

# Predicting Meat Consumption Levels and Investigating the Reduction of Meat Consumption in the United States through Financial Incentives

Project Report - Advance Topics in DS

**Team Members:** Carlos Figueroa, Carla Garcia Medina, Hannah Won

**Team Name:** The Karl-Won Project

## Abstract

Meat consumption makes up approximately 12% of greenhouse gas emissions in the United States and beef and veal are the most highly consumed meat products per capita. With the aim of helping policymakers reduce this consumption and, consequently, decrease the rate of climate change, this project has two distinct goals: (i) to develop a predictor of beef and veal consumption levels given agricultural commodity data, (ii) to perform a feature analysis and investigate which commodities should be subsidized. Our project shows that a KNeighborsRegressor model performs the best at predicting beef and veal consumption, and can make predictions that have errors (RMSE score) as low as .33. For our feature analysis, we employ methods including Pearson correlations, Random Forests, XGBoostRegressor, and XGBoostClassifier, as well as a dimensionality reduction pipelines of SVD and UMAP paired with Random Forests. All models highlight a variety of commodities meaning that policymakers can choose to subsidize different commodities depending on cultural/religious norms and still have a high effect on meat consumption. Additionally, dimensionality reduction shows 2 or 3 features are able to represent data well so even the subsidization of only 2 can be greatly beneficial to the reduction of greenhouse gas emissions.

## P1 - Background

As many countries strive to meet their greenhouse gas reduction targets, the role of meat consumption on climate change has come under scrutiny. The global livestock industry is responsible for approximately 14.5% of worldwide emissions, most of them coming from methane, a gas with 80 times the warming potential of carbon dioxide but with a way lower lifespan to remain in the atmosphere. In the current context of curbing global warming effects as soon as possible, the most evident proposal has been meat taxes since they can curb emissions dramatically without changing cultures and diets around the globe. However, the economic implications of such a tax are complex and hotly debated. In this section, we will discuss the feasibility of a meat tax, its economics, how this problem has it been approached in the past, and how we are approaching it in our project.

In their latest Summary for policymakers for the Special Report on Climate Change and Land, the Intergovernmental Panel on Climate Change (IPCC) concluded with “high confidence” that shifting

towards diets emphasizing plant-based foods “present major opportunities for adaptation and mitigation [of climate change] while generating significant co-benefits in terms of human health” (IPCC, 24). Furthermore, the agricultural sector was reported to account for approximately 11% of total greenhouse gas emissions in the US (EPA), with most of its emissions originating from meat production. There is a clear need to shift consumption away from meat products, towards plant-based foods in the near future, if we want to reduce greenhouse gas emissions.

Dissuading citizens from consuming meat products has proven extremely difficult due to cultural norms, however, financial and health incentives appear to be promising incentives for change in behavior. According to a study by Neff et al. for the Cambridge University Press, the main drivers of meat consumption reduction in the country, particularly for “red and processed meat” were “cost and health; [while] environment and animal welfare lagged.” Therefore, we investigate potential policy-based approaches that reduce meat consumption in an indirect manner by subsidizing different commodities related to food demand and supply. We focus on the specific reduction of beef and veal since these have been shown to be the greatest meat sources of greenhouse gas emissions (FAO).

### ***Economics background***

First, let us talk about the role of taxes in curbing production. From an economic perspective, this connects with the tax incident and Pigouvian taxes. First, the tax incident refers to the impact of taxes in reducing incentives to either produce or consume a product. The tax burden is not dependent on whether the state collects the revenue from the producer or consumer but on the price elasticity of supply and demand.

To understand the dynamics better, let us consider the example where the meat supply is inelastic, and demand is relatively elastic. In this situation, the firm will produce the same quantity regardless of the price because supply is inelastic. However, because demand is elastic, the consumer is susceptible to price. In other words, a slight increase in price leads to a significant drop in the quantity demanded. In this case, the producer cannot pass the tax onto the consumer, and the tax incidence falls on the producer.

On the other hand, if consumption is inelastic, the consumer will consume nearly the same quantity, no matter the price. As a result, the entirety of the tax will be borne by the consumer. Even so, consumers and producers are often neither perfectly elastic nor inelastic, so the tax burden is shared between the two parties in varying proportions. Furthermore, in the case of beef, supply is often elastic, but demand strongly depends on the type of beef we are talking about. Beef types like ground beef, with strong substitutes like chicken, fall under elastic demand. In contrast, fancy cuts, such as tomahawks and rib eyes, are relatively inelastic demand (their findings reveal nonlinear demands for meat products, with demand being more inelastic at higher prices (Lusk, Jayson, and Tonsor, Glynn. (2016.) "How Meat Demand Elasticities Vary with Price, Income, and Product Category. Applied Economic Perspectives and Policy, vol. 38(4), 673-711.)).

Second, we have the concept of Pigouvian taxes, which refers to any tax on a market activity that results in negative externalities or external expenses not reflected in the market price. In the case of meat, it is argued that its environmental impact is not accounted for in its price. Therefore, a meat tax could help

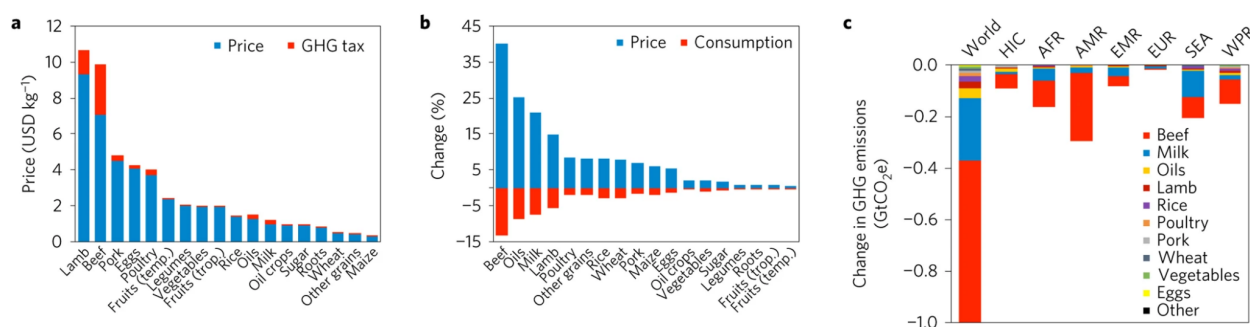
internalize the environmental costs of meat production by pricing in the negative externalities associated with greenhouse gas emissions, deforestation, water pollution, and other environmental impacts. This concept is familiar to markets: other examples of Pigovian taxes are taxing tobacco, alcohol, and other goods.

It is worth mentioning that the idea of a Pigovian tax is independent of its price elasticity since it tries to raise funds to cover the externality but not reduce its demand or supply per se. This makes it complementary to what we talked about before; if its demand is inelastic, and raising the price will not change consumption that much, the tax would be at least raising funds that can be directed to offsetting the negative impact it has on global warming, through the channels of funding better meat substitutes, campaigns against its consumption, and many others. Proponents of a meat tax argue that it could be a crucial tool in curbing meat consumption, which could reduce greenhouse gas emissions.

### *How much of an increase will cause a meaningful change in greenhouse emissions?*

One of the most renowned papers in this area is the research done by Springmann, M., Robinson, S., Wiebe, K., Godfray, H. C., Rayner, M., & Scarborough, in their paper "Mitigation potential and global health impacts from emissions pricing of food commodities" for the Nature Climate Change Oxford Martin Program, their proposals are a little rash, but their calculations are very insightful. They calculated that the price of GHG emissions is about \$52 per ton of carbon dioxide. Therefore, their proposal for the tax is to increase the price of certain products according to their GHG emissions, meaning that if a product produces a lot of GHG, its price will increase accordingly. Whether this is the best proposal or not, is outside the scope of this paper, but we will use this framework in order to expose how these changes are being proposed currently. Graph A shows the price change for specific products by implementing the tax in USD per kilogram.

Aside from covering its GHG price, they also account that this proposal will shift the demand for these products by using their corresponding price elasticities, and they show a table of how much demand will change with this tax program being implemented. In this chart, they show that the price of beef would have to increase by 40% of its current price, consequently leading to a 13% drop in consumption.



*Figure X: Barplot of price per product together with tax regime, change in price and consumption under the regime, and change in GHG emissions across regions and worldwide*

Moreover, as we can see on the third chart, this proposal will reduce worldwide GHG emissions by approximately one-seventh of current GHG emissions produced by livestock worldwide (one giga-ton

of carbon dioxide specifically), which is a significant amount in the context of curbing emissions. One approach to mitigating the regressive effects of a meat tax is to use the revenue generated to compensate producers and incentivize more sustainable farming practices, such as carbon sequestration. This would help offset the higher costs for farmers and create incentives for them to adopt more environmentally-friendly practices. Additionally, the revenue could be used to support lower-income consumers through targeted subsidies or other social policies, to ensure that the tax burden is not disproportionately borne by those who can least afford it.

