International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 08, August-2021

Predicting Optimal Dietary Habits for Fast Recovery from COVID19 using Machine Learning Techniques

Dhaval Garg
SRM Institute of Science and Technology,
Kattankulathur,603203, Tamil Nadu;

***Abstract*: 1) Background: COVID19 is a life-threatening disease which has affected a lot of people all over the world. This disease becomes deadlier if not taken care of properly by healthy diet and enough rest. 2) Objective: This paper focuses on finding the most important food items that help in recovery from COVID19. (3) Method: This was done by calculating effects of intake of different category of food items. After calculating the Recovery rate from the dataset, each country was divided into categories depending on their recovery rate which made it easier to observe the food intake habits. Boruta Algorithm was applied to find out the most important factors contributing to recovery of a person. Multivariate Analysis was carried out to compare different food items to Recovery Rate. (4) Results: The Research concluded that percentage of protein taken from Vegetable oils, percentage of fat from cereals except Beer and percentage of kcal from cereals, meat and milk were the most important factors in recovery of a COVID19 infected person.**

Keywords: COVID19; dietary habits; recovery rate; protein; fat; kcal; Boruta Algorithm; multivariate analysis

1. INTRODUCTION

At the time of writing this article COVID-19 had already taken over the world affecting millions of people all over the world, currently there are 171,244,796 total cases with 154,526,625 people recovered.
Many countries are going through 2nd wave of COVID19 and some countries have already faced the 2nd wave.
People have been living isolated in their homes for the past one year, wearing masks was used to prevent catching the virus through air, sanitizers were used to avoid getting it through touch. Nobody had imagined that such a pandemic could occur and disrupt lives like that.
Offices, Universities, shops, factories etc. all closed to prevent spread of the virus, even though everything is reverting back to normal now the process is slow and the virus has evolved in the past so it's possible that it changes again.
COVID19 has made a devasting impact on the economy as well as healthcare of some countries especially developing countries. In such times all countries have come together in unity and supported each other whenever there is need.
Vaccines have been invented but they are not completely effective and their production is slow especially for developing countries with vast population, in such times people were recommended to maintain a healthy lifestyle like doing yoga, exercise, eat healthy food etc.
The positive thing that came from COVID-19 was people came together, they donated blood, plasma etc. whenever need be. And it showed how countries adapted to the situation and allotted more resources for vaccine researching companies and opened hospitals specially for the virus victims and medical workers all around the world worked day and night to save everyone possible.
Being a victim myself I realized that diet played an extremely important role in my recovery. There is great awareness everywhere that eat healthy for fast recovery and this study tells exactly what category of food items will help in that.

2. MATERIALS AND METHODS

2.1 Dataset

This Study used the COVID-19 Healthy Diet Dataset [[7]] to find what category of food items lead to most possible recovery rate of a COVID-19 infected person also, to study different patterns of diet consumed by various countries.
The COVID-19 Healthy Diet Dataset has information related to 170 different countries and their percentage intake of nutrients from different type of food categories like-alcohol, animal products, aquatic products, cereals, eggs, seafood, fruits, meat, miscellaneous, milk, offal, oil crops, pulses, spices, starchy roots, stimulants, sugar crops, sugar and sweeteners, treenuts, vegetal products, vegetable oils, and vegetables [[1]].
This research used three csv files to study the three main nutrients of food namely -protein, fat, carbohydrate.

