Computer and Information Science; Vol. 11, No. 2; 2018
ISSN 1913-8989 E-ISSN 1913-8997
Published by Canadian Center of Science and Education

Finding Nutritional Deficiency and Disease Pattern of Rural People Using Fuzzy Logic and Big Data Techniques on Hadoop

Sadia Yeasmin[1], Muhammad Abrar Hussain[1], Noor Yazdani Sikder[1] & Rashedur M Rahman[1]

[1] Department of Electrical and Computer Engineering, North South University, Dhaka, Bangladesh

Correspondence: Rashedur M Rahman, Department of Electrical and Computer Engineering, North South University, Dhaka, Bangladesh. E-mail: rashedur.rahman@northsouth.edu

Received: December 27, 2017 Accepted: January 7, 2018 Online Published: March 20, 2018

doi:10.5539/cis.v11n2p11 URL: http://dx.doi.org/10.5539/cis.v11n2p11

Abstract

Over the decades there is a high demand of a tool to identify the nutritional needs of the people of Bangladesh since it has an alarming rate of under nutrition among the countries of the world. This analysis has focused on the dissimilarity of diseases caused by malnutrition in different districts of Bangladesh. Among the 64 districts, there is no single one found where people have grown proper nutritional food habit. Low income and less knowledge are the triggering factors and the case is worse in the rural areas. In this research, a distributed enumerating framework for large data set is processed in big data models. Fuzzy logic has the ability to model the nutrition problem, in the way helping people to calculate the suitability between food calories and user's profile. A Map Reduce-based K-nearest neighbor (mrK-NN) classifier has been applied in this research in order to classify data. We have designed a balanced model applying fuzzy logic and big data analysis on Hadoop concerning food habit, food nutrition and disease, especially for the rural people.

Keywords: big data, hadoop, map reduce-based K-nearest Neighbor (mrK-NN), fuzzy logic, disease, food nutrition, rural area

1. Introduction

Everyone has daily nutritional needs and to be fit, active and healthy, they need to maintain a nutritious food chart. As food is one of the basic needs of human being, hence the importance of taking proper nutrition comes along with it to keep human body free from diseases. As living in Bangladesh, where most of the people are not conscious about the nutrition aspect of foods, they fall victim to deficiency or overdose of it. So diseases break out quite often. Although nutrition is not the only determinant factor of having diseases, still maintaining proper nutrition can minimize some disease possibilities. The rural people of Bangladesh have many diseases as they do not have proper knowledge to fight against them. Besides, medical facilities are less available in rural areas compared to urban areas. To fight against these diseases, the first task is to educate people about the cause and symptoms of the diseases. Some of the diseases occur from virus and bacterial attack, again some from lacking nutrition in daily food habit. In this study we are focusing on the food aspect of disease outbreak in rural area of Bangladesh.

One of our objectives is to bring change in the food consumption of rural people to fight against frequent diseases that they go through. If nutritional knowledge can be reached in every house in rural areas and if they are given a cost efficient solution to get proper food, the rate of morbidity will be surely seen a downfall. The Bangladesh Integrated Household Survey (BIHS), a nationally representative survey designed by the International Food Policy Research Institute (IFPRI) has been conducted in Bangladesh over 6,500 households (Ahmed, Tauseef & Ghostlaw, 2016). The collected survey dataset is focused as the intake choice and daily food habit in this research.

To fulfill the above purpose this research applies big data analysis(MapReduce-Based-K-NN) and Fuzzification on the food habit of people of different districts which varies more or less from one another and it is also focused to find out the particular diseases of any district under a division to prevent and fight that disease in that area.

MapReduce is a programming model and an associated implementation for processing and generating large datasets that is responsive to a broad variety of real-world tasks. If a dataset is too much huge to start with, then running it through one or two servers sequentially (Anchalia & Roy, 2014) would take too long time. In our dataset, as we did not know in advance how much input we will have, that is where MapReduce is useful to share these discrete units of work. For counting staffs (eg., food items frequencies per day, crops variations in different districts

etc.) MapReduce is better choice because of its very cheap reduction & linear scalability. Fuzzy Inference System (FIS) is used to get more accuracy on nutrition recommendations (Priyono & Surendro, 2013) for different age group people in a family.

