COVID Death Rates and Diet

Adriana Siqueiros, Sana Khader, Ananya Krishnan, Keilyn Yuzuki, Nishant Mishra

Can we accurately predict COVID-19 death rates using country diet data?

Project Context and Motivation

Is diet partly responsible for differences in COVID-19 death rates between and within countries?

Jean Bousquet,[1,2,3,4] Josep M. Anto,[5,6,7,8] Guido Iaccarino,[9] Wienczyslawa Czarlewski,[10,11] Tari Haahtela,[12] Aram Anto,[10] Cezmi A. Akdis,[13] Hubert Blain,[14,15] G. Walter Canonica,[16] Victoria Cardona,[17] Alvaro A. Cruz,[18] Maddalena Illario,[19,20] Juan Carlos Ivancevich,[21,22] Marek Jutel,[23] Ludger Klimek,[24] Piotr Kuna,[25] Daniel Laune,[26] Désirée Larenas-Linnemann,[27] Joaquim Mullol,[28] Nikos G. Papadopoulos,[29,30] Oliver Pfaar,[31] Boleslaw Samolinski,[32] Arunas Valiulis,[33] Arzu Yorgancioglu,[34] Torsten Zuberbier,[1,2,3,4] and The ARIA group

Relation of Dietary Factors with Infection and Mortality Rates of COVID-19 Across the World

Deldar Morad Abdulah[1] and A. B. Hassan[2]

Dataset

Data Information

- **Data**: Kaggle COVID-19 Healthy Diet Dataset
- 4 Data Sets:
    - Fat_Supply_Quantity_Data
    - Food_Supply_Quantity_Data (one by kcal, one by kilograms)
    - Protein_Supply_Quantity_Data
- Rows: 170 (1 for each country in the data)
- Columns (for each dataset d = [Fat, Food, Protein]):
    - Percentage of d intake from alcoholic beverages
    - Percentage of d intake from animal products
    - Percentage of d intake from animal fats
    - Percentage of d intake from aquatic products
    - Percentage of d intake from cereals - excluding beer
    - Percentage of d intake from eggs
    - Percentage of d intake from fish, seafood
    - Percentage of d intake from fruits - excluding wine
    - Percentage of d intake from meat
    - Obesity, Death rate, Recovered, Confirmed, Active

Our Process

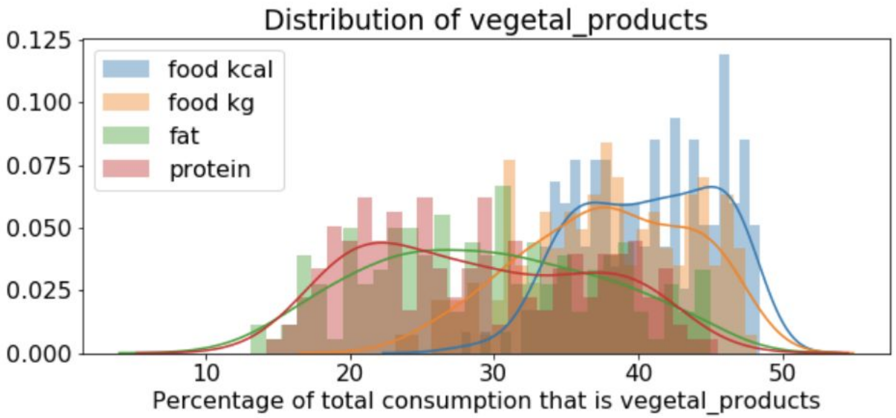
- Trained independent models to explore possible areas of interest
  - Fat content
  - Protein content
  - Caloric content
  - Food Quantity
- Trained and optimized models on combined data sets to predict death rate (continuous)
  - Linear regression
  - Neural Network
- Trained models on combined dataset to predict categorical death rate (high/medium/low)
  - Logistic Regression
  - Neural Network (Tensor Flow)

