Melih Elibol – Project Final

Aim:

In this project, I aim to examine the relationship between the prevalence of obesity and the prevalence of undernourishment throughout the world.

Introduction to Datasets:

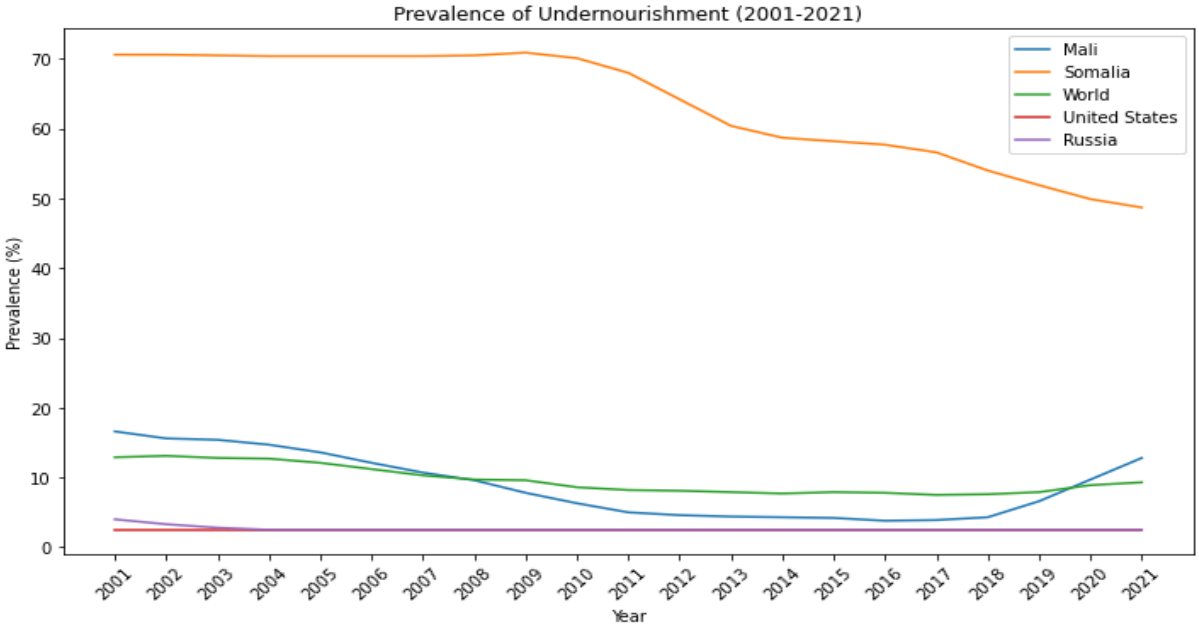
I utilized four different datasets for this project. The first one indicates the prevalence of undernourishment between 2001-2022 in each country. The second one indicates death rates due to malnutrition between 1990-2019 in each country. The third one shows the prevalence of obesity among adults (18+ years) between 1975-2016 in each country. The last dataset shows death rates due to obesity between 1990-2019 in each country. By using these datasets, I attempted to generate insights to test my hypothesis.

Hypothesis

Null Hypothesis (H0): As the obesity rate increases worldwide, the rate of undernourished people is also increasing.

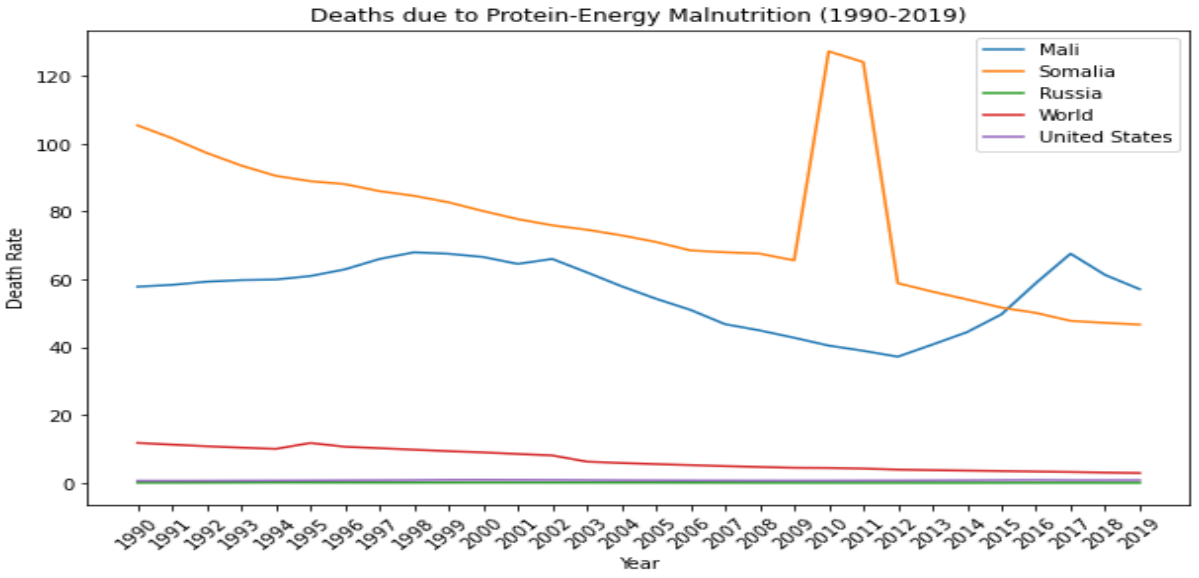
Alternative Hypothesis (H1): As the obesity rate increases worldwide, the rate of undernourished people is decreasing or remaining unchanged.