However, we believe that the methodology of utilizing the International Model for Policy Analysis of Agricultural Commodities and Trade (IMPACT) to predict their change in consumption, which is an economics formulas of multiple equilibrium, should not be the most reliant methodology to use since it is not considering changes in substitute bundles, and in the country as a whole. Therefore, we would like to generate a model that uses data across different products (substitutes and complements of meat consumption) in real-time in order to predict meat consumption worldwide, and therefore calculate a better estimation of the impact a price increase in the form of a tax will have. Specifically, the problems we have with the IMPACT model approach are:

1. Too complicated for the public to use it
2. Based on economic equilibrium and not very adaptable to constantly changing variables
3. This model system supports longer-term scenario analysis only, through the integration of these multidisciplinary modules
4. Doesn't take into account multiple factors from different products at the same time, only considering a static equilibrium where the markets clear given elasticities of price in the same product (meat in this case)

Hence, we will want our modeling to be adaptable to different frameworks in regard to the size of the tax, easy to use, adaptive to the changing variables of international markets across products, and that can be extrapolated to different nations that don't want to increase prices so abruptly since most of them will probably be searching for an equilibrium in curbing meat demand progressively. We want to clarify this since the proposal we discussed before, where they calculated that the price of GHG emissions is about \$52 per ton of carbon dioxide and wanted to pass that price to the customer to cover the price completely, might be a little too harsh for the public to accept it. In other words, also want our results to be more realistic and adaptive to different scenarios.

Consequently, this project has two main distinct desired outcomes that would serve as helpful tools for policymakers:

1. Develop a predictor of beef and veal consumption levels given agricultural commodity data
2. Perform a feature analysis to investigate which commodities should be prioritized in subsidization changes

## P2 - Methods

### *Data Processing*

For the data collection, it took a few tries in order to get the data in the dimensions, and with the information we wanted. Our first trial was with the World Bank Data, with indicators on ratios per expenditure. The problem we had after cleaning the data was that for most of the columns, data was only available for one year. While that one year could have been sufficient in order to generate predictions of food consumption based on prices, we thought that our insights wouldn't be as valuable if we were only using that one year. Moreover, we also had data for multiple countries, which means there were even some countries where all the columns were missing values. So we had to look for a more complete, cleaner, and longer dataset in order to tune our models and rely on its predictions with more confidence.

Our second trial used OECD data, from all the countries belonging to the organization across the years, only focused on statistics involving meat consumption. This source was particularly better because it included total consumption, producer's prices, and indexes of consumption per capita, across different products aside from meat. OECD countries are required to provide this data in order to account for projects they have as part of the organization to curb certain productions and promote investment in other industries. For the OECD, clean data is the milestone for bigger investments across industries and therefore enforced to be of quality, which is very handy to the analysis we were trying to pursue. Additionally, something excellent about the data is that it also included forecasted values for each variable until 2031 which are according to the projections each nation has committed to as part of their goals as an organization.

However, when trying to work with the dimensionality across countries, things got complicated, and it was not feasible for prediction unless we concatenated by regions or in general. Additionally, all the data was based on the different characteristics of the meat industry during those years, and didn't take into account the different products such as vegetarian food options. So the data was hard to work with since it was multiple countries, and didn't have that many useful variables for prediction.

Finally, we decided to narrow our investigation to the United States across the years, since we consider it to be the country with the most complete data (some countries joined the organization in the early 2000s, so there were a lot of missing values for some variables), and relevant in order to evaluate the global meat market, since the US is one of the biggest producers and consumers of meat worldwide.

One advantage of this approach is that in many papers that have tried to predict meat consumption per capita, GDP per capita is one of the strongest estimators, but in the paper *Are We Approaching Peak Meat Consumption? Analysis of Meat Consumption from 2000 to 2019 in 35 Countries and Its Relationship to Gross Domestic Product* by Clare Whitton, Diana Bogueva, Dora Marinova and Clive J. C. Phillips, they found evidence of a tipping point of around USD 40,000 of GDP per capita, after which increases in economic well-being do not lead to increased meat consumption. Since the current GDP per capita in the US is well above 40,000 USD, there was no real need for adding GDP per capita as a predictor, given that in the case of the US, increases at the current level of GDP per capita do not

increase meat consumption significantly, so we can focus on the rest of the variables as substitutes and complements without having to merge datasets. However, if this analysis was to be done in a poorer nation, economic indicators would be essential for this analysis.

In order to get the dimensions right, we had to transpose the tables that were output by the data source, so that years were the rows, and concatenate the variables together with the food product they were related to, since all the food products had the same variables. In order to do this, we mostly used Excel formulas, and resulted in the final dataset we will use for our predictions.

For our last dataset, we are using data from 1990 to 2031, and this is the list of variables that we are using per product:

- Production (Tonnes, Thousands)
- Ethanol production from commodity (Liters, Millions)
- Biodiesel production from commodity (Liters, Millions)
- Imports (Tonnes, Thousands)
- Consumption (Tonnes, Thousands)
- Ending stocks (Tonnes, Thousands)
- Exports (Tonnes, Thousands)
- Area harvested (Hectares, Thousands)
- Feed (Tonnes, Thousands)
- Food (Tonnes, Thousands)
- Biofuel use (Tonnes, Thousands)
- Other use (Tonnes, Thousands)
- Yield (Tonnes per hectare)
- Producer price (National currency per tonne)
- Human consumption per capita (Kilograms per capita)
- Direct GHG emission (Tonnes of CO<sub>2</sub> equivalent, Millions)
- Total Calorie availability (KCal/person/day)
- Food Protein availability (g/person/day)
- Food Fat availability (g/person/day)

And this is the list of products we have:

- Wheat
- Maize
- Other coarse grains
- Rice
- Distiller's dry grains
- Soybean
- Other oil seeds
- Protein meals
- Vegetable oils
- Molasses

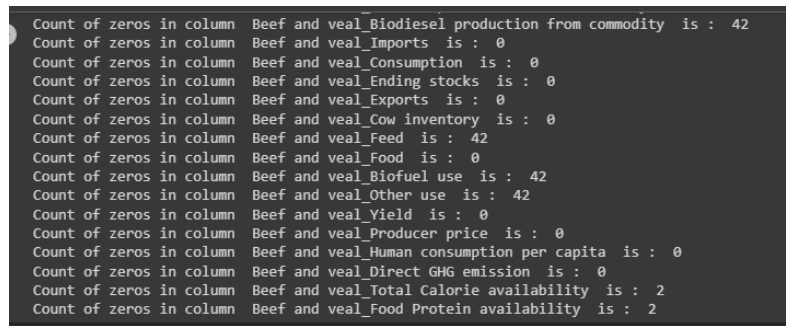
- Sugar
- Raw sugar
- White sugar
- High fructose corn syrup
- Sugar beet
- Sugar cane
- Beef and veal
- Pig meat
- Poultry meat
- Sheep meat
- Milk
- Fresh dairy products
- Butter
- Cheese
- Skim milk powder
- Whole milk powder
- Whey powder
- Casein
- Fish
- Fish from capture
- Fish from aquaculture
- Fish meal
- Fish oil
- Agricultural sector
- Total

Therefore, the dataset's dimensions are 42 rows  $\times$  637 columns. We made the variables from each product differentiable from each other by concatenating the product with its specific variables such that the production of wheat will be "wheat\_production" and so on, in order to have multiple covariates to predict meat consumption from. We used Excel for this task since it allowed for easy concatenation using separation in only one line. Here is what the final result looks like:

	Year	Wheat_Production	Wheat_Ethanol production from commodity	Wheat_Biodiesel production from commodity	Wheat_Imports	Wheat_Consumption	Wheat_Ending stocks	Wheat_Exports	Wheat_Area harvested	Wheat_Feed	...	TUBERS_Producer price
0	1990	74296.96	0	0	990.66	37149.84	23626.21	29110.23	27964.00	13129.00	...	0.00
1	1991	53890.40	0	0	1107.69	30797.63	12927.60	34899.08	23391.00	6654.31	...	109.00
2	1992	67136.43	0	0	1905.12	30686.04	14443.53	36839.58	25414.30	5269.02	...	122.00
3	1993	65220.42	0	0	2961.10	33736.95	15472.30	33415.80	25373.90	7394.59	...	136.00
4	1994	63168.34	0	0	2501.15	35013.38	13787.63	32340.77	25009.66	9373.19	...	123.00
5	1995	59401.64	0	0	1847.97	31026.24	10233.22	33777.78	24645.44	4164.05	...	149.00
6	1996	61981.72	0	0	2512.04	35397.13	12073.02	27256.82	25414.34	8382.53	...	108.00
7	1997	67536.50	0	0	2582.80	34213.23	19663.56	28315.53	25414.34	6817.61	...	124.00
8	1998	69324.60	0	0	2803.25	37588.02	25743.61	28459.77	23876.53	10630.57	...	123.00

Figure 2: Table with covariates by year in the US, example of the data we used for the models

Then, when thinking about null values, we found that there are no null values, but zero values are the ones to worry about in this data since zero is the default value for missing values. After analyzing the number of zeros per column, we see that most of them are in the biodiesel, feed, and other uses, which aren't that meaningful to our analysis, and we will see proof of this in the feature analysis section of the study.



Count of zeros in column	Beef and veal_Biodiesel production from commodity is :	42
Count of zeros in column	Beef and veal_Imports is :	0
Count of zeros in column	Beef and veal_Consumption is :	0
Count of zeros in column	Beef and veal_Ending stocks is :	0
Count of zeros in column	Beef and veal_Exports is :	0
Count of zeros in column	Beef and veal_Cow inventory is :	0
Count of zeros in column	Beef and veal_Feed is :	42
Count of zeros in column	Beef and veal_Food is :	0
Count of zeros in column	Beef and veal_Biofuel use is :	42
Count of zeros in column	Beef and veal_Other use is :	42
Count of zeros in column	Beef and veal_Yield is :	0
Count of zeros in column	Beef and veal_Producer price is :	0
Count of zeros in column	Beef and veal_Human consumption per capita is :	0
Count of zeros in column	Beef and veal_Direct GHG emission is :	0
Count of zeros in column	Beef and veal_Total Calorie availability is :	2
Count of zeros in column	Beef and veal_Food Protein availability is :	2

*Figure 3: Table with the count of zeros per column focused on the variables for beef and veal, where zero is substitute of missing values*

The original source of the data is OECD/FAO (2023), "OECD-FAO Agricultural Outlook (Edition 2022)", OECD Agriculture Statistics (database), <https://doi.org/10.1787/13d66b76-en> (accessed on 27 April 2023). Moreover, you can access the CSV output after the data cleaning process in the following link: [https://docs.google.com/spreadsheets/d/1V9vvkjZoDzWqUk75fVIOVoxHxKBy1\\_rtaTQBGqxjpFA/edit?usp=sharing](https://docs.google.com/spreadsheets/d/1V9vvkjZoDzWqUk75fVIOVoxHxKBy1_rtaTQBGqxjpFA/edit?usp=sharing).