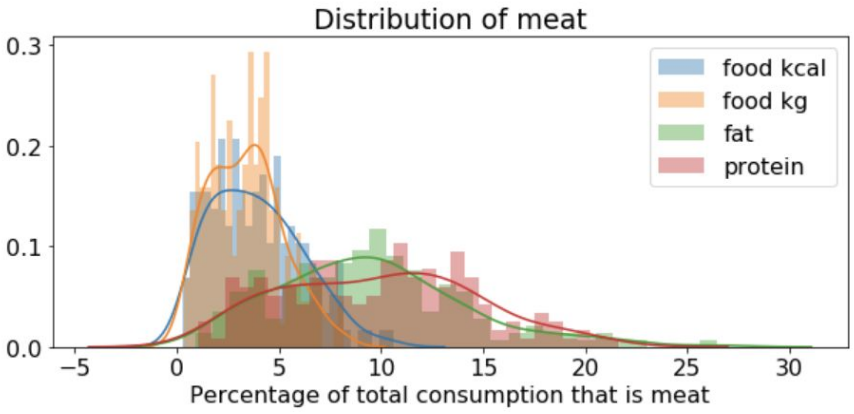
Understanding the Data

Combined the food kcal, food kg, fat, and protein supply datasets

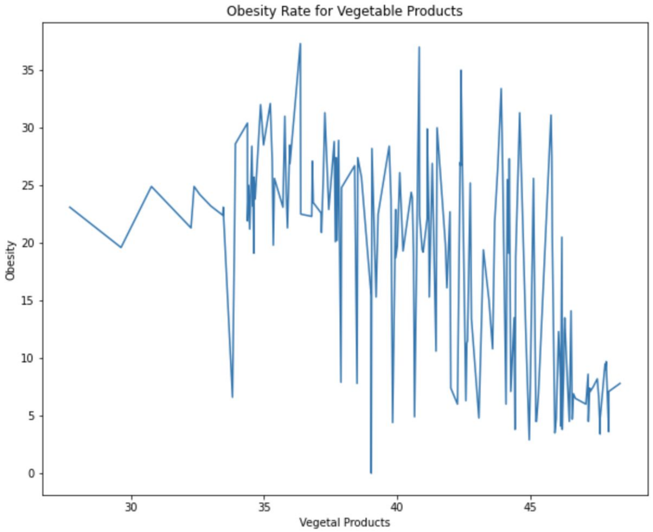
- Vegetable products typically make up between 30-50% of the diet, both by weight and by calories
- Countries are fairly spread out in terms of how much fat/protein they get from vegetal products

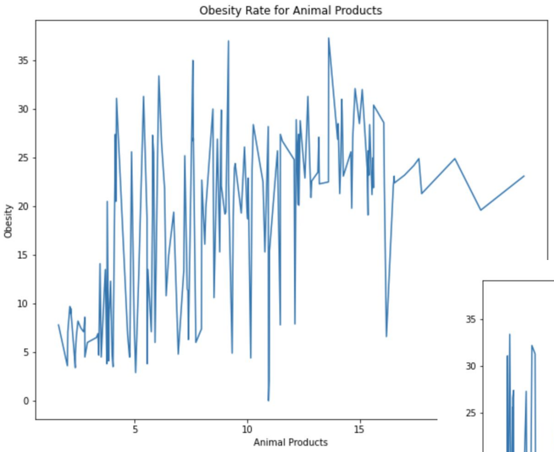
- For almost all countries, less than 10% of the total food intake is meat
- Meat provides typically between 0-20% of fat and protein content in the diet

