An prediction of Healthy Diet required to Ease the recovery from Covid-19 using the approach of Machine Learning.

Rencita Maria Colaco¹, Shreya², Dr. Subba Reddy N V¹, Dr. U Dinesh Acharya⁴
Department of Computer Science and Engineering
Manipal Institute of Technology, Manipal, 576104
Udupi district Karnataka, India
¹rencitacolaco@gmail.com
²itsshreya83@gmail.com
³nvs.reddy@manipal.edu
⁴dinesh.acharya@manipal.edu

Abstract: Global terror that has shaken the world named, COVID-19 virus has taken away huge number of lives. According to the research there are lot of recovery cases also. Most important thing to survive from this disease is having good immunity. Everyone does not have same level of immunity. One main factor on which immunity depends is having a healthy diet. If the routine of having healthy diet is maintained, then the immunity to fight against this virus increases. It is much required that people need to be informed about having an healthy diet. Using the dataset of healthy dietary and using various machine learning algorithms we can determine what type of diet one person needs to have. By using algorithms like Random Forest, KNN, logistic regression and Support Vector Machines we can determine the type of diet and probability of recovery. The dataset required for analysis needs to have all the information regarding the diet. Based on the dataset the prediction is taken place by using Decision Tree algorithm. This method of finding the appropriate diet of a particular person based on amount of Sugar level, Blood Pressure and BMI can be the most useful research in this pandemic time.

1. Introduction
The virus that has led to Covid-19 is also called as SARS CoV-2 virus (Severe Acute Respiratory Syndrome Coronavirus 2)[1]. This was first discovered in Wuhan, China in the year 2019. But then the virus was spread from people to people, states and then countries. This has caused terror in the minds of people and millions of people have suffered and are still suffering. This virus is transmitted through droplets produced from sneezing, coughing and exhaling from the infected person. Since these droplets are heavy they gently fall on the ground and the one who comes in contact, may also have the chance of getting infected. Since this virus was recently discovered there are no vaccines available. Only one weapon to fight against virus is having good immunity, and there are lot of vaccines produced to improve the immunity level, which are quite effective. Another way of boosting the immunity is to have an healthy diet[1]. Especially in countries like India people hardly follow diet routines and they stick to their traditional ways of consuming diet, which has become habitual and difficult to change. Immunity
can be boosted by consuming high nutritional and rich protein food. There are many who have lack of knowledge regarding healthy diet.

There are lot of researches taking place regarding the diet to be consumed during this pandemic. Since lockdown has been imposed it is very difficult to get all the required food supplies and follow the diet. Scientists have researched and already mentioned about various food stuffs to be taken, but there are lot of myths and also facts going around, which has created lot of confusion among people. Since there is faster peer to peer communication, people believe what recommendation have been given by their friends or relative which may or may not be right.

Therefore we have come up with the machine learning model to predict what healthy diet can be consumed to build stronger immunity. We have taken into consideration different conditions patients suffer with, like Blood sugar and Blood Pressure and based on BMI (Body Mass Index) the predictions are taken place. Considering centralized data repository of Kaggle, dataset have been chosen and modified. Using various algorithms like KNN(K-Nearest Neighbor), Logistic Regression, Support Vector Classifier, Random Forest, Naïve Bayes and Decision Tree we have observed the accuracy of prediction. Based on the accuracy we have decided to consider Decision Tree for predicting. Then the user needs to enter the input such as Blood Sugar Level, Blood Pressure, Age and BMI for further calculation. The type of diet needs to be consumed will be predicted.

2. Related Work
Predicting the probability of Recovery after consuming the healthy diet in South Asian countries using the mechanism of Machine Learning. Using the data from the centralized data Repository of Kaggle the dataset was chosen. By using data mining and various machine learning algorithms it was possible to predict. Algorithms such as KNN, Random Forest and Support Vector Machine were used. Input were taken as the Kcal consumption of Meat, Alcoholic Beverages, Animal Products, Vegetable Products and Cereals and result was prediction of recovery rate. According to the result they found that the people who consumed more plant based product than animal product were more likely to be recovered soon[1].