To estimate which food is more preferable in which area we have implemented MapReduce in K-Nearest Neighbor classification (K-NN). Though K-NN is a very well-known method in data mining for its effectiveness and simplicity which provides a simple non-parametric procedure to class the labels of input pattern, this method lacks of scalability, takes longer run time and consumes more memory in big training dataset with high dimensionality. In this scenario, this paper proposes a MapReduce-based approach for K-Nearest Neighbor (K-NN) classification which allows to classify large volume of hidden cases (test examples) against a big dataset (training) simultaneously. It deals with the weakness of K-NN algorithm by reducing the runtime and memory consumption of RAM. Different MapReduce framework implementation are possible but in this paper we are using Hadoop implementation (Kumar, N.K.& S.K, 2016) because of its flexibility, time effectiveness, fault tolerant, scalable, open source and distributed file system nature. The problem is not the lack of data but the lack of information. And that is the main reason we are using Hadoop to acquire the correct information of food pattern intake of 64 districts to classify diseases based on various nutrient deficiencies.

2. Literature Review

Many researches have been done on nutrition intake and its relation with diseases. Bazzano et al. (2002) have done a research on having relation of fruits and vegetables with cardiovascular disease. In their study they have showed an inverse association between fruit and vegetable intake and the risk of subsequent cardiovascular disease. These findings have important clinical and public health implications. Increased fruit and vegetable intakes have been recommended to prevent morbidity and mortality from cardiovascular disease.

Another cross-sectional study involving 62 patients with coronary diseases (CD) have been done to figure out some nutrient relation on having their diseases. In that study zinc and calcium intake was found to be correlated with reduced femoral neck bone mineral density (BMD). In the conclusion (Araújo et al., 2017) stated “Low calcium and zinc intake, glucocorticoid use, and active disease phase are favorable conditions for bone loss in patients with Crohn's disease”.

However generating useful information from surveyed data, medical reports, news articles, electronic reports cannot be easily done with traditional data mining system. Kumar et al. (2016) observed the frequent changes in the behavior of cancer disease and found that the disease generates a massive volume of data. They experimented microarray-based analysis on these data. The particular identification of genes of interest that are accountable for causing cancer are crucial in microarray data analysis. MapReduce based algorithms are proposed to select features and after feature selection, a MapReduce-based K-nearest neighbor (mrK-NN) classifier is used to classify microarray data. Implementing the algorithms in Hadoop framework, a comparative analysis is done on these Map Reduce-based models using microarray datasets of different dimensions. It is observed from their obtained results that the MapReduce-based models consume much less execution time than conventional models in processing big data.

Data analysis is an important thing to generate the most accurate result. Since data is found in several forms, it is essential to consider all forms of data. While considering such, big data challenges come in front of us. Counting algorithm is used to solve large amount of data and carry out the counted result. (Dai, Zhang, Wang &Ding, 2012) proposed a system for pre-processing big data based on Hadoop platforms. Usually MapReduce programming models are used to perform application processing that involves several deployment and calculation strategies such as collection and storage of data in distributed storage nodes, reading data from storage nodes by computation nodes and performing map operations on it, commutation between compute nodes and performing reduction operation to obtain computation result. During data collection and storage process storage nodes perform I/O operation. But the computing resources of those nodes are not fully utilized. (Dai et al., 2012) proposed a system that can utilize idle computing resources in cluster storage nodes to perform pre-processing works in parallel with the time of data collection and on the basis that I/O performance is not affected. This process reduces the data size of disk transfer and network communication and also the runtime of applications. They conducted experiments based on WordCount which showed that the proposed system could effectively decrease data transmission rate in the computing phase, reduce computing time, and improve application performance.