To begin with, I examined the data that shows the prevalence of undernourishment between 2001-2022. In this dataset, I found out that especially in some parts of Africa, there were very high rates occurring. After that, I wanted to compare these rates with the world's situation and some other parts of the world, especially the ones that are developed countries such as the USA and Russia.

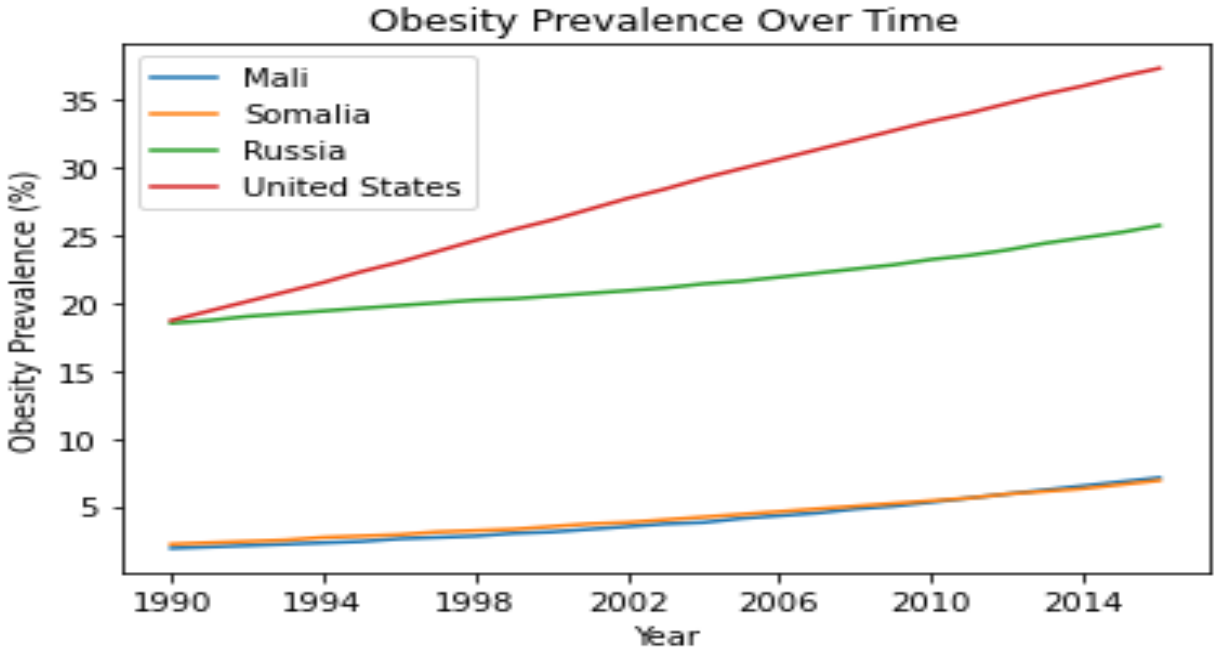


Here, we can clearly see that compared to the USA, Russia, or the world in general; Somalia shows a great peak for the prevalence of undernourishment. Yet, there is an obvious decrease in Somalia in the last years. This is important since later I'll be talking about this situation of Somalia.

After that, I also wanted to endorse my hypothesis, so I searched for deaths due to protein-energy malnutrition dataset. This dataset corresponds to the previous one, so we are on the right track.