- Percentages of protein consumed from each type of food item given.
- Percentages of fat consumed from each type of food given.
- Percentages of energy (in kilocalories) consumed from each type of food given.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 08, August-2021

2.2 Data Preprocessing

As mentioned in the above section the Dataset had 170 countries but it was reduced to 169 countries since French Polynesia had almost all values missing. The 'Undernourished' column had '<' operators which were removed and replaced with the suitable values. All missing values of the column 'Obesity' were filled with the mean of all values in that column. Dropna was used before training the model or implementing the Algorithm. To implement the Boruta Algorithm Recovery Rate Percentage was needed which was calculated using formulae->

Number of people having COVID= percentage of people having COVID *Population of each country
Number of people recovered= percentage of people recovered * Population of each country
Recovery Rate = Number of people recovered /Number of people having COVID
Recovery Rate percentage = Recovery Rate * 100
The Recovery Rate percentage column was inserted at the end of all columns. To train model using Random Forest classifier 'category' was created that categorized all countries from A to F. A having the highest Recovery Rate.

2.3 Methods

Maria et al. [[1]] conducted a study which said "It was observed that countries with a high degree of obese people and with a higher average daily caloric intake, are related to a higher risk of death from COVID-19 while countries with a high number of undernourished people do not show an increase in these percentages". The above mentioned study is focused on mortality rates and deaths , in comparison to that study this research is based on recovery rate and finding what category of food items play the most important role in a person's recovery. In this Study Random Forest Classifier is used by keeping 'Category' column as Y (Predicted value) and rest of the columns as X.

While Boruta Algorithm uses 'Recovery Rate Percentage' as Y (Predicted Value) and rest as X, which will return a list with each column' weight. This Algorithm uses Random Forest Regression.

Given below is a brief explanation of the two used methods.

2.3.1 Random Forest

A Random Forest is made up of a huge number of individual trees that work together as a unit. Each tree produces a class prediction, with the class with the highest votes becoming the final output [[2]].

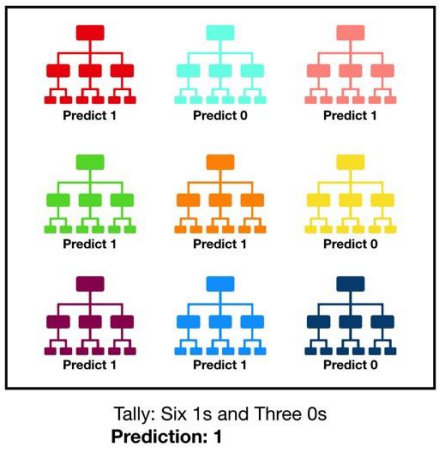

Figure 1. mage reference: This image was taken from https://towardsdatascience.com/understanding-random-forest-58381e0602d2 (1.6.2021).

The trees shield each other from their own mistakes (as long as they don't all make the same mistake) [[2]].
The training algorithm for random forests applies the general technique of bootstrap aggregating, or bagging, to tree learners. Given a training set $X = x_1,\ldots, x_n$ with responses $Y = y_1, \ldots, y_n$, bagging repeatedly (B times) selects a random sample for replacement of the training set and fits trees to these samples:
For $b = 1,\ldots, B$:

1. Sample, with replacement, n training examples from X, Y; call these X_b, Y_b.
2. Train a classification or regression tree f_b on X_b, Y_b.

After training, predictions for unseen samples x' can be made by averaging the predictions from all the individual regression trees on x':

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 08, August-2021

$$\hat{f} = \frac{1}{B}\sum_{b=1}^{B} f_b(x')$$

Figure 2. Image Reference: https://en.wikipedia.org/wiki/Random_forest; Image address: https://wikimedia.org/api/rest_v1/media/math/render/svg/b54befce12aefdb29442bfc71cb5ad452364e8d8 (1.6.2021)

Or by taking the majority vote in the case of classification trees.
Because it reduces the variance of the model without increasing the bias, this bootstrapping approach improves model performance. As long as the trees are not correlated, this means that while a single tree's predictions are highly sensitive to noise in its training set, the average of several trees is not. Simply training multiple trees on a single training set would result in extremely correlated trees (or, if the training technique is deterministic, the same tree multiple times); bootstrap sampling is a method of decorrelating the trees by exposing them to different training sets. [[3]].
"The approach described here is the original tree bagging algorithm. Random forests differ from this basic scheme in only one way: they employ a modified tree learning algorithm that selects a random subset of features at each candidate split during the learning process. This process is sometimes called "feature bagging". The reason for doing this is the correlation of the trees in an ordinary bootstrap sample: if one or a few features are very strong predictors for the response variable (target output), these features will be selected in many of the *B* trees, causing them to become correlated. An analysis of how bagging and random subspace projection contribute to accuracy gains under different conditions is given by Ho" [[4]].