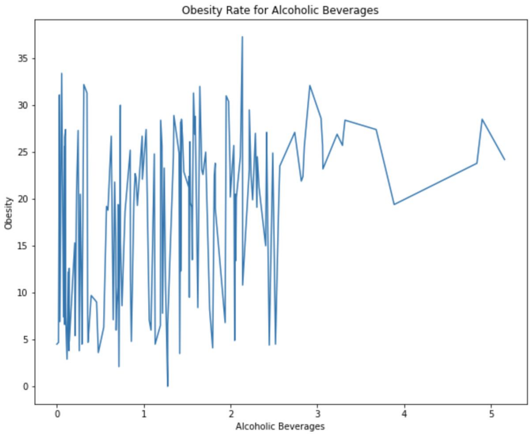




Obesity and Diet







- Inverse relationship between obesity rates and vegetable product consumption

- Positive relationship between obesity rates and Alcohol and Animal product consumption

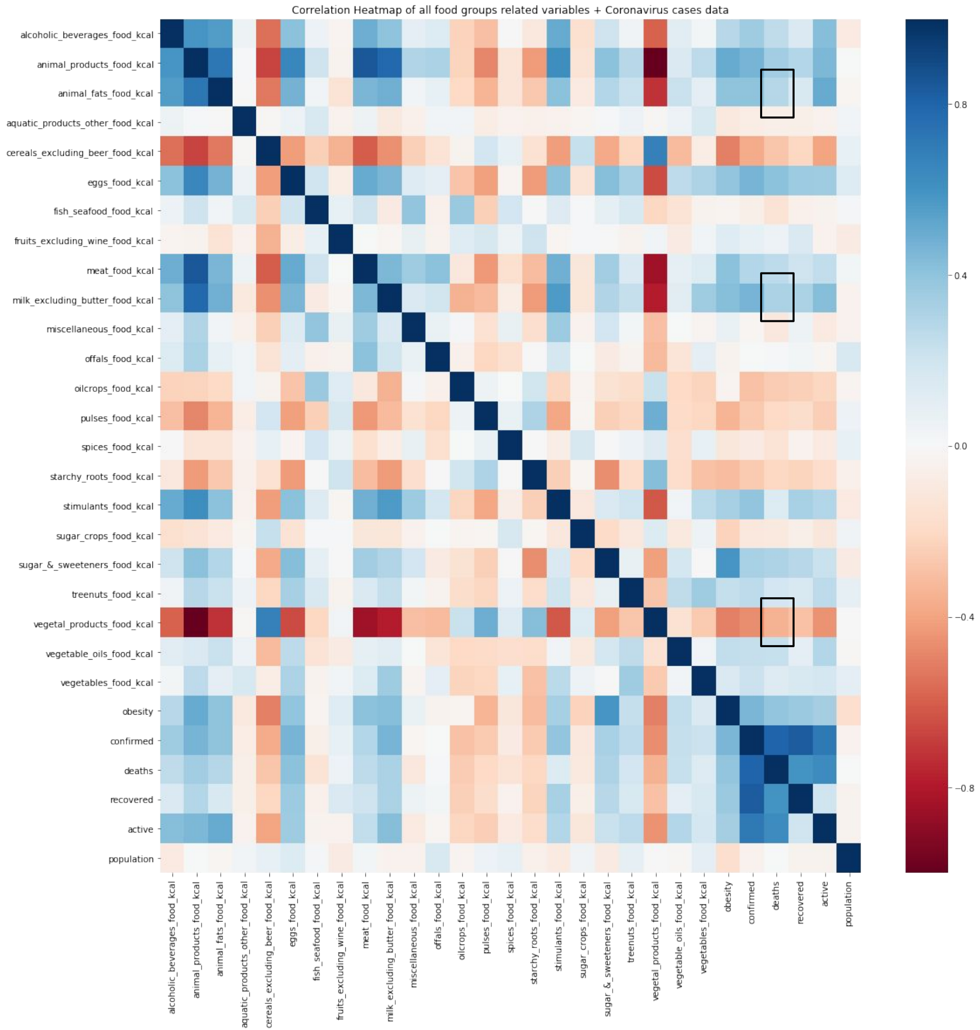
Death Rate vs Obesity Rate

Death Rate

Obesity

Correlation Heat Map

- Key Takeaways
  - positive association with deaths and animal products, milk (excluding butter)
  - negative association with deaths and vegetable products