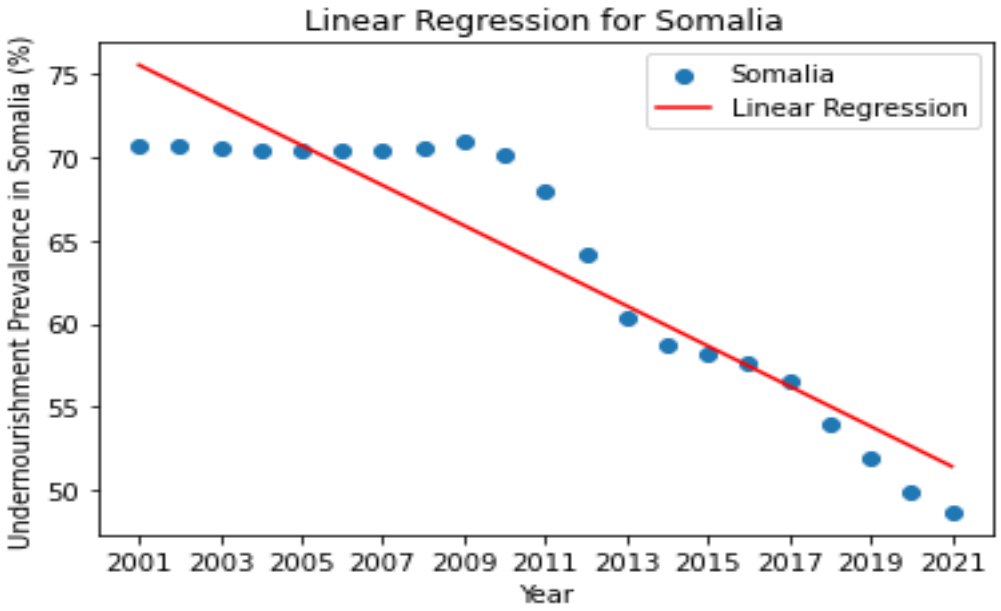


Now, I need to move forward and look for the obesity part of my hypothesis. So, by using the same countries that I have chosen before (above you can find those countries), I examined the dataset that consists of obesity prevalence over time. Beforehand, I had expected Russia, USA, and so on to have very high rates as well as countries like Somalia to have very low rates. When I saw them in the graph, I realized that although countries like Somalia cannot even be compared to the USA or Russia, there was an increase in their obesity prevalence. So, I was kind of surprised at first because when I was constructing my null hypothesis, I would not expect this. In fact, I would expect it to be otherwise. Hence, our hypothesis is about to shape into its last version.

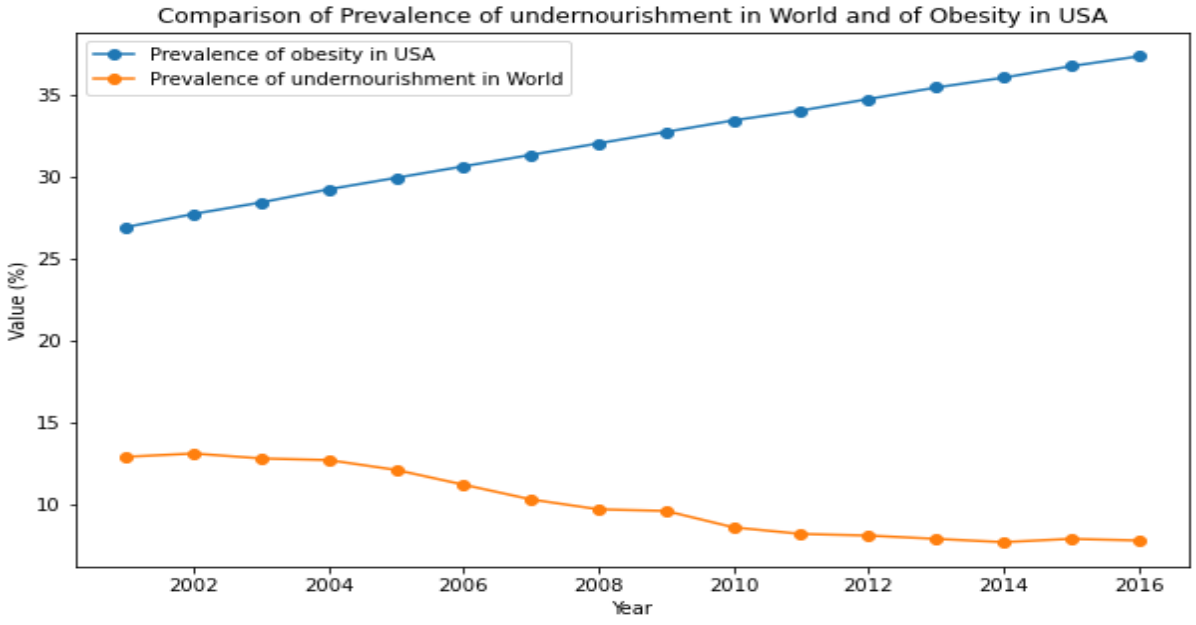


So far, what I have all done was to create a ground for my hypothesis. I especially underlined Somalia because when I was examining these datasets, I found out that Somalia had one of the greatest prevalences of undernourishment and on the other hand, I especially benefitted from the data of the USA and Russia due to their increase in prevalence of obesity.