Given below is the screenshot of the accuracy and error obtained while training Random Forest Classifier: -

```
from sklearn import metrics

print('Mean Absolute Error:', metrics.mean_absolute_error(y_test, y_pred))
print('Mean Squared Error:', metrics.mean_squared_error(y_test, y_pred))
print('Root Mean Squared Error:', np.sqrt(metrics.mean_squared_error(y_test, y_pred)))
```

```
Mean Absolute Error: 0.375
Mean Squared Error: 0.625
Root Mean Squared Error: 0.7905694150420949
```

```
print("Accuracy:",metrics.accuracy_score(y_test,y_pred))
```

```
Accuracy: 0.75
```

Figure 3. This Image shows the Output for the Random Forest Classifier from the protein csv file.

2.3.2 Boruta Algorithm

Boruta Algorithm is a feature extraction Algorithm. It was first introduced in R language [[5]]. Instead of competing against themselves, features in Boruta compete against a stochastic version of themselves [[5]].
In practice, a new dataframe is constructed by randomly shuffling each attribute starting with X. Shadow features are these permuted features. The shadow dataframe is now joined to the original dataframe, yielding a new dataframe (named X_boruta) with double the number of columns as X [[5]].
Now we compare the relevance of each of the original attributes to a threshold. This time, the threshold is defined as the shadow feature with the highest feature relevance. This is referred to as a "hit" when the relevance of a characteristic exceeds this threshold. The premise is that a feature is only helpful if it can outperform the best randomized feature [[5]].
"In random forest, the Z score is computed by dividing the average accuracy loss by its standard deviation. It is used as the importance measure for all the variables. But we cannot use Z Score which is calculated in random forest, as a measure for finding variable importance as this Z score is not directly related to the statistical significance of the variable importance. To work around this problem, boruta package runs random forest on both original and random attributes and compute the importance of all variables. Since the whole process is dependent on permuted copies, we repeat random permutation procedure to get statistically robust results" [[6]]. There is one disadvantage of Boruta Algorithm and that is it takes too long to execute. Suppose the dataset is large and number of iterations are more it may take hours to execute. Increasing the iterations will show more weight in the output to make it clearer but the result remains the same.

Results Obtained from the Boruta Algorithm: -

```
print(hits)
```

```
[ 0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.  0.
  0.  0.  0. 12.  0.  5.  0.  0.]
```

Figure 4. This is the output for Protein csv file.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 08, August-2021

```
print(hits)
```

```
[ 0.  0.  0.  0.  0. 39.  0.  0.  0.  0.  0.  1.  6.  0.  0.  0.  0.  0.
  0.  0.  0.  0.  0.  0.  0.  0.]
```

Figure 5. This is the output for Fat csv file.

```
print(hits)
```

```
[ 2.  1.  0.  0.  0. 31.  0.  0.  0. 30. 18.  1.  0.  0.  2.  0.  0.  0.
  0.  0.  0.  0.  5.  0.  0.  0.]
```

Figure 6. This is the output for energy(kcal) csv file.

Each value in the above images denote the importance of each column written in the order as given in the csv file.

2.3.3 Multivariate Analysis

After applying the algorithm and training the model, multivariate analysis was performed by plotting scatter graphs to check if all graphs are similar for the Output given by Boruta Algorithm.