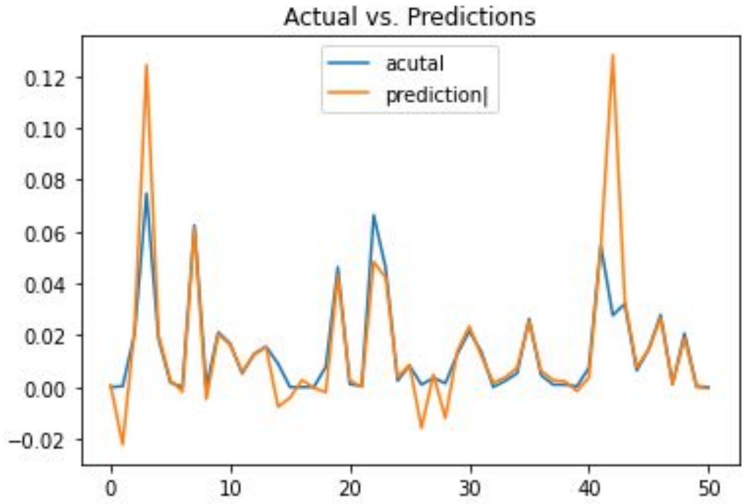


Approaches

Linear Regression

sklearn.linear_model.LinearRegression

- MSE = 0.000281
- R^2 = 0.196



Linear Regression, contd.

on complete protein dataset:

(1) Lasso Regression

- MSE: 0.000327
- R^2: 0.0433

(2) Ridge Regression

- MSE: 0.000294
- R^2: 0.14

on selected protein dataset:

(1) Lasso Regression

- MSE: 0.000353
- R^2: 0.166

(2) Ridge Regression

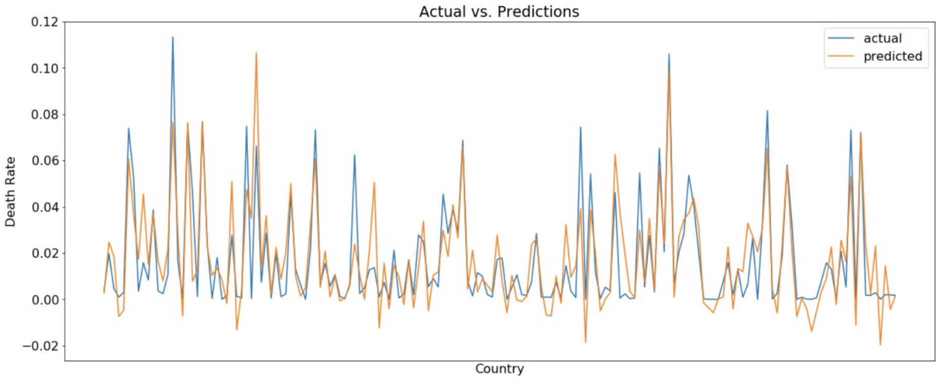
- MSE: 0.000273
- R^2: 0.166

Neural Network trained on combined dataset

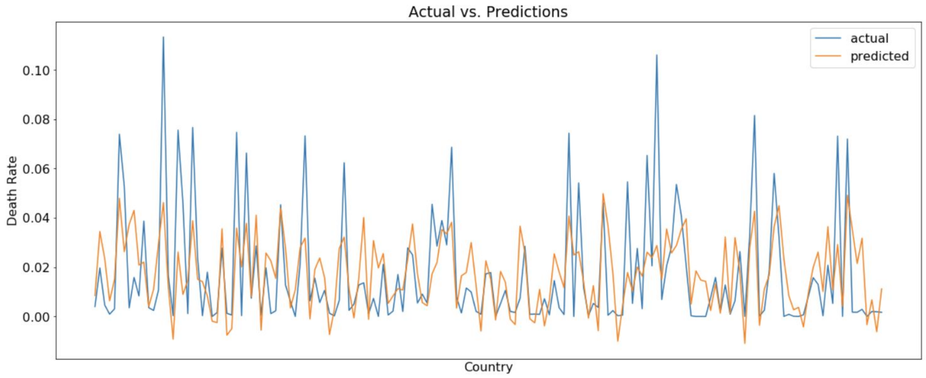
MLPRegressor - Activation: ReLU, Solver: LBFGS

