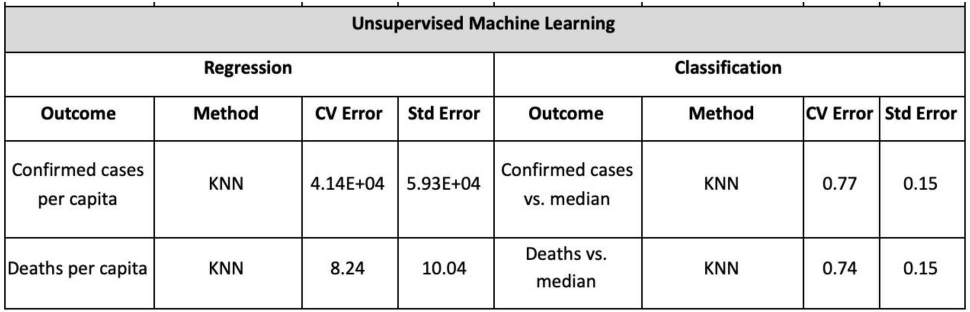
- K-means predicted 7 clusters (outcomes)

- High degree of overlap between clusters indicates that food consumption data are poor predictors of COVID-19 outcomes



Results & Discussion - Unsupervised

- Overall, unsupervised KNN models had higher biases and variances than all supervised machine learning models, including KNN
    - Contributors: non-linear associations, high variability in original dataset, substantial overlap between dichotomous outcomes



Unsupervised Machine Learning							
Regression				Classification			
Outcome	Method	CV Error	Std Error	Outcome	Method	CV Error	Std Error
Confirmed cases per capita	KNN	4.14E+04	5.93E+04	Confirmed cases vs. median	KNN	0.77	0.15
Deaths per capita	KNN	8.24	10.04	Deaths vs. median	KNN	0.74	0.15

Final Conclusions

- **Significance of Findings:**
  - Diet is not a strong standalone predictor of COVID-19 cases/deaths at the country level, though it does have some impact
  - Stimulants and Sugars/Sweeteners consistently were most strongly associated with higher COVID cases/deaths
    - Supported by previous studies - link between obesity and COVID-19 severity
- **Limitations:**
  - Left out other potentially important factors like availability of healthcare, GDP, messaging from government public health organizations, health system infrastructure
  - Higher calorie foods were weighted more heavily - made models unable to completely characterize the impact of food intake patterns on health
  - Lacked data on macronutrients, vitamins, minerals, etc.
- **Further Analyses:**
  - Incorporate other potential factors related to COVID-19 prevalence/mortality
  - Examining how diet influences COVID-19 risk at the individual level

Works Cited

https://builtin.com/data-science/step-step-explanation-principal-component-analysis
https://cran.ism.ac.jp/
https://coronavirus.jhu.edu/map.html
https://www.ncbi.nlm.nih.gov/books/NBK554776/
https://www.health.harvard.edu/diseases-and-conditions/if-youve-been-exposed-to-the-coronavirus
https://pubmed.ncbi.nlm.nih.gov/?term=covid+AND+machine+learning&size=100
https://pubmed.ncbi.nlm.nih.gov/33070540/
https://pubmed.ncbi.nlm.nih.gov/33027032/
https://pubmed.ncbi.nlm.nih.gov/32311498/
https://www.kaggle.com/mariaren/covid19-healthy-diet-dataset?select=Food_Supply_Quantity_kg_Data.csv

Thank you!