The statistical comparison is carried out based upon the diet consumed and infection rate using the RStudio. Pearson correlation was used for statistical analysis for two variables to measure linear dependence between them. The dataset was selected from the repository of Kaggle and comparison was carried out on density of populations similarity. According to comparison they found that in the countries where lockdown was imposed the obesity was increased and this may be decreased once the lockdown was lifted. The country where people have consumed high energy and protein intakes were the one’s that were having least number of cases[2].

The study was made taking into consideration the amount of protein intake and covid infections. By using deep learning network mechanism regression tool the predictions were carried out. Considering the food habits of United Nations this comparison was carried out, the food group supplies as nutrition values, undernourished percentages and obesity. According to the result they found out that there is no direct strong relationship between protein intake and death rate. The disease also depends on age, health condition, sex and food plays a small role[3].

Healthy eating habits can bring about the positive change in behaviour of infected person. Logistic regression was used to analyse the positive impact of healthy diet. Best Worst Scaling were used to measure FCQ(Food Choice Questionnaire). According to the result they obtained that the one’s who consumed amount of vegetables(14%), water(7%), fruits(13%), meat(4%), beans and grains(3%) were less likely to be infected[4].

The research was conducted to find out the impact of nutrient food on covid patients. Hypothesis testing was carried out on various factors like peer influence, social media influence, fear, peer influence,
attitude and healthy nutrition. After the study was made they have come up with conclusion that social influence have small impact but healthy nutrition consumption had high impact and this depends on many other factors depending upon the individual[5].

3. Proposed Architecture

![Diagram of methodology]

Figure 1. Methodology

The important step in order to process the data is to collect the data. Collected data is pre-processed and then data modelling is carried out. After the model evaluation is done, model will be used predict the required result. In the figure 1 the complete methodology is shown.

3.1. Data Collection

Data is the crucial part in order to train the machine to get the accurate result. The data collected should be perfect and there must not be any faults in it. It is always important to choose the data from best repository and the dataset required for our analysis was considered from centralized repository of Kaggle.com. The dataset consisted the information regarding the different healthy diets, diabetes condition and blood pressure. Figure 2, figure 3 and figure 4 shows the example of the samples of data considered. Figure 2 consists of diabetes dataset whereas figure 3 consists of food supply in Kcal and figure 4 consists of supply food data description, all these 3 dataset were used to predict the diet for particular person.
3.2. Data Visualization

We understand the data better when we visualize it. It is nothing but graphical representation of information. Graphical representation by means of maps, graphs and charts. In the figure 5 we can see the result of data visualization of diabetes data in form of bar graph.

![Figure 5. Result of Data Visualization](image)

3.3. Pre-Processing Data

In the Pre-processing step the data from diabetes data set is processed, i.e., only Glucose level, Blood Pressure, Skin Thickness, Insulin and BMI are considered other data is set to NILL. Next the Count Mean, Standard Deviation, minimum value, 25%, 50%, 75% and maximum value is calculated and new table is Formed. Then the features Glucose, Insulin, BMI and Age are selected and data is divided as training data and test data, where test data considered as 20% of data.

![Figure 4. Supply Food Data](image)