- 2 models - 5 (left) versus 7 (right) hidden layers
- Tradeoff with accuracy on countries with higher death rates versus overall RMSE
- Widely varying R^2 values between models and test/train data

5 Hidden Layers
Coefficient of Determination, Train: 0.857
Coefficient of Determination, Test: 0.204
RMSE: 0.020



7 Hidden Layers
Coefficient of Determination, Train: 0.407
Coefficient of Determination, Test: 0.349
RMSE: 0.018



Binary/Ternary Classification

Binary

- Logistic Regression
  - Accuracy: 75%
  - AUC: 0.74
- Keras Tensor Flow NN Model with 2 64 node relu layers and sigmoid output layer
  - Accuracy: 82%
  - AUC: 0.9444
- Most relevant features:

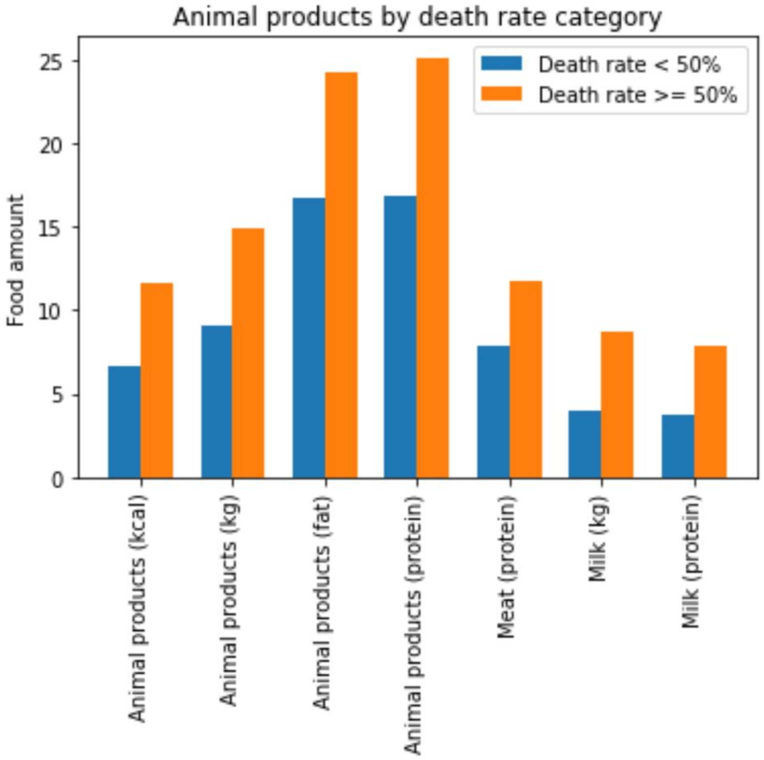
Features	Coef
obesity	2.217811e-14
animal_products_protein	6.418858e-15
animal_products_fat	5.184291e-15
milk_excluding_butter_protein	4.720170e-15
animal_products_food_kcal	4.692738e-15
sugar_&_sweeteners_food_kcal	4.654641e-15
milk_excluding_butter_food_kg	4.086446e-15
animal_products_food_kg	4.065999e-15