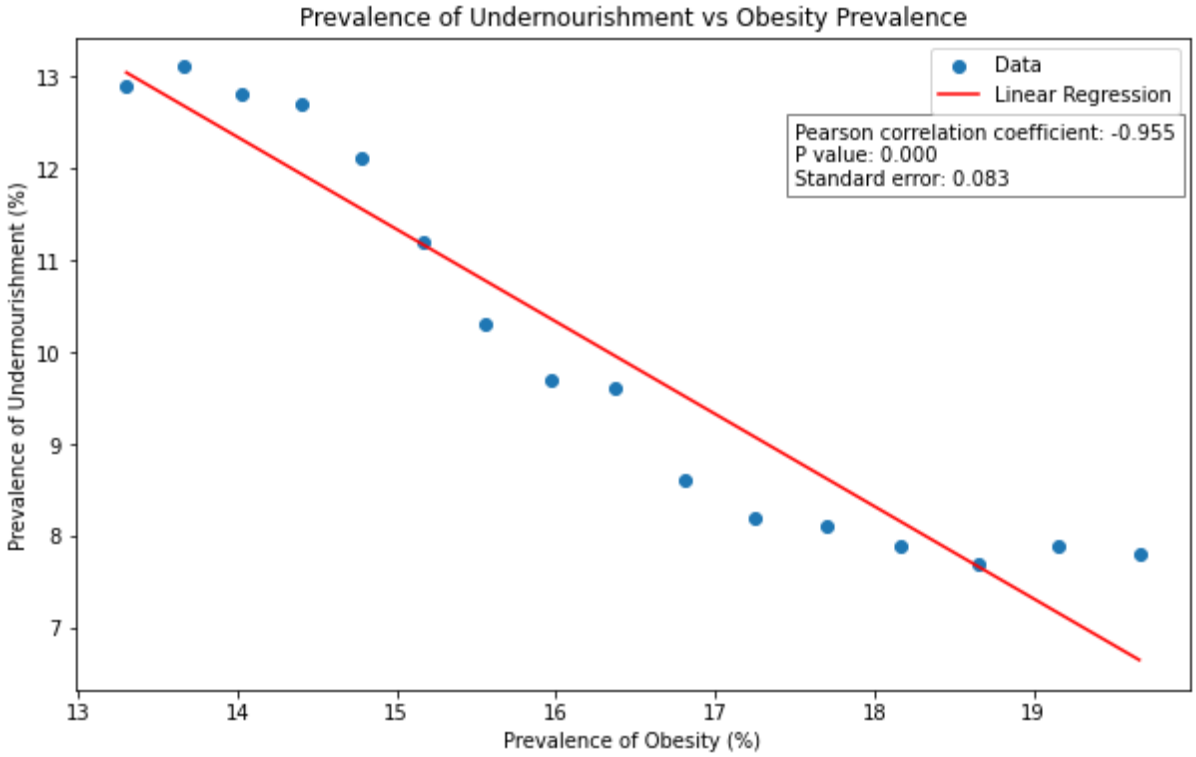
Before moving onto the hypothesis stage, I wanted to highlight another point: the trend in undernourishment prevalence in Somalia. Even for a country like Somalia which shows great rates when it comes to undernourishment prevalence, there is a negative trend, i.e. undernourishment prevalence has a tendency to decrease. This is very crucial because it may support not the null hypothesis but the alternative one. Keep in mind that Somalia was the country that I have chosen to be showing the most effects of undernourishment prevalence throughout the world. So, it is highly above average as you can see from the previous graphs.



As I'm comparing undernourishment and obesity, I believe that it is a wise idea to show the other side of the story, which is how things are changing in the USA (which was showing one of the highest obesity rates). Here, you can see that although obesity in the USA increases, there is a decrease in the undernourishment prevalence in the world. This supports the alternative hypothesis.



Now, time has come to talk about hypothesis testing. As we can see from the below graph as well, the prevalence of undernourishment and the prevalence of obesity are negatively correlated, meaning that although our null hypothesis is wrong, our alternative hypothesis seems to be true. It occurs mainly due to the -0.955 Pearson correlation coefficient. As you can see, the p-value is so small that when I round it to three decimals, it rounds to 0.000. Hence, we can conclude that although there is an increase in the obesity rate in the world, the undernourishment rate seems to be decreasing.