| Categories                  | Items                                                                 |
|------------------------------|----------------------------------------------------------------------|
| Alcoholics                   | Alcohol, Non-Food, Beer, Beverages, Alcoholic, ...                   |
| Animal Fats                  | Butter, Cheese, Cream, Fats, Animals, Fat, Fish, ...                |
| Aquatic Animals              | Aquatic Animals, Others, Aquatic Plants, Bovine, ...                |
| Aquatic Plants               | Meat, Bovine Meat, Meat, Other, Mutton & Goat Meat, ...             |
| Cereals - Excluding Beer    | Barley and products, Cereals, Other, Maize and ...                  |
| Eggs                         | Fish, Seafood                                                       |
| Fruits - Excluding Wine      | Cephalopods, Crustaceans, Dermal Fish, Fresh, ...                   |
| Meat                         |                      |
| Milk - Excluding Butter      | Infant Food, Miscellaneous                                       |
| Miscellaneous               |                                                                 |
| Offals                       |                                                                 |
| Oilseeds                     |                                                                 |
| Oils                         |                                                                 |
| Pulses                       |                                                                 |
| Spices                       |                                                                 |
| Stimulants                   |                                                                 |
| Sugar & Sweeteners           |                                                                 |
| Sugar Crops                  |                                                                 |
| Sugar Maltose                |                                                                 |
| Tea                           |                                                                 |

![Figure 2. Diabetes dataset](image)

| Pregnancies | Glucose | Blood Pressure | Skin Thickness | Insulin | BMI | Diabetes | Gender | Age | Outcome |
|-------------|---------|----------------|----------------|---------|-----|----------|--------|-----|---------|
| 0           | 0       | 148            | 72             | 35      | 0.336| 0.027    | 0      | 71  |
| 1           | 0       | 85             | 66             | 20      | 0.266| 0.351    | 0      | 31  |
| 2           | 0       | 103            | 64             | 0       | 0.233| 0.972    | 0      | 32  |
| 3           | 1       | 89             | 66             | 23      | 0.64 | 0.157    | 21     | 71  |
| 4           | 1       | 137            | 40             | 35      | 0.18 | 2.238    | 33     | L1  |

![Figure 3. Food supply in Kcal](image)

| Alcoholic Beverages | Alcohol, Non-Food, Beer, Beverages, Alcoholic, ... |
| Animal Fats         | Butter, Cheese, Cream, Fats, Animals, Fat, Fish, ... |
| Aquatic Animals     | Aquatic Animals, Others, Aquatic Plants, Bovine, ... |
| Aquatic Plants      | Meat, Bovine Meat, Meat, Other, Mutton & Goat Meat, ... |
| Cereals - Excluding Beer | Barley and products, Cereals, Other, Maize and ... |
| Eggs                | Fish, Seafood |
| Fruits - Excluding Wine | Cephalopods, Crustaceans, Dermal Fish, Fresh, ... |
| Meat                |                      |
| Milk - Excluding Butter | Infant Food, Miscellaneous |
| Miscellaneous       |                                                                 |
| Offals              |                                                                 |
| Oilseeds            |                                                                 |
| Oils                |                                                                 |
| Pulses              |                                                                 |
| Spices              |                                                                 |
| Stimulants          |                                                                 |
| Sugar & Sweeteners  |                                                                 |
| Sugar Crops         |                                                                 |
| Sugar Maltose       |                                                                 |
| Tea                 |                                                                 |
3.4. Data Modelling
For data modelling different algorithms were used like KNN(K-Nearest Neighbor), Logistic Regression, (SVM)Support Vector Classifier, Random Forest, Naïve Bayes and Decision Tree. These algorithms were used to find out which model works best for this data model. Now let us consider each of the algorithms:

3.4.1. KNN(K-Nearest Neighbor)
The KNN algorithm is supervised learning algorithm used to classify the similarity in the data[6]. The algorithm is as follows:

Step-1: Select K number of the neighbors.
Step-2: Calculate Euclidean distance of K neighbors. The formula is given by equation (1)

$$\sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

(1)

Where (x1,y1) and (x2, y2) are the points of 2 neighbors.

Step-3: As per the Euclidean Distance calculated consider the K nearest neighbors.
Step-4: Calculate the number of data points in the each category from these K neighbors.
Step-5: New data points are required to assign to that category for which the number of neighbor is maximum.
Step-6: Model is ready.

3.4.2. Logistic Regression
Logistic regression is a statistical model used to model a binary dependent variable using logistic function in its basic form[7]. The steps include:

1. Data Pre-processing step
2. Fitting Logistic Regression to the Training set
3. Predicting the test result
4. Test accuracy of the result(Creation of Confusion matrix)
5. Visualizing the test set result.