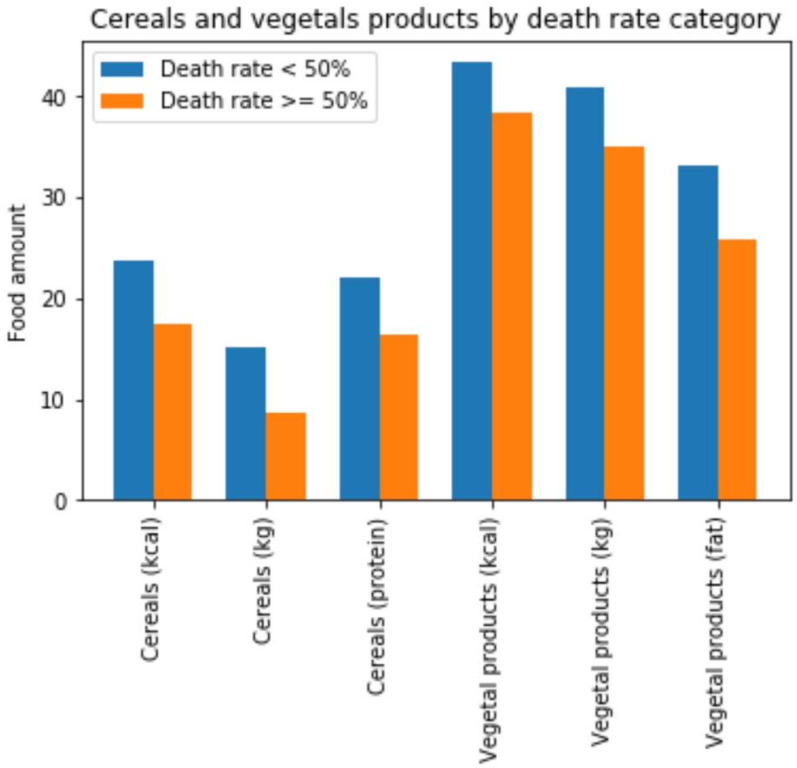
Ternary

- Logistic Regression
  - Accuracy: 39%
- Keras Tensor Flow NN Model with 2 64 node relu layers and sigmoid output layer
  - Accuracy: 47%
- Most relevant features:

Features	Coef
population	3.194849e-09
vegetal_products_food_kg	3.055966e-14
vegetal_products_food_kcal	2.916529e-14
vegetal_products_protein	2.883876e-14
vegetal_products_fat	2.821662e-14
reals_excluding_beer_food_kcal	1.553541e-14
starchy_roots_food_kg	1.546364e-14
cereals_excluding_beer_protein	1.413555e-14
vegetable_oils_fat	1.229698e-14
:ereals_excluding_beer_food_kg	1.222656e-14
starchy_roots_food_kcal	9.637387e-15

Binary Classifier Exploration





Results and Conclusions

Summary of Findings

Predicting continuous death rate:

- Low R^2 values indicate poor predictors
- Low RMSE value due to smaller data values (proportions between 0-1)
- High overfitting due to discrepancy in loss between train and test

Predicting categorical (high/low/medium) death rate:

- Higher accuracy, but not as informative
- Binary classification much stronger than ternary
- Key features that distinguish different levels of death rate that is consistent with correlation map
    - Features to distinguish different death rates
        - Vegetable Products
        - Milk
        - Animal Products

Can we accurately predict COVID-19 death rates using country diet data?

- Predicting exact death rates from this dataset is not possible
- No definitive relationship between diet and death from COVID-19, but there exists an association
- Possible to distinguish high/low death rate for a country with a specific diet

How could we have done better?

- Larger data set
- More granular data→ death rate and diet by individual would have been more predictive
- There are ultimately larger factors at play that influence death rate (government policy, poverty, healthcare systems, comorbidities, etc.)

Real World Applications

- These models can be used to anticipate high death rates in countries known to have certain diets
- Changes in diet can’t change people’s overall health or COVID death rates in the short-term, but other measures can (social distancing, lockdown, contact tracing, etc.)
- In the long term, changes in diet could be used to reduced death rates

Thank you!
Q&A