Other Hypothesis Testing Criteria:

T-test:

- T-test statistic: 8.571464329310436
- P value of the t-test: 1.456563270708865e-09
- Since the p-value is extremely small (1.46e-09), much smaller than the commonly chosen threshold of 0.05. This suggests strong evidence against the null hypothesis, meaning that it supports alternative hypothesis.

Spearman Correlation Coefficient:

- The Spearman correlation coefficient is -0.986, indicating a very strong negative monotonic relationship between the variables.
- The p-value associated with the coefficient is 2.70e-12, which is extremely small.
- With such a small p-value, it suggests strong evidence against the null hypothesis and supports the alternative hypothesis.

Kendall’s Tau Correlation Coefficient:

- The Kendall's Tau correlation coefficient is -0.946, indicating a very strong negative association between the variables.
- The p-value associated with the coefficient is 3.53e-07, which is very small.
- With such a small p-value, it suggests strong evidence against the null hypothesis and supports the alternative hypothesis.

If we have to make a prediction about how the future would look like, we should look for linear regression trend. It seems that when the prevalence of obesity is increasing, the prevalence of undernourishment seems to be decreasing. That’s why one may conclude that in the upcoming years, the prevalence of obesity will increase whereas the prevalence of undernourishment will decrease.

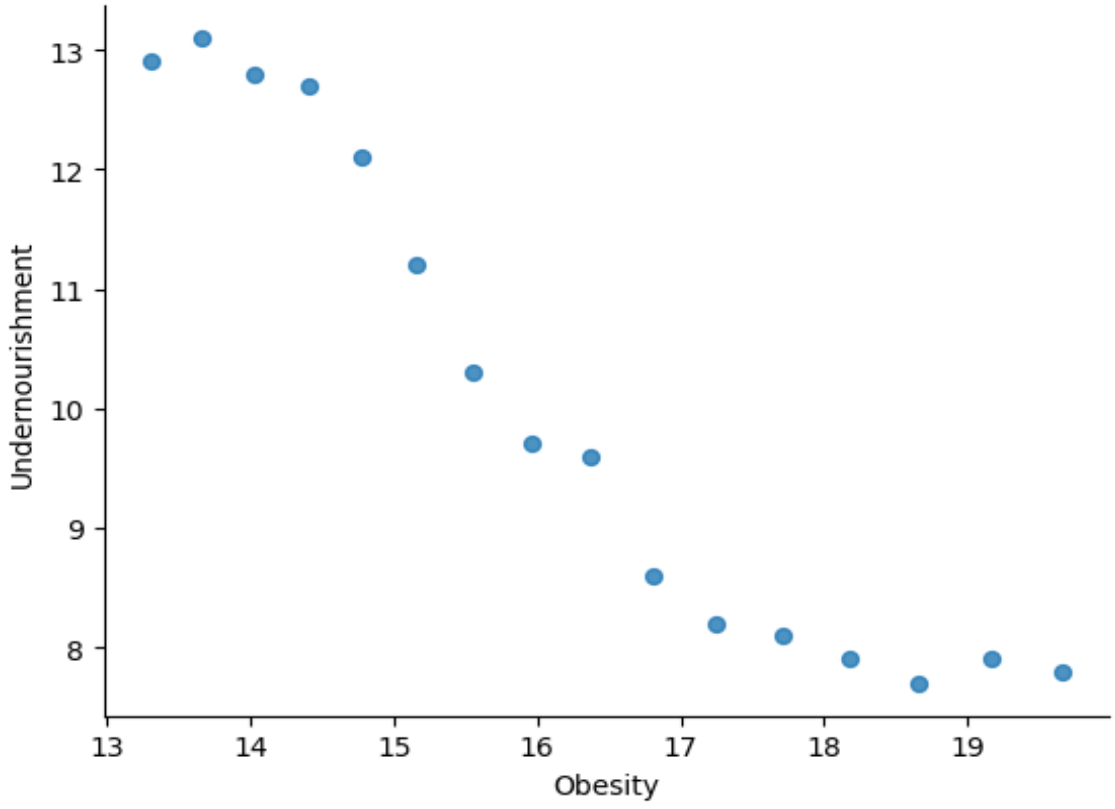
Machine Learning Part

1-General Introduction to Machine Learning & Importance of Topic:

Machine learning is a powerful tool that enables computers to learn from data and make predictions or decisions without being explicitly programmed. It plays a crucial role in various fields, including healthcare, finance, and environmental science. In this project, machine learning techniques are employed to predict undernourishment rates based on obesity rates, leveraging the relationship between these two factors. This project can be used in order to inform public health initiatives and interventions regarding the related topic. Hence, it carries high importance.