Also, we will attach the notebooks in which we did data cleaning for the three trials described, and notes as the process went by in regard to how we wanted the data to fit in order to make our modeling replicable and understandable from downloading the original data, to the dimensions we adjusted for purposes of modeling. In a nutshell, the target variable will be beef and veal consumption per capita.

## ***Prediction***

For the prediction model, we want to predict the 'Beef and veal\_Human consumption per capita' of the United States by using features of the dataset. Some of the features are imports, exports, GHG emission, and a lot more. Our question is simple: given X, product related features in the dataset, and the target Y, beef and veal human consumption per capita. To estimate the Y, the variable we want to find out by changing X, we will use machine learning models such as: linear regression, KNeighborsRegressor, and MLP Regressor. The loss function, RMSE, calculates the average magnitude of the squared differences between the predicted and actual values.

To evaluate the performance of the models, we use stratified k-folds cross validation on our dataset. We split the data into 10 subsets, equally. By simply repeating the cross-validation procedure 3 times, the estimated performance of the Machine Learning model improves. The mean result across the



folds will be reported multiple times and this result is a more accurate estimate of the true mean performance of the model on the dataset. We chose stratified k-folds cross validation over the traditional train-test split because the dataset contains a limited number of rows. The dataset contains 42 rows, where one row represents each year.

We use the linear regression, multi-layer perceptron regressor (MLP regressor), and k-nearest neighbor regressor (KNN regressor) models in our experiments. All model implementations are based on the Scikit-learn Python library and are hyper parameterized. The best estimator is found by using the best hyperparameters that are found by GridSearchCV.

For linear regression, 'fit intercept' and 'normalize' hyperparameters are given for GridSearch to find the best of them, and essentially the best estimator. This hyperparameter 'fit intercept' determines whether the model includes an intercept term in the regression equation. By default, linear regression models in Scikit-learn include an intercept term, but this hyperparameter can be set to be False if there's prior knowledge that the intercept should be zero. For our dataset, there is no reason for this parameter to be False, but for sanity check, this is included. The hyperparameter 'normalize' determines whether the features in the dataset are normalized before the regression is performed. Normalization involves scaling the features so that they have a mean of zero and a standard deviation of one. Since the features from the dataset have different scales or units, we choose to include this hyperparameter.

For MLP Regressor model specifically, we select the following hyperparameters to be tuned:

- **hidden\_layer\_sizes:** This hyperparameter specifies the number of hidden layers and the number of neurons in each hidden layer. We choose to tune this hyperparameter because the number of hidden layers and neurons can significantly impact the model's performance.
- **Solver:** This hyperparameter specifies the optimization algorithm used to minimize the loss function. We choose to tune this hyperparameter because different solvers can have different convergence rates and can perform better or worse depending on the dataset. We will search over 'lbfgs', 'adam', and 'sgd' solvers. According to sklearn guidebook, the default solver 'adam' works pretty well on relatively large datasets (with thousands of training samples or more) in terms of both training time and validation score. For small datasets, however, 'lbfgs' can converge faster and perform better. For our dataset that contains a very limited number of rows and myriad features, we suspect that default 'lbfgs' won't be the best solver for this particular model.
- **Alpha:** This hyperparameter controls the regularization strength, which can prevent overfitting. We choose to tune this hyperparameter because it is very important we avoid overfitting due to the data problem we have that is mentioned earlier.

For KNeighborsRegressor, we select the following hyperparameters to be tuned:

- **n\_neighbors:** This hyperparameter specifies the number of neighbors to use in the prediction. We choose to tune this hyperparameter because selecting the right value impacts the model's variance and its bias.

- **p:** This hyperparameter specifies the distance metric used to compute the distance between the new sample and the training set samples. We choose to tune this hyperparameter because different distance metrics can lead to different performance results. We compare the two distance formulas, Euclidean ( $p=2$ ) and Manhattan ( $p=1$ ).
- **Algorithm:** This hyperparameter specifies the algorithm used to compute the nearest neighbors. We choose to tune this hyperparameter because different algorithms have different computation times and can perform better or worse depending on the dataset.

For each model, the best estimator and hyperparameters are printed, and the evaluations are done by recording the cross validated RMSE with  $k=10$ , repeated 3 times. Originally, Mean Squared Error (MSE) was going to be used but MSE gives equal weight to large and small errors while RMSE gives more weight to large errors than MSE. Tying this to the question we want to answer (what is the best predictor of beef and veal consumption levels given agricultural commodity data), our recommendation will be given to a government, or big corporations that oversee the entire country, even as large as global scale. We decide large errors are more significant, and those errors cannot be present for making the nationwide decision.

### ***Feature Analysis***

The methods to be used for the feature analysis section were: Pearson correlation between the features and the dependent variable (beef and veal human consumption per capita), examining feature importance in Random Forests, XGBoostRegressor, as well as XGBoostClassifier, and, finally, performing dimensionality reduction (SVD and UMAP) to reduce the number of features used for Random Forests. The majority of the hyperparameters for the models were chosen to be the default implementation parameters provided by their respective libraries. Because the focus of this section was to identify the commodities and approaches that should be prioritized by policymakers when choosing to subsidize products, we examined the effects of changing the number of components for each of the models on overall accuracy on a repeated stratified  $k$ -fold cross-validation with 2 splits and 3 repeats. The decision to do cross-validation was made due to the fact that our dataset was sparse, only containing 42 rows for our 637 features (columns). The evaluation metrics used in this case were accuracy for classification and MSE for regression. It is also worth noting that our  $X$  features were standardized to avoid instability in our models.

Since many of our methods, namely Random Forests, XGBoostClassifier, and the dimensionality reduction techniques were designed to be used with categorical data, we binned our dependent variable into five quantiles. These bins are composed by the dependent variables (beef and veal) by its quantiles. It was decided to use quantiles since the data was not equally distributed, as illustrated in the whiskers and box plot, and histogram below.

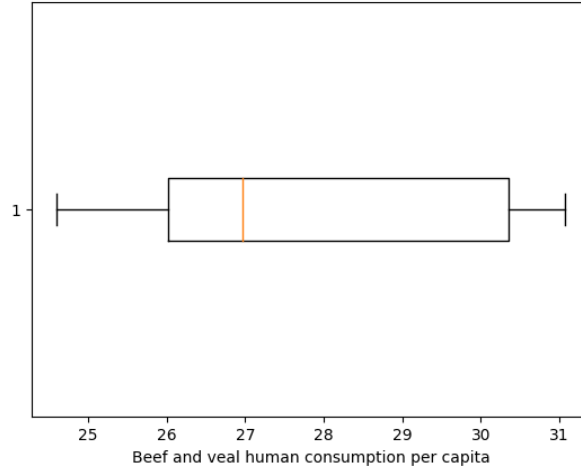


Figure 4: Box and whiskers plot of beef and veal human consumption per capita (in tonnes).

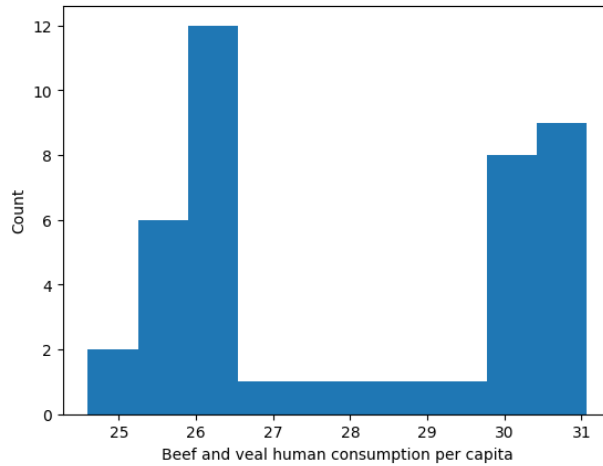


Figure 5: Histogram of beef and veal human consumption per capita (in tonnes).

Additionally, we investigate and evaluate our results visually. For the Pearson correlations, we plot a heatmap of the correlations, such that we could obtain an overall idea of the distribution of highly correlated features in our dataset, and also display bar plots sort by correlations of the feature, to obtain an idea of what the most highly correlated features were and what the distribution of high to low correlated features was in the data. For the Random Forests, XGBoostRegressor, as well as XGBoostClassifier, we displayed similar bar plots of the feature importance, providing insight into the least number of variables that were needed, for the models to obtain their results, in addition to which variables exactly were most useful.

Finally, repeat these plots for Random Forests after the three (specify above) dimensionality reduction techniques, and include two more types of plots. The first one was a box and whiskers plot of accuracy achieve with each dimensionality technique across a varying number of components. The second was a scatter plot of the resulting embeddings return by the dimensionality reduction in 2D and 3D, color-code by label. The objective of plotting the latter was to examine whether the embeddings after dimensionality reduction, correctly capture the distribution of each label in their latent spaces.