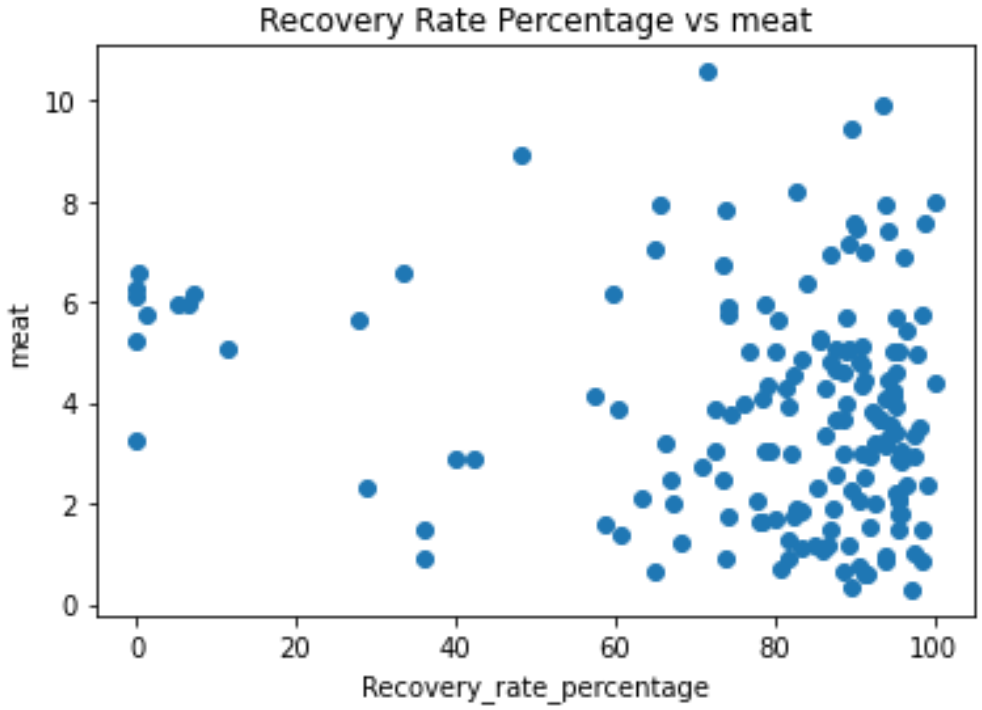

Figure 7. This Figure shows the relationship between Recovery Rate Percentage and the percentage of carbohydrates taken by each country from Meat.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 08, August-2021