2-Introduction to Merged Data:

	Year	Obesity	Undernourishment
0	2001	13.301523	12.9
1	2002	13.660914	13.1
2	2003	14.028426	12.8
3	2004	14.406091	12.7
4	2005	14.779695	12.1
5	2006	15.160914	11.2
6	2007	15.556853	10.3
7	2008	15.965990	9.7
8	2009	16.377157	9.6
9	2010	16.812183	8.6
10	2011	17.252284	8.2
11	2012	17.706091	8.1
12	2013	18.171066	7.9
13	2014	18.653807	7.7
14	2015	19.160406	7.9
15	2016	19.665482	7.8



3-Overview of Models

1. k-Nearest Neighbors (kNN)

Why kNN was Chosen:

- kNN is chosen for its simplicity and ease of implementation.
- It does not assume any underlying data distribution and can capture complex relationships in the data.
- In our project, where we aim to predict undernourishment rates based on obesity rates, the kNN algorithm is suitable for capturing the relationship between these variables without making strong assumptions about the data distribution.

2. Random Forest

Why Random Forest was Chosen:

- Random Forest is chosen for its robustness and ability to handle high-dimensional data with complex interactions.
- It can capture non-linear relationships between input features and target variables effectively.
- In our project, where the relationship between obesity rates and undernourishment rates may be non-linear and complex, Random Forest is well-suited to capture these patterns and provide accurate predictions.

4-Model Training and Evaluation

1. k-Nearest Neighbors (kNN):

- The model is trained by fitting the kNN regressor to the training data using the “fit()” method.
- To evaluate the performance of the kNN model, we employed cross-validation.
- We compute the Mean Squared Error (MSE) for each fold and average the MSE values to obtain an overall evaluation metric for the model.

2. Random Forest:

- For Random Forest, the training process involves building an ensemble of decision trees based on bootstrapped samples of the training data.
- Each decision tree is trained independently on a subset of the features and data points.
- RandomizedSearchCV searches through a specified range of hyperparameter values and selects the combination that yields the best performance based on a specified evaluation metric (e.g., MSE).
- In our project, we tune hyperparameters such as the number of estimators, maximum depth, minimum samples split, and minimum samples leaf to improve the model's performance.

- To evaluate the performance of the Random Forest model, we employed cross-validation.
- By splitting the dataset into multiple folds and training the model on different subsets of the data, cross-validation provides a more robust estimate of the model's performance.

5-Results and Comparison:

1-kNN:

```
kNN Mean Squared Error for 2 number of neighbors with 5-fold cross-validation: 0.5201666666666679
kNN Mean Squared Error for 3 number of neighbors with 5-fold cross-validation: 0.9103888888888909
kNN Mean Squared Error for 4 number of neighbors with 5-fold cross-validation: 0.9602916666666672
kNN Mean Squared Error for 5 number of neighbors with 5-fold cross-validation: 1.3970066666666674
kNN Mean Squared Error for 6 number of neighbors with 5-fold cross-validation: 1.6696481481481478
kNN Mean Squared Error for 7 number of neighbors with 5-fold cross-validation: 2.300404761904761
kNN Mean Squared Error for 8 number of neighbors with 5-fold cross-validation: 2.7493020833333333
kNN Mean Squared Error for 9 number of neighbors with 5-fold cross-validation: 3.5438991769547306
kNN Mean Squared Error for 10 number of neighbors with 5-fold cross-validation: 4.418519999999999
kNN Mean Squared Error for 11 number of neighbors with 5-fold cross-validation: 5.042670798898067
kNN Mean Squared Error for 12 number of neighbors with 5-fold cross-validation: 5.804152777777777

Best kNN Model
Number of Neighbors: 2
Mean Squared Error: 0.5201666666666679
```