## P3- Results

### *Prediction*

We will compare three models and check which hyperparameters work the best for each model. Below are the images for the best estimator using GridSearchCV:

### *Linear Regression*

```
Best Estimator: LinearRegression(normalize=True)
Best RMSE Score: 0.4530652902210355
Best Hyperparameters: {'fit_intercept': True, 'normalize': True}
```

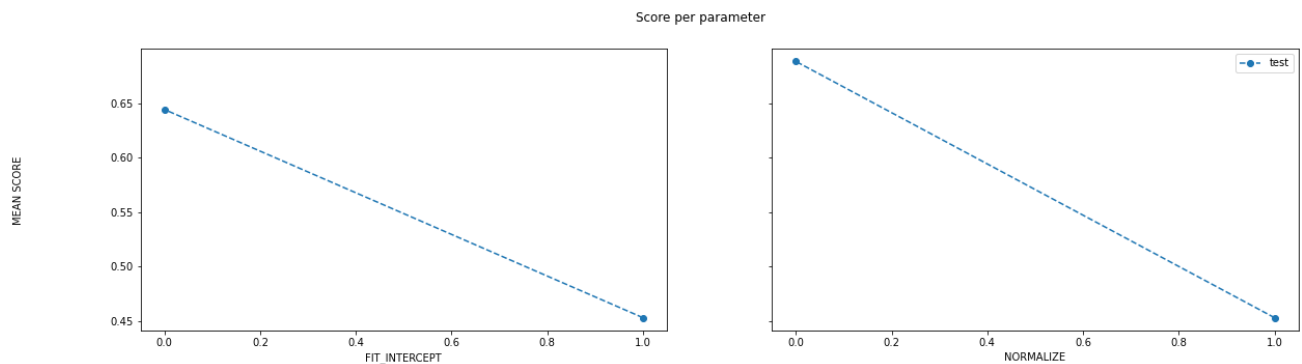


Figure 6: RMSE score per parameters 'FIT\_INTERCEPT' and 'NORMALIZE'

The RMSE score drops significantly when the fit\_intercept and normalize are true. The errors go from 0.65  $\rightarrow$  0.45.

### *KNeighborsRegressor*

```
Best Estimator: KNeighborsRegressor(n_neighbors=3, p=1)
Best RMSE Score: 0.33933333333333332
Best Hyperparameters: {'algorithm': 'auto', 'n_neighbors': 3, 'p': 1}
```

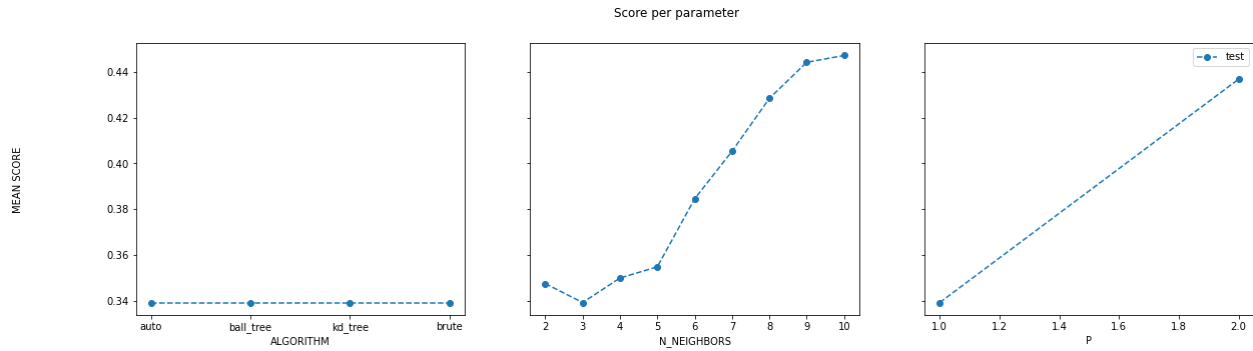


Figure 7: RMSE score per parameters ‘ALGORITHM’, ‘N\_NEIGHBORS’ and ‘P’

The type of algorithm doesn’t change the RMSE score. However, RMSE score goes very high when there’s a high #of neighbors. The best KNeighborsRegressor has the # of neighbors of 3 and uses the Manhattan distance.

### MLP Regressor

```
Best Estimator: MLPRegressor(activation='tanh', alpha=0.0005, hidden_layer_sizes=40,
                             solver='lbfgs')
Best RMSE Score: 0.6039159365762994
Best Hyperparameters: {'activation': 'tanh', 'alpha': 0.0005, 'hidden_layer_sizes': 40, 'solver': 'lbfgs'}
```

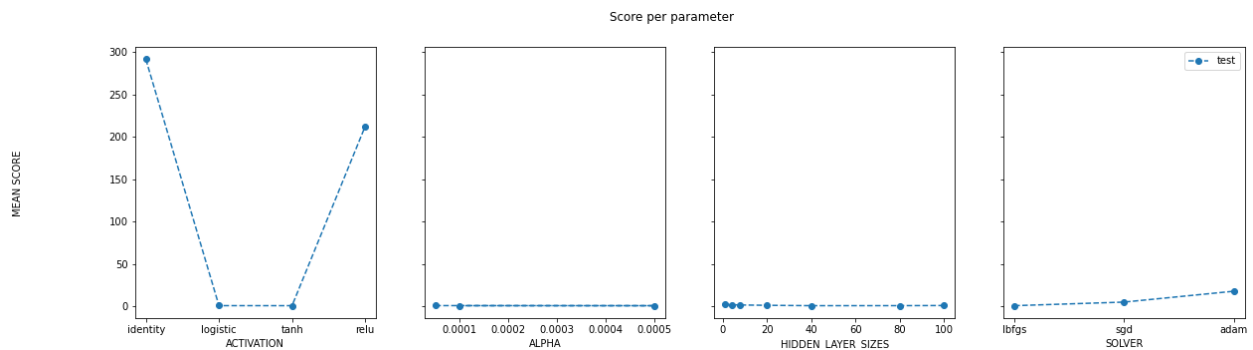


Figure 8: RMSE score per parameters ‘ALGORITHM’, ‘ALPHA’, ‘HIDDEN\_LAYER\_SIZES’ and ‘SOLVER’

For the MLPRegressor model, the most interesting thing is that the hyperparameters like alpha or hidden layers didn’t change the model’s performance. The best model has the activation function of ‘tanh’, alpha of 0.005, hidden layer size of 40, and the solver ‘lbfgs’.

Our results show testing the ability of machine learning models to predict beef and veal consumption from data with the agricultural commodity features. To evaluate, we show cross validated errors RMSE of all models tested on the entire dataset.

The figure below shows that KNeighborsRegressor over performs the other models in the metric we decide to use, RMSE of .33.

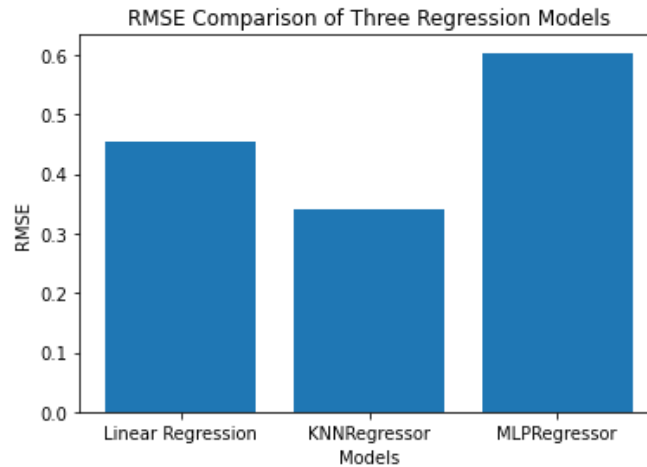


Figure 9: RMSE score comparison for three regression models

The linear regression and MLPRegressor perform comparatively poorly, with a RMSE score of 0.45 and 0.60 respectively.

### Feature Analysis

The heatmap obtained for the Pearson correlations between the features and the dependent variables can be seen below. Its legend can be seen to the right of the plot. The columns are ordered by commodity type and demand/supply indicator (e.g. imports, consumption, exports, etc. in tonnes) as provided in the dataset by default; the x-labels indicate the feature/column index numbers.

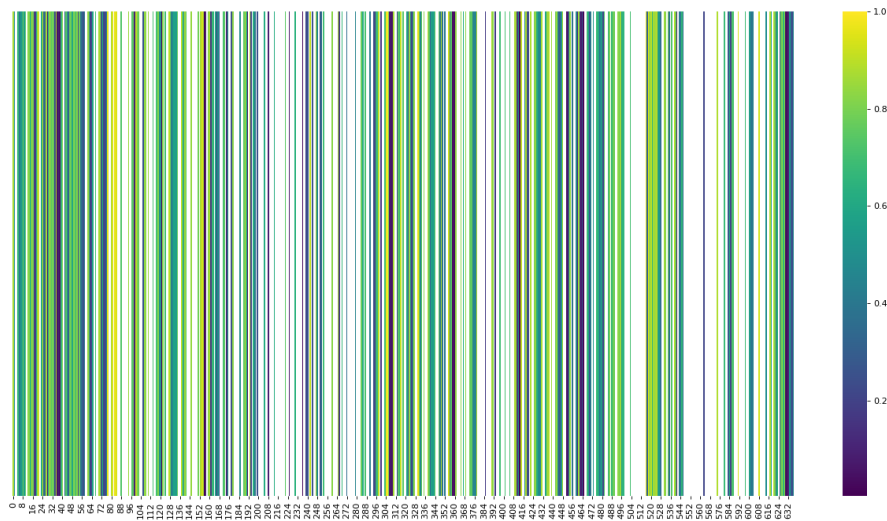


Figure 10: Histogram of correlation between all features and beef and veal human consumption per capita.