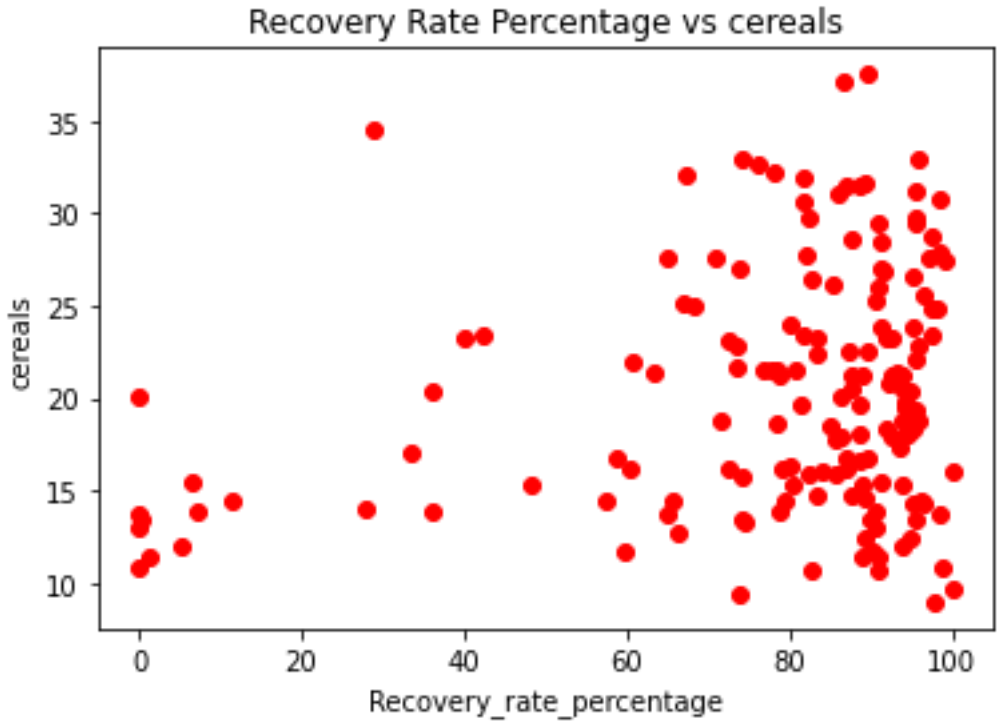

Figure 8. This figure shows the relationship between Recovery Rate Percentage and the percentage of carbohydrates taken from cereals.

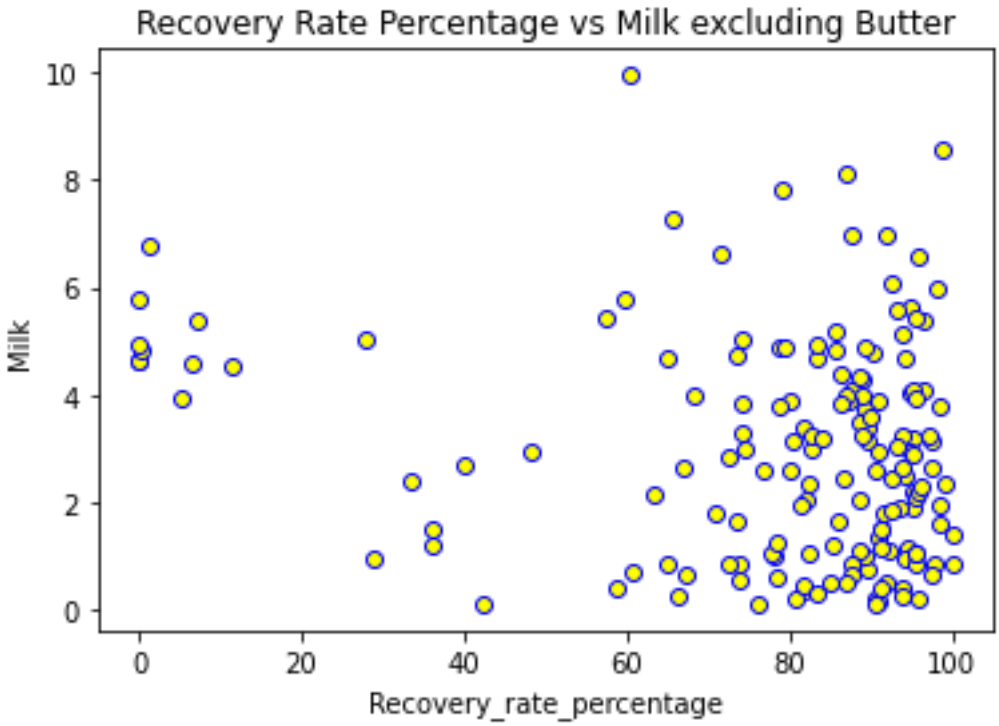

Figure 9. This Figure shows the relationship between Recovery Rate Percentage and the percentage of carbohydrates taken by each country from milk.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 08, August-2021