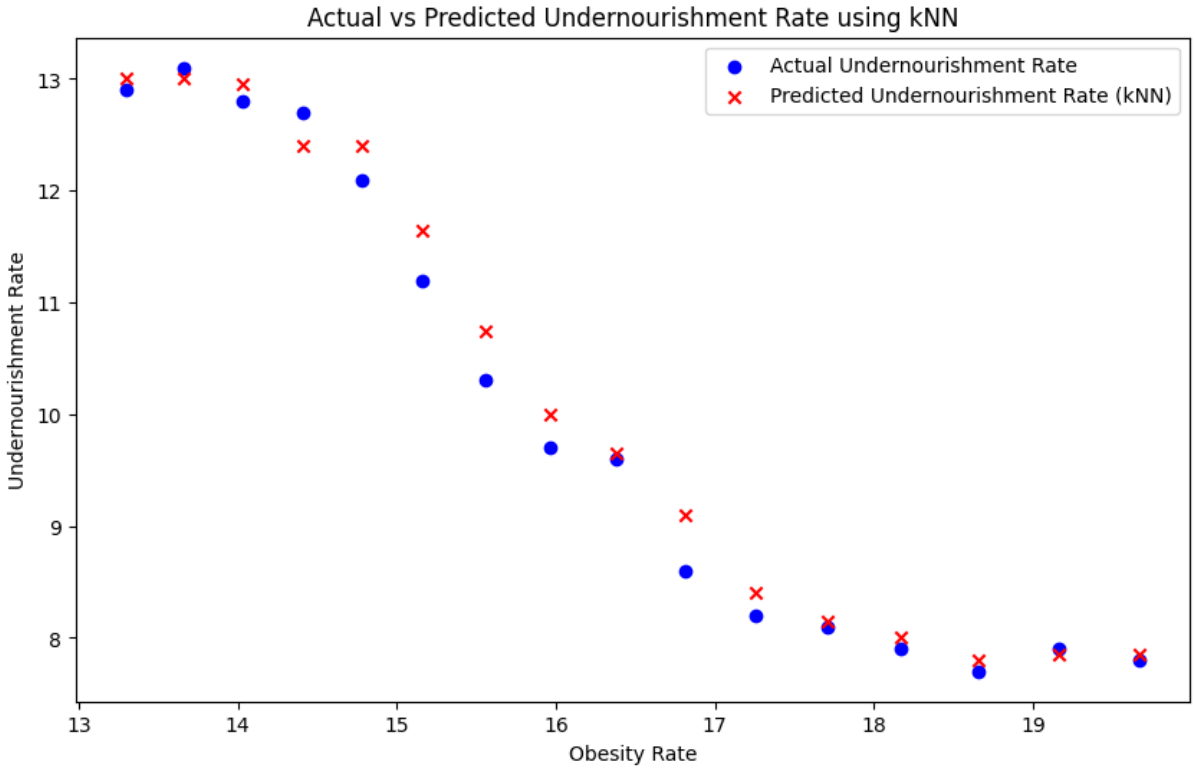
2-Random Forest:

```
Random Forest Mean Squared Error with 5-fold cross-validation: 0.5101380000000051
Standard Deviation of MSE: 0.5888945243503029
Best Parameters: {'n_estimators': 200, 'min_samples_split': 2, 'min_samples_leaf': 1, 'max_depth': 30}
```

1. k-Nearest Neighbors (kNN):

Effect of Number of Neighbors (k):

- **General Trend:** As the number of neighbors increases, the Mean Squared Error (MSE) generally tends to increase as well. This is because using more neighbors can lead to a smoother decision boundary, which might oversimplify the model and result in higher errors.
- **Initial Observation:** Initially, with a small number of neighbors (e.g., 2 or 3), the model might overfit the training data, resulting in low MSE.
- **Optimal Range:** However, as you increase k beyond an optimal value, the model starts to underfit the data, leading to higher MSE. In this case, the optimal value of k appears to be lower, as indicated by the lowest MSE achieved with a smaller number of neighbors.
- **Interpretation:** The observed trend suggests that for this dataset, using a lower value of k in the kNN model leads to better performance in terms of predicting undernourishment rates based on obesity rates. This implies that the model benefits from considering fewer neighbors for making predictions, possibly capturing more localized patterns in the data.



Here, best kNN output is found with the following properties:

- kNN Mean Squared Error for 2 number of neighbors with 5-fold cross-validation: 0.5201666666666679.