The large variation in colors across the heatmap demonstrates a large variation in the types of commodity demand/supply indicators that are most highly correlated with beef and veal consumption per capita. When sorting the features by correlation value, we can see that about half of the demand and

supply indicators for the different commodity types have no correlation with beef and veal consumption, and could thus be disregarded during policymaking. This is illustrated by the following figure.

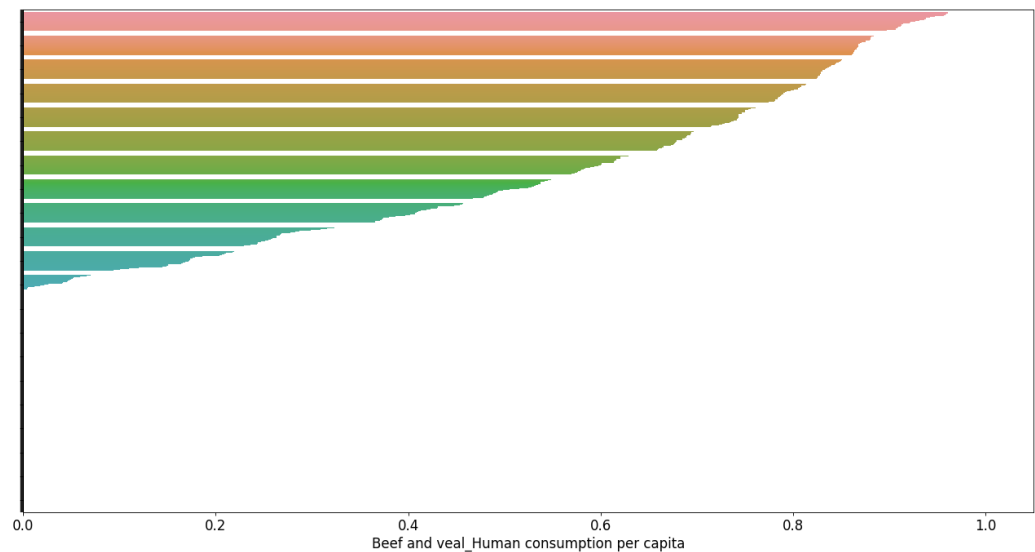


Figure 11: Barplot of correlation between all features and beef and veal human consumption per capita sorted in descending order.

A zoomed-in version of the plot allows us to see that the highest three correlated features are: distiller’s dry grain exports, poultry meat imports, and ethanol production. These are followed by commodities related to other types of grains, as well as meats, and alcohols/fuels.

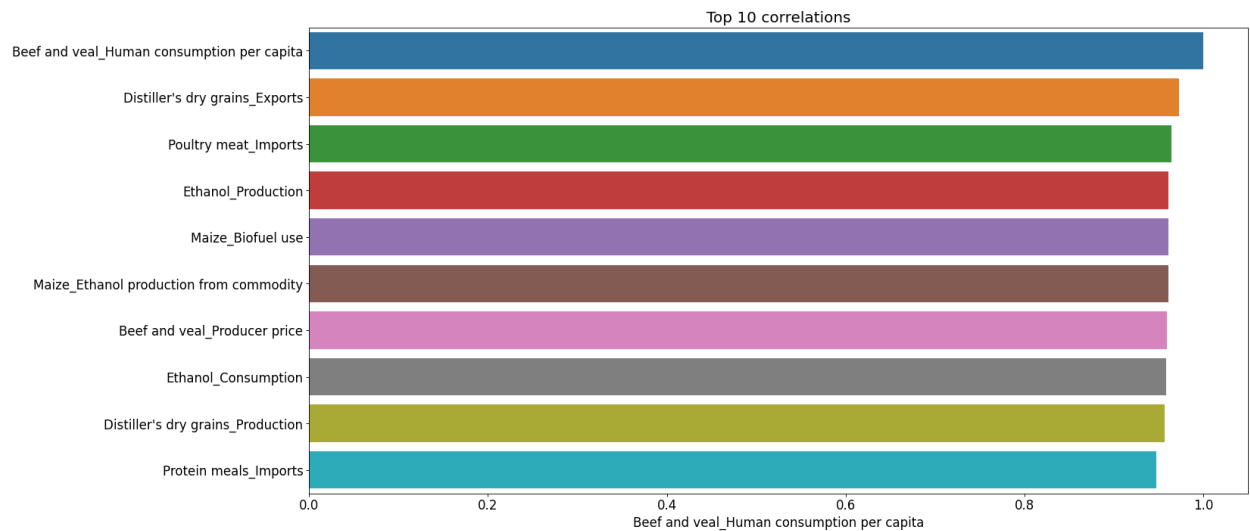
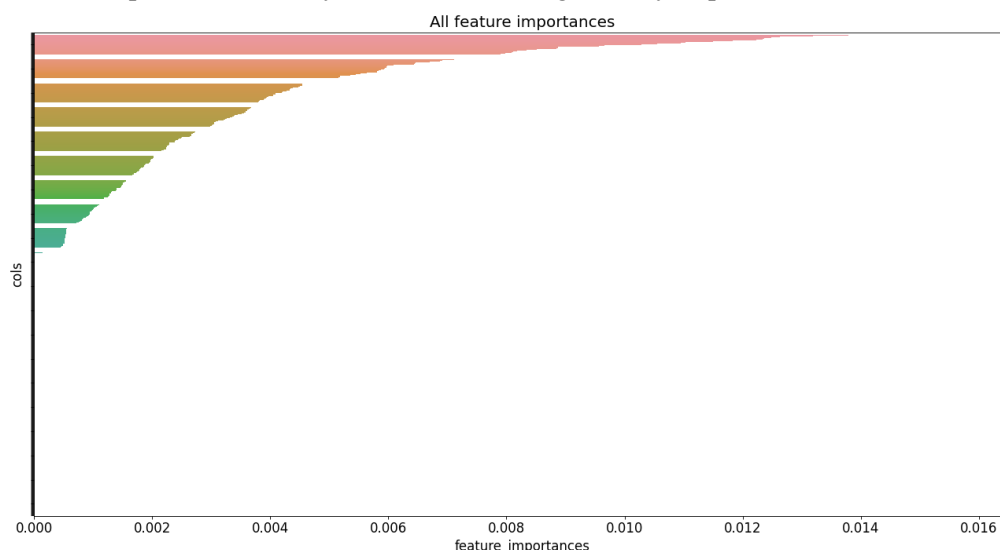


Figure 12: Barplot of correlation between the ten highest correlated features with beef and veal human consumption per capita sorted in descending order.

While these plots can be helpful in identifying highly correlated features and their distribution in the dataset, they alone are not sufficiently reliable for policymaking given that there may be correlations

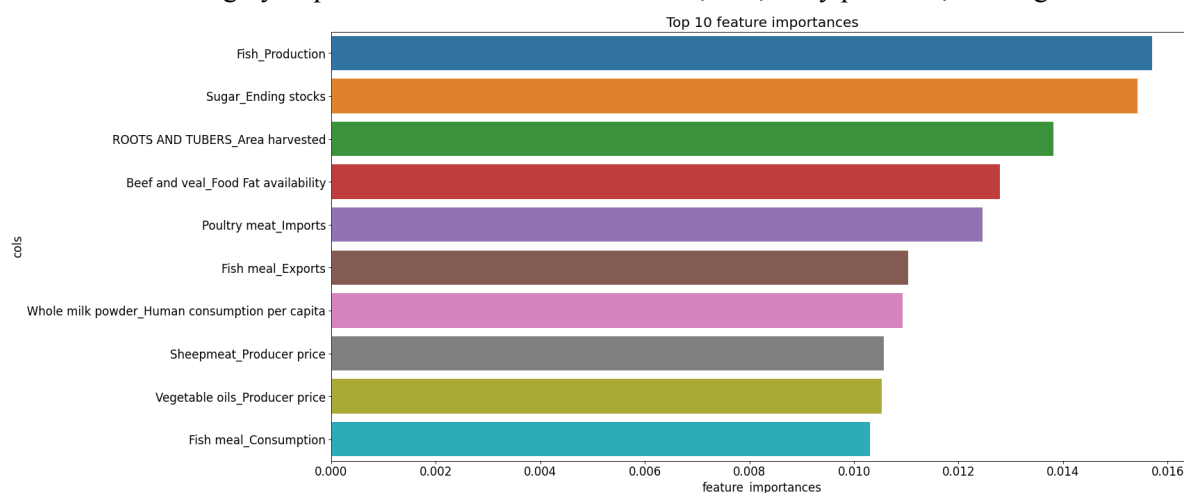
between features themselves, which could make these results very noisy. Thus, we also look at the feature importance results given by other models.

The bar plot of feature importance provided by the Random Forests classifier shows that the model gives higher importance to fewer features as can be seen by the logarithmic-looking curve created by the outline of the plot. Additionally, less than half are given any importance.



*Figure 13: Barplot of feature importance results for all features provided by the Random Forests classifier.*

Zooming into the plot allows us to see that fish production, sugar ending stocks, and the area harvested for roots and tubers were the most important features according to the models. The remaining features deemed highly important are related to other meats, fish, dairy products, and vegetable oils.



*Figure 14: Barplot of feature importance results for ten most important features provided by the Random Forests classifier.*



Now, using the XGBoost regressor gives even greater importance to a substantially smaller subset of features as shown in the next plot.

