Project 4

Overview of Data / Highlights

- Data describing diet tendencies (percentage of diet from meat, vegetables, etc.) on a national level, included in the data set were percentage of COVID-19 deaths.
- Data was last collected 02/06/2021

https://www.kaggle.com/mariaren/covid19-healthy-diet-dataset

Overview of Data/Highlights

First analysis was to look at the data in Tableau to explore data coverage and patterns

https://public.tableau.com/app/profile/jill.peloquin/viz/Healthdata_16369294240380/Story1?publish=yes

Overview of Data / Highlights

Additional processing:

1. Drop NaNs
2. Convert the y axis (death) data from continuous to category by binning them into distinct ranges from 0-0.20 so we can use them with Random Forest Classifier.
   a. bins=[0, 0.05, 0.10, 0.15, 0.2]
   b. labels=[0,1,2,3]
3. Trimmed the original data set down based on the experimental run
4. Scaled data

Multiple Linear Regression

Alcoholic Beverages
Animal fats
Animal Products
Aquatic Products, Other
Cereals - Excluding Beer
Eggs
Fish, Seafood
Fruits - Excluding Wine
Meat
Milk - Excluding Butter
Miscellaneous
Offals
Oilcrops
Pulses
Spices
Starchy Roots
Stimulants
Sugar & Sweeteners
Sugar Crops
Treenuts
Vegetable Oils
Vegetables
Vegetal Products
Obesity
Undernourished
Confirmed
Deaths
Recovered
Active
Population
Unit (all except Population)

Experimenting with Features to include:

1. **Small:** Alcoholic Beverages, Animal Products, Meat, Fish, Seafood, Vegetables
2. **Medium:** Alcoholic Beverages, Animal fats, Animal Products, Aquatic Products, Other, Cereals - Excluding Beer, Eggs, Fish, Seafood, Meat, Milk - Excluding Butter, Sugar & Sweeteners, Vegetable Oils, Vegetables, Vegetal Product
3. **All:** All columns except: Confirmed, Recovered, Active, Population, Unit (all except Population),Obesity,Undernourished

Multiple Linear Regression

Suggests the more features (data) we give the model, the better explanation we have toward understanding the effects of nutrition on COVID-19 survival. However, none of these fits for Linear Regression are exceptionally high.

Experiment	Score
Small	0.388
Medium	0.429
All	0.480

Logistic Regression with and without reduced features with Random Forest Classifier

Experiment	Training/Testing	Training/Testing (with feature reduction)
Small	Training Score: 0.715 Testing Score: 0.609	Reduced to 1 Feature Training Score: 0.674 Testing Score: 0.536
Medium	Training Score: 0.756 Testing Score: 0.634	Reduced to 4 Features Training Score: 0.699 Testing Score: 0.512
All	Training Score: 0.813 Testing Score: 0.634	Reduced to 8 Features Training Score: 0.780 Testing Score: 0.560

Feature	S	M	A	Feature	S	M	A	Feature	S	M	A
Alcoholic Beverages	No	No	Yes	Meat	No	No	No	Stimulants			No
Animal fats		No	Yes	Milk - Excluding Butter		Yes	Yes	Sugar & Sweeteners		No	No
Animal Products	Yes	Yes	Yes	Miscellaneous			Yes	Sugar Crops			No
Aquatic Products, Other		No	No	Offals			No	Vegetable Oils		No	No
Cereals - Excluding Beer		No	No	Oilcrops			Yes	Vegetables	No	No	No
Eggs		Yes	Yes	Pulses			No	Vegetal Products		Yes	Yes
Fish, Seafood	No	No	No	Spices			No				
Fruits - Excluding Wine			No	Starchy Roots			No				

Summary

Our analysis suggests the more features (data) we give the model, the better explanation we have toward understanding the effects of nutrition on COVID-19 survival. However, with just 8 features we can explain a country’s death rate in a fairly compelling way.