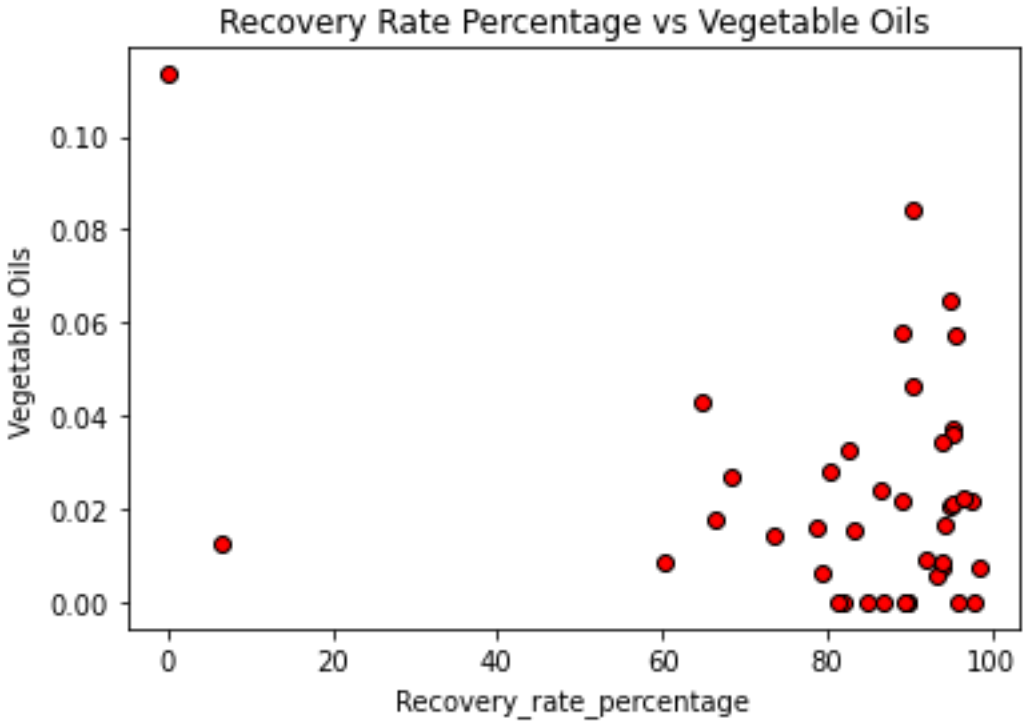

Figure 10. This Image depicts the relationship between Recovery Rate Percentage and the percentage of protein taken by each country from Vegetable Oils.