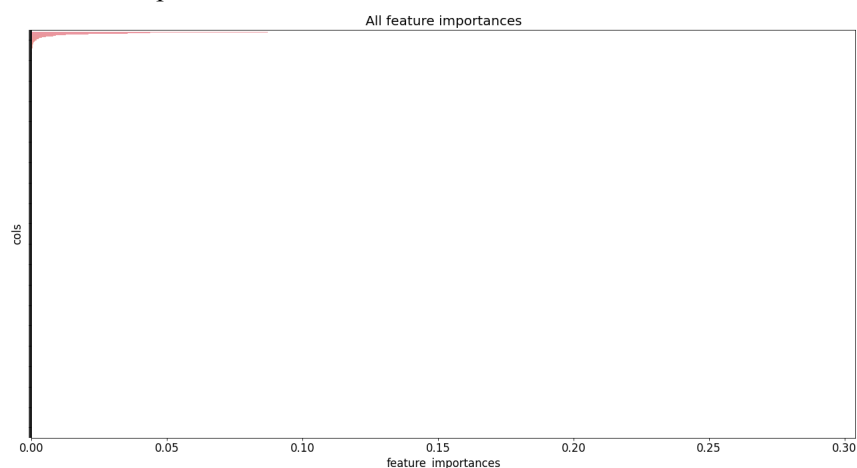


Figure 15: Barplot of feature importance results for all features provided by the XGBoost regressor.

Almost all features are disregarded and zooming into the plot allows one to see, that high fructose corn syrup exports, the area harvested for maize, and the producer price of poultry meat are deemed most important for the XGBoost regressor. Other commodity types include oil seeds, cheese, meats, grains, and biofuel.

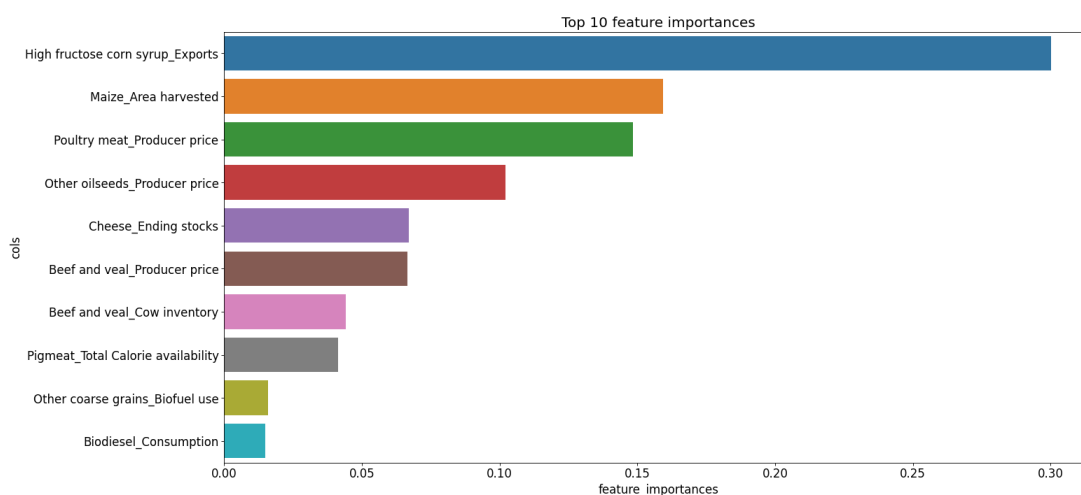


Figure 16: Barplot of feature importance results for the ten most important features provided by the XGBoost regressor.

The XGBoost classifier also allocates relatively high importance to a small portion of features (though greater than its regressor version). The figure below demonstrates this.

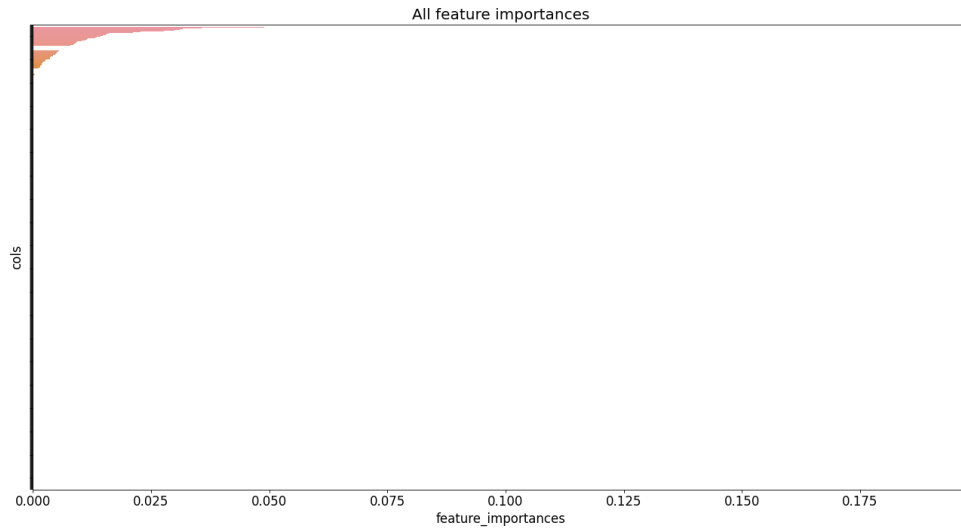


Figure 17: Barplot of feature importance results for all features provided by the XGBoost classifier.

A closer look at the plot (shown below) indicates that cheese imports, protein meal imports, and the producer price of ethanol are of greatest use to the classifier. The other useful commodities include fish, sugars/syrups, oilseeds, roots, and tubers, as well as grains.

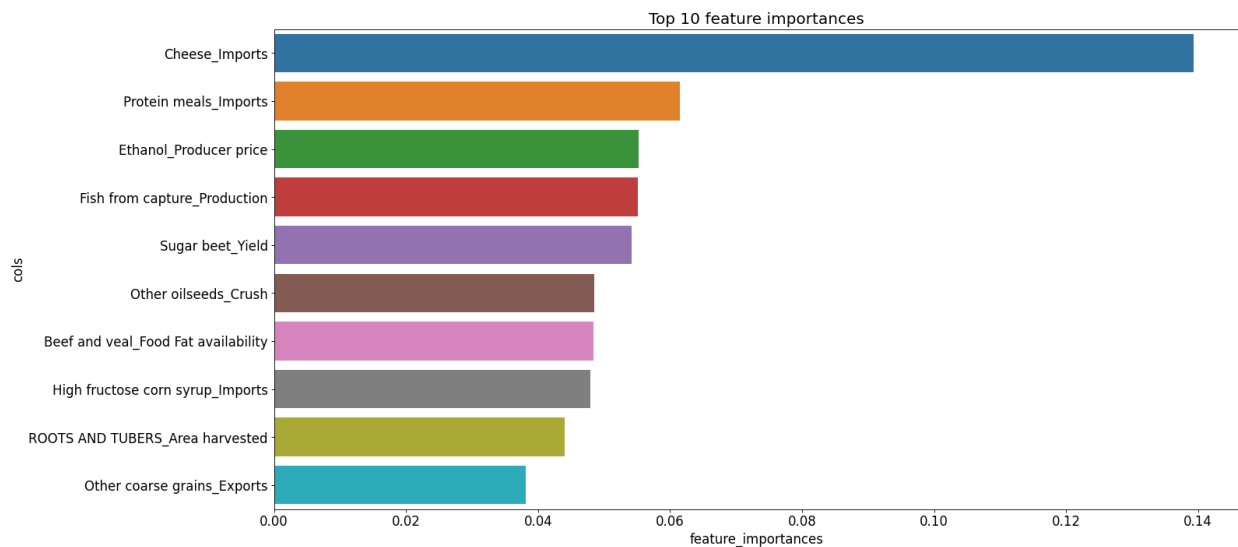


Figure 18: Barplot of feature importance results for the ten most important features provided by the XGBoost classifier.

The following table shows the evaluation metrics obtained by each model and a summary of their most important features.

*Table 1: Summary of results for feature importance models without dimensionality reduction.*

Model	Evaluation Metric Value	Evaluation Metric Type	Features with Highest Importance Values
Random Forests classifier	0.563 (0.115)	Accuracy (std.)	fish production sugar ending stocks area harvested for roots and tubers
XGBoost classifier	0.540 (0.081)	Accuracy (std.)	cheese imports protein meal imports producer price of ethanol
XGBoost regressor	0.751 (0.112)	MSE (std.)	high fructose corn syrup exports, the area harvested for maize producer price of poultry meat

Overall, a similar accuracy score was obtained for the Random Forests and XGBoost classifiers (because we only implemented one regressor, these results are not comparable with the other two models run in this section, but is among the MSE results we achieved in the prediction section). While the accuracy XGBoost is significantly more complex than Random Forests, thus we would expect a higher accuracy score. Nevertheless, it involves several parameters that can alter its performance. It strikes us that doing additional hyperparameter tuning would help XGBoost attain a higher score

Looking particularly, at how varied the types of features selected by the models as most important are, could show that a large variety of different systems could be applied to reduce beef and veal consumption while achieving similar results.

Next, the results obtained for the Random Forests classifier with different dimensionality reduction techniques will be outlined. The accuracy scores achieved by the model while using SVD features with varying numbers of components are displayed in the series of box and whiskers plots below.