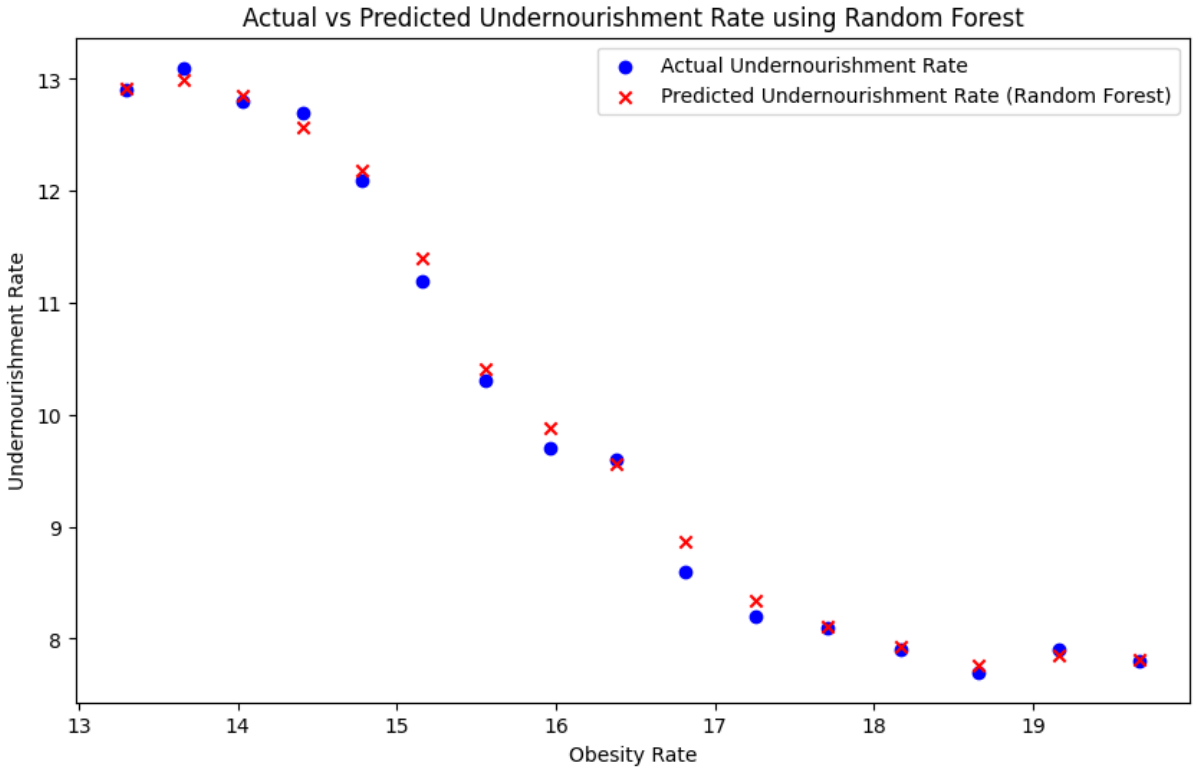
2. Random Forest Regressor:

- **Number of Trees (n_estimators = 200):**
  - Using 200 trees in the Random Forest helps capture a high level of detail and complexity from the data, reducing the model's bias. The large number of trees leads to more stable and accurate predictions by averaging the results of multiple trees. This has contributed to a lower Mean Squared Error (MSE) of mean_cv_mse, indicating that the model is well-tuned.

- **Maximum Depth of Trees (max_depth = 30):**
  - With a maximum depth of 30, the trees in the forest can grow quite deep, allowing the model to learn complex relationships in the data. This depth ensures that the model captures intricate patterns and reduces bias. However, setting the depth to 30 also risks overfitting, but in this case, the model seems to handle it well, as reflected in the low MSE.

- **Minimum Samples Required to Split an Internal Node (min_samples_split = 2):**
  - The parameter value of 2 means that any node with at least 2 samples will be split. This low value allows the trees to grow very deep and complex, capturing more detailed patterns in the training data. While this can lead to overfitting, the combination with other hyperparameters seems to balance it well, maintaining a low MSE.

- **Minimum Samples Required at a Leaf Node (min_samples_leaf=1):**
  - Setting min_samples_leaf to 1 means that each leaf node can contain as few as a single sample. This setting ensures that the model captures fine-grained details in the data, allowing the trees to model the training data very closely. While this can also lead to overfitting, the overall model performance indicates that it successfully generalizes to the test data.



Here, best Random Forest output is found with the following properties:

- Random Forest Mean Squared Error with 5-fold cross-validation: 0.5101380000000051.
- Standard Deviation of MSE: 0.5888945243503029.
- Best Parameters: {'n_estimators': 200, 'min_samples_split': 2, 'min_samples_leaf': 1, 'max_depth': 30}

While both models offer viable means to predict undernourishment rates based on obesity rates, the Random Forest model exhibits a slightly better performance with a notably lower MSE. Hence, for this specific task, the Random Forest model is preferred due to its heightened accuracy and reliability as inferred from the lower MSE.

6-Prediction

In this part, we'll be examining how we can use what we have found so far. I strongly believe that it is of utmost importance in the future that undernourishment rate can be predicted with the given obesity rate. Therefore, I have extended my project a little further and conducted some tests.

In the following, we'll be exploring what can be the undernourishment rate with a given obesity rate for two models we have talked about.

1- k-Nearest Neighbors (kNN):
    a. Predicted undernourishment rate for an obesity rate of 20 using the best kNN model: 7.85.
2- Random Forest:
    a. Predicted undernourishment rate for an obesity rate of 20: 7.812499999999993.

These predictions provide insights into the expected undernourishment rates based on different machine learning approaches. We should note that since Random Forest model's MSE was slightly lower compared to kNN, it is going to produce more accurate results.