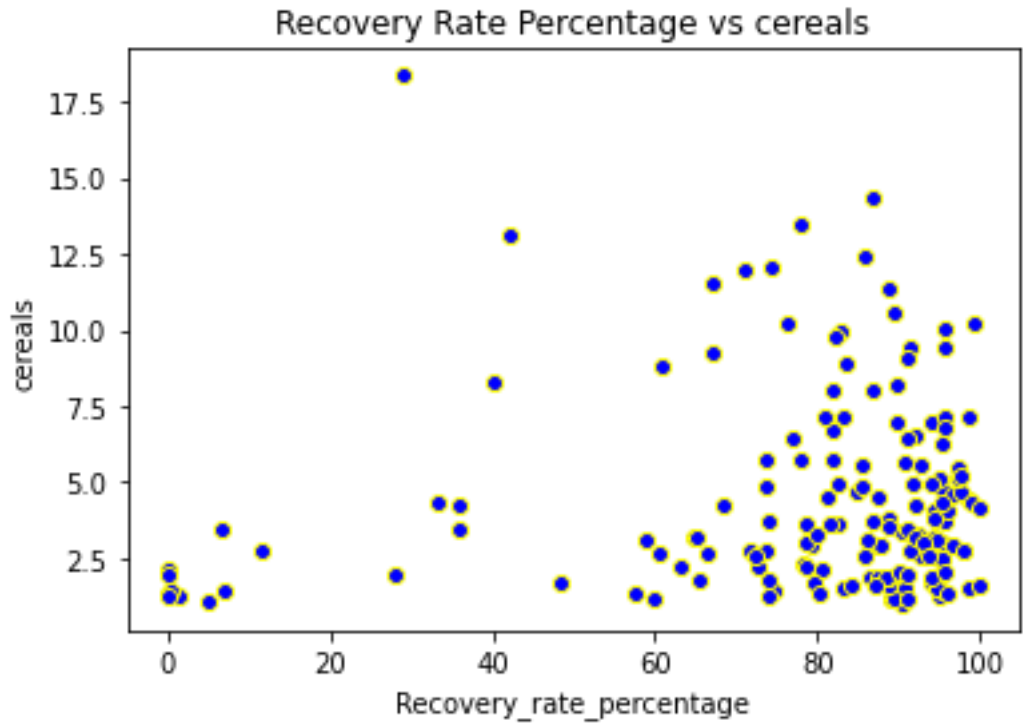

Figure 11. This Image shows the relation between Recovery Rate Percentage and the percentage of fat taken by each country from Cereals.

Furthermore, it was found out that from the Dataset that Nepal and Djibouti had the highest Recovery Rates (98.53 and 98.50 respectively).

3. RESULTS AND DISCUSSIONS

In this work, a study was carried out on the Recovery Rate of people in a country.
First the Dataset was cleaned and Recovery Rate Percentage was found out for each country using the confirmed, Recovered and Population columns.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 08, August-2021

Using simple Python coding a new column was made that categorized all countries from A to F (A denoting recovery rate from 95 to 100) depending on their Recovery Rate Percentage, Using Train Test split the columns were divided into Train Data and Test Data with 'Category' column as the Output/predicted or Y column while rest columns as input or X. Random Forest Classifier was implemented and an accuracy of 0.75 or 75% was achieved(for the protein csv file , for other csv files the accuracy was in close proximity).
As mentioned in the Method section just having a score is not enough, to find out the most important features, Boruta Algorithm was applied and it was found that Percentage of protein from Vegetable Oils, Percentage of Fat from Cereals except Beer and Percentage of carbohydrates from Cereals, Milk and Meat were the most important features for Recovery from COVID19.
In the methods section information about Random forest classifier as well as regression is given since first code used the Random Forest classifier and the Boruta Algorithm used Random Forest regression model.
Finally, to confirm the results scatter plots were formed and all above mentioned features had a similar pattern.
In future work, a more detailed study can be done by analyzing the dietary patterns of the countries with the highest recovery rates and compare that patterns with low Recovery rate countries to know what food to have and what not to have.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

Experimental Subjects/Animals: No Subjects or Animals were used in this study.

Acknowledgement: I want to thank Mrs. Jeysudha my faculty in-charge for helping me in this Article.

REFERENCES

[1] Maria, T.G.; Natalia, A.; Carmen, B.; Oscar, G.-O.; Jose, A.B.-A. Evaluation of Country Dietary Habits Using Machine Learning Techniques in Relation to Deaths from COVID-19. *Healthcare* **2020**, *8*, 371. [CrossRef]
[2] Tony, Y. *UnderStanding Random Forest How the Algorithm Works and Why It Is So Effective*; Medium,2019. [CrossRef]
[3] Random Forest—Wikipedia, Main Article—Bootstrap Aggregating (Wikipedia). [CrossRef] [CrossRef]
[4] Tin, K.H. A Data Complexity Analysis of Comparative Advantages of Decision Forest Constructors. *Pattern Anal. Appl.* **2002**, 5, 102–112. [CrossRef]
[5] Samuele, M. *Boruta Explained Exactly How You Wished Someone Explained It to You*; Medium,2020. [CrossRef]
[6] Deepanshu, B. *Feature Selection: Select Important Variables with Boruta Package*; Listen Data ,2017. [CrossRef]
[7] COVID-19 Healthy Diet Dataset|Kaggle Mari'a Ren. Available online: https://www.kaggle.com/mariaren/covid19-healthy-diet-dataset (accessed 1 June 2021)