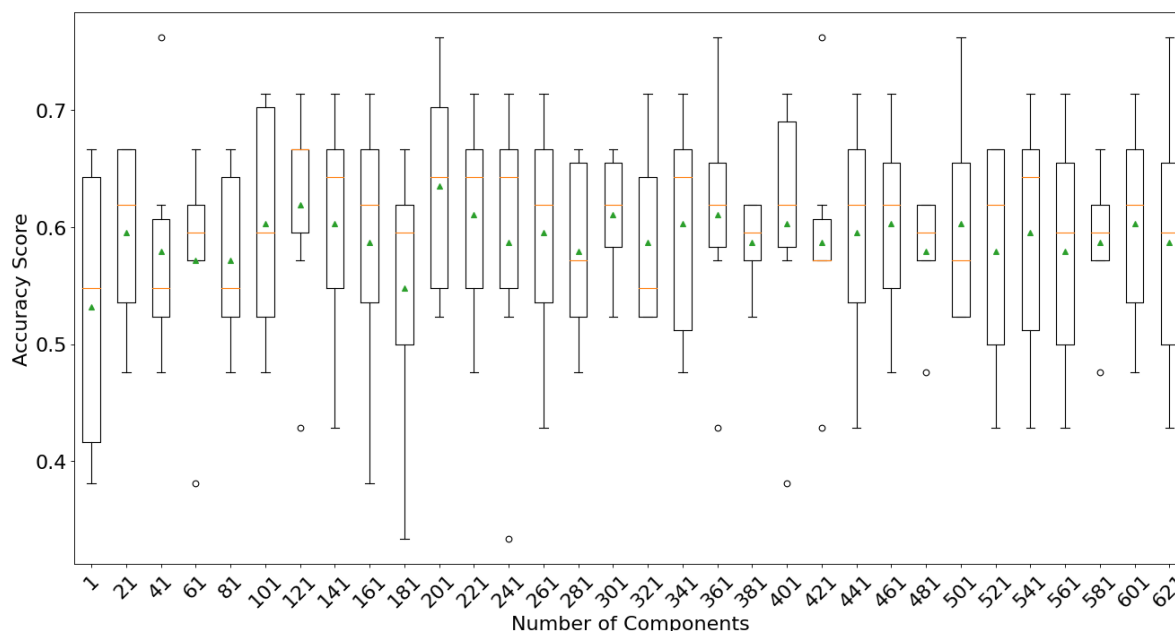


Figure 19: Box and whiskers plots of the accuracy scores achieved by the Random Forests classifier with SVD features while varying numbers of components.

This plot is interesting due to the fact that no clear pattern seems to be displayed as one varies the number of components kept from the data. The fact that accuracy does not increase with a number of features after 21 components, could imply that many of the features being used are redundant, and that very few do, in fact, have an effect on beef and veal consumption levels per capita. It could, however, be argued that higher scores are achieved between 301 components and 461.

The bar plot of explained variance ratios for each component is included below. It is possible to appreciate up to 42 components could be used to capture the entire variance in the dataset.

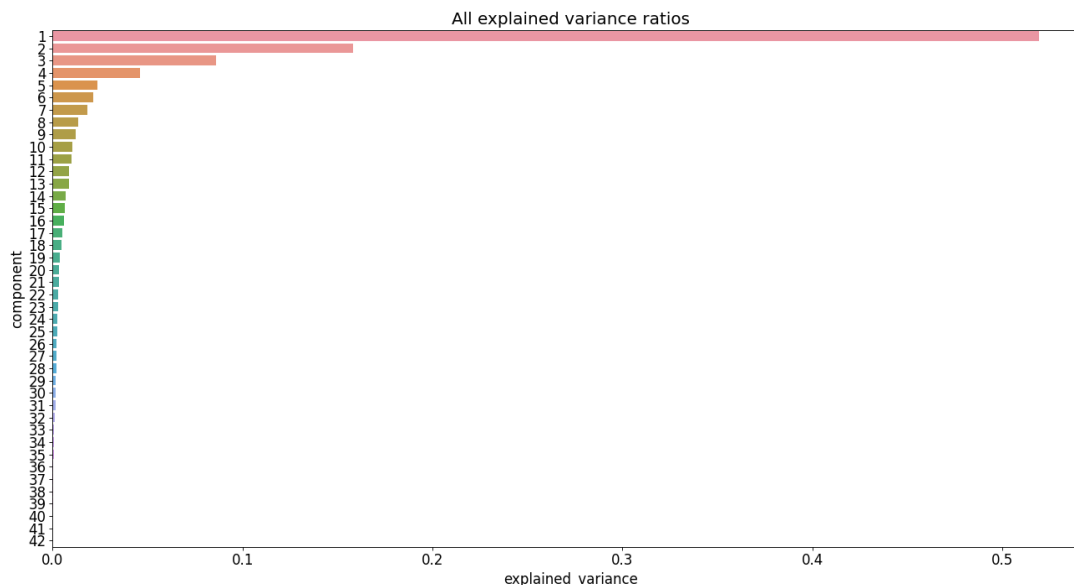


Figure 20: Barplot of the explained variance ratios for SVD components.

The same process for component numbers smaller than 21 in one-step increments was repeated to obtain a clearer idea of what happens when a smaller number of components are used.

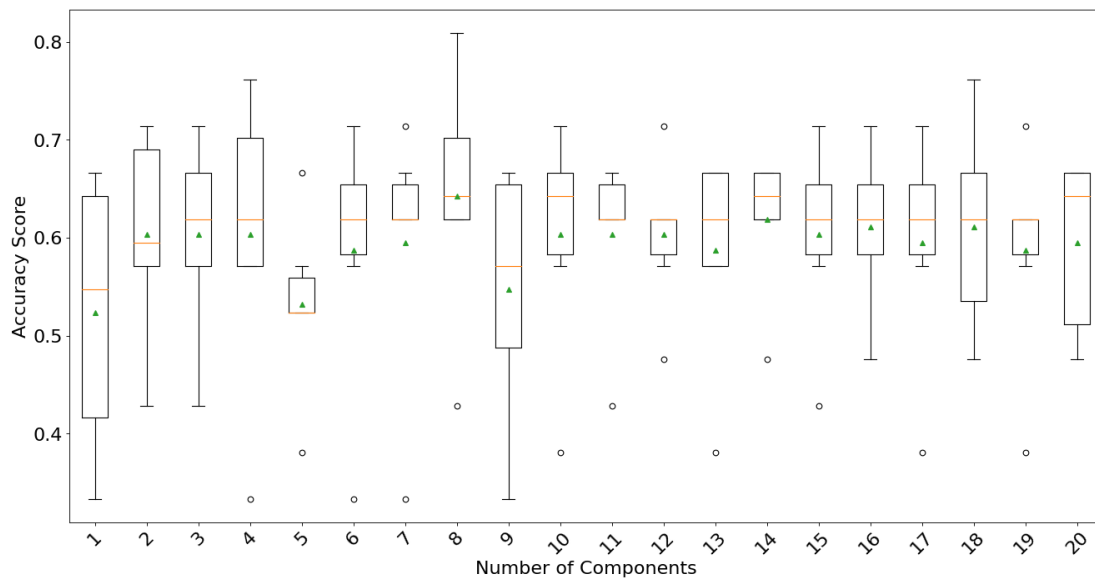


Figure 21: Box and whiskers plots of the accuracy scores achieved by the Random Forests classifier with PCA features while varying numbers of components.

In this case, we also see similar results across component numbers higher than 1 in terms of both median score and variance. Due to algorithmic constraints (UMAP only allowing for up to 20 components to be used due to our number of features to samples ratio), we could only run our experiment for the smaller number of components.

Finally, UMAP scores can be seen in the next figure.