In the figure 7 the best value for n-estimator is shown which is obtained from logistic regression. The highest value obtained is 0.3766237662337664
Figure 7. Result of finding best value for n-estimators

3.4.3. Support Vector Machine(SVM)
SVM is a supervised learning model used to analyse data for classification and for regression analysis[8]. It can be applied even on the complex data, where it tries to divide the data based on the class labels provided. There are two types of SVM:
1. Linear SVM
2. Nonlinear SVM
We have used Linear SVM for classification of data.

3.4.4. Naïve Bayes
Naïve bayes algorithm also uses supervised learning model which is used for classification of data. The steps include:
1. Converting the dataset into frequency values.
2. Finding the probabilities of given features and generating the likelihood table.
3. Posterior probability is calculated using Bayes theorem.

3.4.5. Random Forest
Random Forest consists of many decision trees on different datasets, where average is taken to improve the accuracy[9]. The steps include:
1. Random K data points are selected from the training set.
2. Decision trees are built associated with the data points selected (Subsets).
3. Choose the number of decision trees that you want to build.
4. Repeat Step 1 & 2.
5. Find the predictions for new dataset of each decision tree, category that wins the majority votes assign the new data points to that category.

3.4.6. Decision Tree Algorithm
ID3 algorithm is used for prediction which calculates Entropy and Information Gain[10]. The steps include:
1. The root node is considered from S dataset.
2. On each iteration calculates Entropy(H) and Information gain(IG) of this attribute. The formula for Information Gain is as follows in equation (2)
   \[
   \text{Gain}(S, A) = \text{Entropy}(S) - \sum ( \frac{|S_v|}{|S|} \times \text{Entropy}(S_v) )
   \]
   where S is the dataset and A is the attribute. To calculate entropy the equation (3) is
   \[
   \text{Entropy}(S) = \sum_{i=1}^{n} -p(I)\log_2 p(I)
   \]
   where p(I) is the proportion of S belonging to class I.
3. The attribute with smallest Entropy or Largest Information gain is selected.
4. Selected attribute is used to split the dataset S to produce a subset of the data.
5. On each subset produced algorithm recurs, but considering the attribute that were not selected before.

3.5. Model Evaluation
By using different algorithms mentioned above the accuracy was compared to find out the best classification algorithm for our dataset. We found out that Decision Tree algorithm was giving the highest accuracy of 97.40259740259741 as shown in the table 1 below. Therefore the classification made be Decision Tree algorithm was selected for predicting.

| Algorithm           | Accuracy          |
|---------------------|-------------------|
| Logistic Regression | 7.792207792207792 |
| K Nearest neighbors | 19.480519480519483 |
| Support Vector Classifier | 87.01298701298701 |
| Naive Bayes         | 89.6103896103896  |
| Decision tree       | 97.40259740259741 |
| Random Forest       | 73.37662337662337 |

Table 1. Accuracy of different algorithms

Classification of dataset according to decision tree algorithm is shown below in the figure 8.

![Figure 8](https://example.com/fig8)

3.6. Prediction
When the user inputs the information like BMI, Glucose Level, Age and Blood pressure the model has to analyse the information and predict the healthy diet. Then the prediction from decision tree algorithm is considered and the supply food per Kcal dataset is analysed. Figure 10 shows the prediction of food that has to be consumed, which consists both animal and vegetal food products. According to the prediction for given input shown in figure 9 of age is 24, BMI is 21.2, glucose level is 120 and blood pressure 110, cereals has to be consumed more and alcoholic beverages less compared to other food. The diet consists of milk, meat, animal fats, starchy roots, cereals and alcoholic beverages. When the animal product group is considered the prediction is shown in figure 10, Where meat has to consumed more. Same as animal product it is able to predict for the vegetal product group as shown in figure 12, Where cereals needs to be consumed more.
Enter Glucose Level : 120
Enter BloodPressure : 110
Enter BMI : 21.2
Enter Age : 24

Figure 9. Input given.