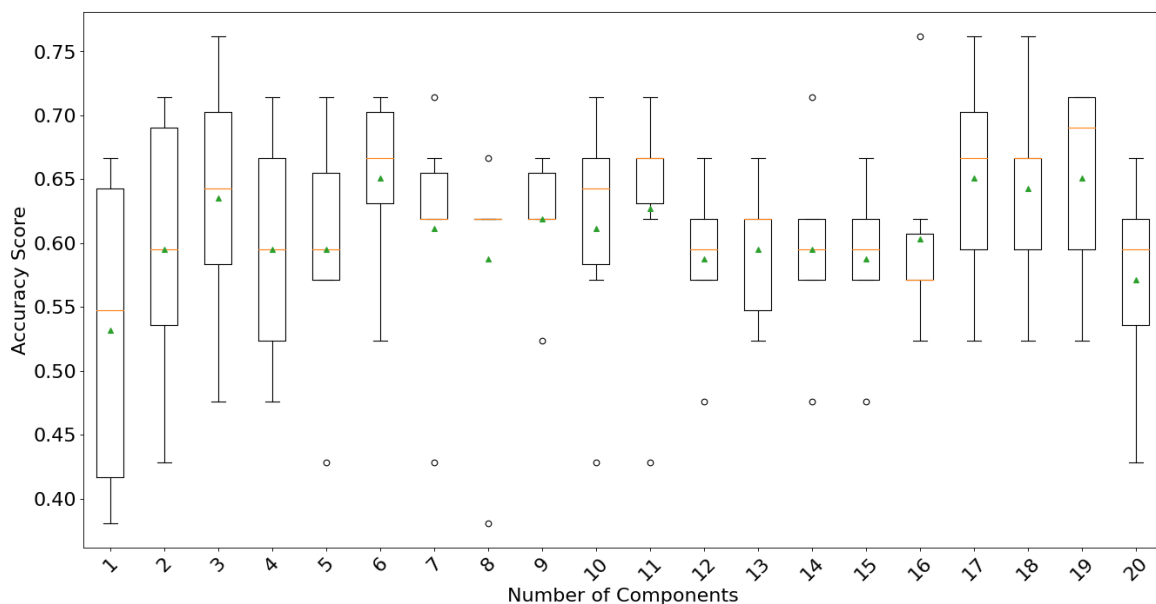


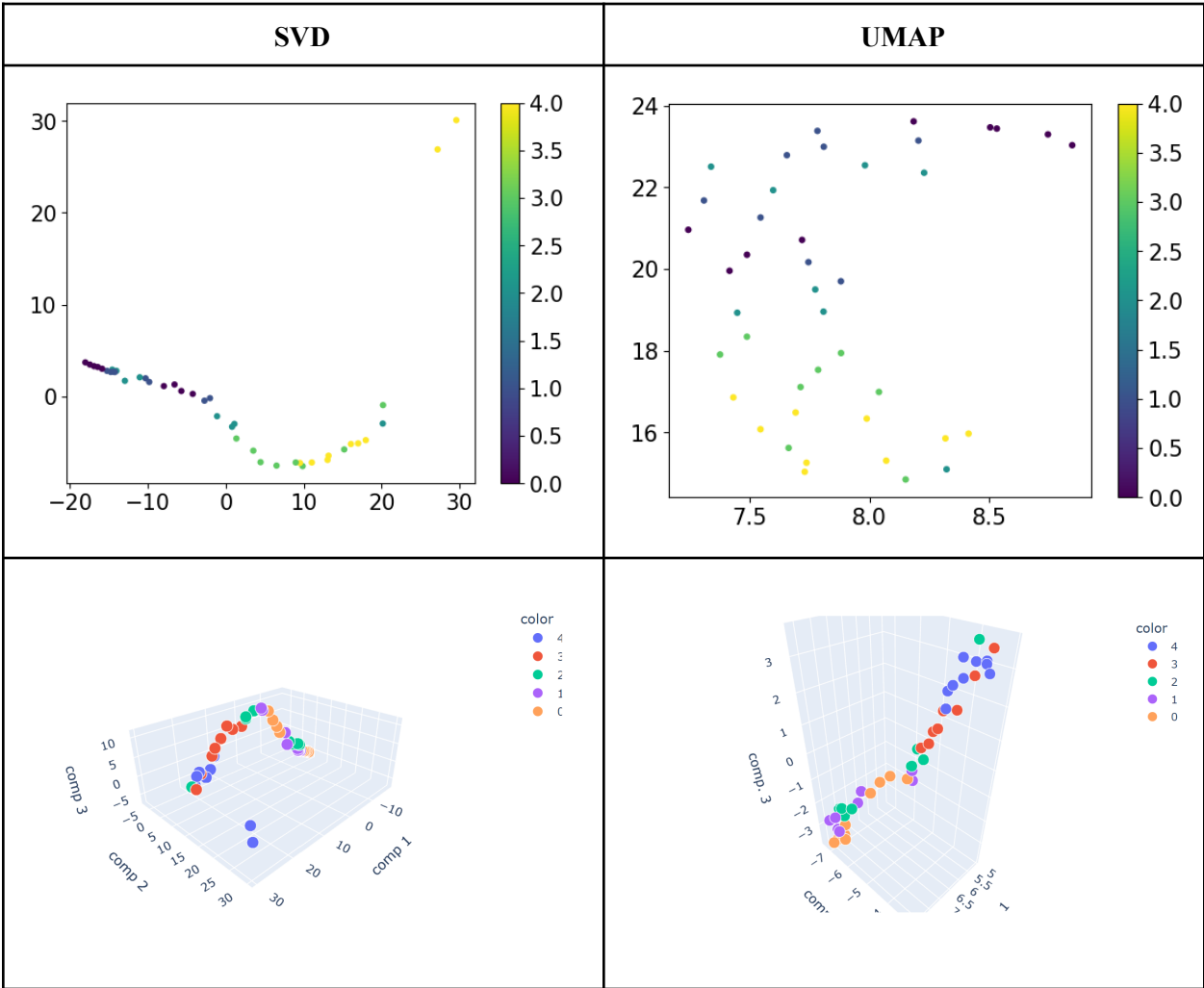
Figure X: Box and whiskers plots of the accuracy scores achieved by the Random Forests classifier with PCA features while varying numbers of components.

We can see that after three components, there seems to be no significant increase in accuracy.

As a non-linear manifold learning technique based on metric spaces rather than feature-based data, UMAP does not have a notion of explained variance the way algorithms like SVD do. Thus, we could not perform this analysis with UMAP embeddings.

Nevertheless, plotting the embeddings resulting from each dimensionality reduction technique in 2D and 3D proved quite interesting, as we could see clear clusters by label in the latent spaces of each method, demonstrating, especially in the 2D case, that groups can be captured reasonably well with some nuances and overlapping (likely explained by the binning of our dependent variable) in two or three dimensions.

Table 2: 2D and 3D embeddings of SVD and UMAP components.



# Discussion

Notwithstanding, such a model can also be used to predict demand given a price decrease of plant-based meat substitutes. But this might be a task outside the scope of our primary investigations, but it is the next route if we find good results when predicting meat demand.

## *Prediction*

To summarize, the three hyper parameterized regressors' RMSE scores tell us that KNeighborsRegressor is the best predictor we developed. Therefore, we recommend the government to consider using KNeighborsRegressor predictor over the over two models to predict the beef and veal consumption while varying the other commodities' units/price/demand. Given more time, it will be worthwhile to develop every regression model that exists to find the absolute best predictor that the world can use to predict food consumption.

## *Feature Analysis*

Due to the fact that evaluation of the Pearson correlation between all variables and the dependent variable does not eliminate inter-feature correlations, that could lead to model instability, we believe the results of the latter methods are more reliable for real-world applications.

When using Random Forests, and XGBoost for feature importance we could see that they all highlight the effects of a variety of features, including commodities like meats, fish, vegetable oils, sugars/syrups, oil seeds, roots, and tubers, grains, dairy products, protein meal, maize, cheese, meats (mainly poultry and pig meat), and alcohols/fuels on beef and veal consumption per capita.

The fact that the types of features selected by the models are relatively varied, might suggest that there are multiple ways to effectively decrease beef and veal consumption. Our recommendation would be to adhere to the top features recommended by a single chosen high-achieving prediction model, such that when policymakers have designed a new subsidization model, its effects on beef and veal consumption can be simulated before implementation using this same model.

Finally, our dimensionality reduction techniques showed that the dimensionality of the features can be reduced reasonably well, while still explaining the majority of the variance captured in the dataset, with a minimum of two features. This is not only helpful to policymakers who can focus on taxing/subsidizing two to three features, but also computationally efficient when testing the effects of a new policy using the prediction methods described above.

It is worth noting that the reason, we decided to use different models for our Predictor Development and Feature Analysis is that models like Random Forests are better suited for feature

importance analysis while models like Linear Regression, KNN Regression, and MLPs are better suited for continuous data prediction, unless using complex methods such as LIME or SHAP. We decided it was best to give follow the methods that were best suited for each tool. Nevertheless, we would encourage future work to evaluate the feature analysis results directly on our prediction models.

## ***Data-wise***

Based on papers like “Are We Approaching Peak Meat Consumption? Analysis of Meat Consumption from 2000 to 2019 in 35 Countries and Its Relationship to Gross Domestic Product” by Clare Whitton, GDP per capita is a very important, but only for developing nations. So for better extrapolation, we would have liked to make the same analysis in a developing country, and add GDP per capita as a variable and see if the model would perform better.

Connecting with the need of testing the model in different countries, it can also be the case that we did not see small RMSE with our models since for big countries like the US, a model like the International Model for Policy Analysis of Agricultural Commodities and Trade (IMPACT) is appropriate due to the absence of market inefficiencies, abundance, and international price setting by the country. In other words, meat consumption in big countries isn't really related to the prices and consumption of other products, since there is enough money in most of the cases for people to over-buy and not simply substitute beef for a plant-based product. So in a small country, we will predict that this approach would be more feasible, but it is outside the scope of our study.

## **References**

*- bibliography of sources you used and cited ~3 pages min for first five sections, plus extra pages for references and figures. Plus pre-run Python notebook (should replicate your figures from data). Due May 5th by midnight*

Caleb Robinson, Bistra Dilkina, Jeffrey Hubbs, Wenwen Zhang, Subhrajit Guhathakurta, Marilyn A. Brown, Ram M. Pendyala, Machine learning approaches for estimating commercial building energy consumption, Applied Energy, Volume 208, 2017, Pages 889-904, ISSN 0306-2619, <https://doi.org/10.1016/j.apenergy.2017.09.060>.

EPA, Environmental Protection Agency: “Inventory of U.S. Greenhouse Gas Emissions and Sinks,” <https://www.epa.gov/ghgemissions/inventory-us-greenhouse-gas-emissions-and-sinks>.

IPCC, 2019: Summary for Policymakers. In: Climate Change and Land: an IPCC special report on climate change, desertification, land degradation, sustainable land management, food security, and greenhouse gas fluxes in terrestrial ecosystems [P.R. Shukla, J. Skea, E. Calvo Buendia, V. Masson-Delmotte, H.- O. Pörtner, D. C. Roberts, P. Zhai, R. Slade, S. Connors, R. van Diemen, M. Ferrat, E. Haughey, S. Luz, S. Neogi, M. Pathak, J. Petzold, J. Portugal Pereira, P. Vyas, E. Huntley, K. Kissick, M. Belkacemi, J. Malley, (eds.)]. In press.



Neff, Roni A, et al. "Reducing Meat Consumption in the USA: a Nationally Representative Survey of Attitudes and Behaviours." *Public Health Nutrition*, vol. 21, no. 10, 2018, pp. 1835–1844., doi:10.1017/S1368980017004190.

OECD, 2023, Meat consumption (indicator). doi: 10.1787/fa290fd0-en (Accessed on 21 March 2023)

"OECD-FAO Agricultural Outlook (Edition 2022)", OECD Agriculture Statistics (database), <https://doi.org/10.1787/13d66b76-en>