Figure 10. Mean food intake by both animal and vegetal products.

Figure 11. Mean food intake by Animal product group
4. Result
As we have already seen the predicted result in the above, based on that prediction it is also possible to find out the different food items that belongs to that diet class. Figure 13 shows the types of food that can be consumed for the particular person whose input was considered to build the immunity.

Figure 13. Predicted healthy diet that needs to be consumed

5. Conclusion
We all know that this disease has shook the world and how important it is to have good immunity in order to fight the virus. Therefore it is very much required to consume healthy diet to increase the immunity. The study that was carried out to understand the healthy diet which needs to be consumed by the particular individual was predicted using different classification algorithms. Through which we were able to say that decision tree algorithm gave highest accuracy in classification. Therefore using this
algorithm we were able to predict the healthy diet. According to different trials we can say that the vegetal products were more likely to boost immunity than the animal products. Our proposed methodology will help the people, to know what to consume according to their amount of glucose level, blood pressure, BMI and Age in order to boost the immunity.

6. References

[1] Md Showrov Hossen and Dip Karmoker 2020 Predicting the Probability of Covid-19 Recovered in South Asian Countries Based on Healthy Diet Pattern Using a Machine Learning Approach 2nd, Int., Conf., on Sustainable Technologies for Industry 4.0 Dhaka P 1-6

[2] Rohan Ramachandran, Senthil Velan S and Daifa Imtiyaz Wadkar 2021 Statistical Comparison of COVID-19 Infections Based Upon the Food Habits/Diets in Countries Using RStudio 11th, Int., Conf., on Cloud Computing, Data Science & Engineering (Confluence) P 3 7

[3] Tianzhe Fang, Yuheng Shi and Beining Mu 2020 Analysis of Protein in Diet vs COVID-19 with Machine Learning 2nd Int., Conf., on Information Technology and Computer Application (ITCA) P 2 3 5

[4] Sara R. Jaeger, Leticia Vidal, Gaston Ares, Sok Chheang and Sara Spinelli 2021 Healthier eating: Covid-19 disruption as a catalyst for positive change Journal of Food quality and Preference Elsevier Vol 92 P 1 2 5

[5] Nilsah Cavdar Aksoy, Ebru Tumer Kabadayi and Alev Kocak Alan 2021 An unintended consequence of Covid-19: Healthy nutrition Appetite Elsevier Vol 166 P 1-8

[6] Manzilur Rahman Romadhon and Fachrul Kurniawan 2021 A Comparison of Naive Bayes Methods, Logistic Regression and KNN for Predicting Healing of Covid-19 Patients in Indonesia 3rd East Indonesia Conf., on Computer and Information Technology P 2

[7] Furqan Rustam, Aijaz Ahmad Reshi, Arif Mehmood, Saleem Ullah, Byung-won On, Waqar Aslam and Gyu Sang Choi 2020 COVID-19 Future Forecasting Using Supervised Machine Learning Models IEEE Access Journal Vol 8 P 3

[8] M. Y. Shams, O. M. Elzeki, Mohamed Abd Elfattah, Lobna M. Abouelmagd, Ashraf Darwish and Aboul Ella Hassanien 2021 Impact of COVID-19 Pandemic on Diet Prediction and Patient Health Based on Support Vector Machine Advanced Machine Learning Technologies and Applications: Proceedings of AMLTA 2021 P 66

[9] Vishan Kumar Gupta, Avdhesh Gupta, Dinesh Kumar, and Anjali Sardana 2021 Prediction of COVID-19 Confirmed, Death, and Cured Cases in India Using Random Forest Model Journal on Big Data Mining and Analytics Vol 4, No 2 P 120

[10] Ms. Tejaswini U. Mane 2017 Smart heart disease prediction system using Improved K-Means and ID3 on Big Data International Conference on Data Management, Analytics and Innovation P 24