Some work has been done to fuzzify a system when necessary. Since we are approximating weekly family meal intake, fuzzy inference system is a great tool to do the so. Again some research has been done in this area. (Nguyen , Ngo & Pham, 2013) proposed Intuitionistic Interval Type-2 Fuzzy C-means Clustering (InIT2FCM) to handle clustering related problems. They introduced Intuitionistic Fuzzy Sets (IFS) and Intuitionistic Type-2 Fuzzy Sets

(InIT2FS) for handling data uncertainty. The combination of IFS and InIT2FS overcame several problems related to "conventional FCM" algorithm in handling uncertainty. Uncertainty handling process is based on the identification of membership functions and non-membership functions which depend on resistance assessment function. They established InIT2FS, which functions on fuzzy set intuition extension that enables to handle both uncertainty and hesitance in the data. Application of these fuzzy set in any fuzzy clustering algorithm provide better results than other traditional algorithms. Thus application of Intuitionistic sets in place of simple sets provides better results in clustering uncertain data. Similarly (Priyono & Surendro, 2013) suggested a way to calculate the suitability between food calories and user's profile. They said that, though it was possible to measure food calories from the foods which come in packet with proper labeling, it was hard to measure calorie of the foods which were unlabeled and unpacked like the foods they served in restaurants and cafes. They used fuzzy logic to acknowledge those people whether the food they intake was suitable for them or not. As input they gave user profile which includes age, height, weight and sex. With this four inputs they calculated an individual's BMI and BMR. Finally, they applied two FIS (Fuzzy Interface System) models to get the output. One is TSK FIS order 1 which is used to get daily calorie need assessment and another one is Tsukamoto FIS which is used to assert calories in the food they intake. Using these two models they come up with an output which says whether the food one person wills to take is suitable for him or not.

Since this research is aimed for the people of Bangladesh, we have gone through some studies regarding local food and nutrition intake of people. A Food Composition Table (FCT) for Bangladesh has been conducted by (Shaheen et al., 2013). The nutrient composition of 381 foods representing 15 food groups including 20 key foods and selected cooked recipes are included in the FCT and in the related new Food Composition Data Base (FCDB). All the nutrients (available) in each of those foods have been calculated in 100 gram of those foods. Eighty seven food items have been analyzed for both nutrients and other nutritionally important food constituents. Nutrient composition has also been analyzed for 37 single ingredient and 11 multi-ingredient recipes. The updated values for energy, macronutrient and micro-nutrient contents of foods are useful to improve food and dietary analysis and planning in Bangladesh.

3. Data

In the collected household survey data, a dataset containing consumed food chart of last 7 days is available with the information of food quantity, food price, consumers' preference of foods, etc. Again the whole samples are categorized into 7 different divisions and 64 districts having feature dimensions of 4185 and the size of downloaded file is 1.96GB. This dataset is focused to their intake choice and daily food habit. Since, we are analyzing food habit of the rural people of Bangladesh; a family usually takes same amount of food weekly. From the BIHS data set we have acquired weekly consumption details of food by the rural people all over Bangladesh. A community survey supplements the BIHS data to provide information on area-specific contextual factors. The sample is statistically representative at following levels: nationally representative of rural Bangladesh, representative of rural areas of each of the seven administrative divisions of the country: Barisal, Chittagong, Dhaka, Khulna, Rajshahi, Rangpur, and Sylhet and representative of the Feed the Future (FTF) zone of influence (Ahmed et al., 2016). It is to be mentioned here, though Bangladesh has currently 8 divisions but previously it had 7. Mymensingh division is the new one which was under Dhaka division and in this research, Mymensingh division is not considered.

Every year, (Directorate General of Health Services [DGHS], 2016) under the Ministry of Health & Family Welfare, Bangladesh publish local health bulletins report of almost all districts containing all the health-related information of a year which tries to disseminate the overall health related activities under each district. In this research, DGHS reports and bulletins of the recent years have been analyzed to find out the top most diseases of any district and common diseases of the districts under a division. Among the diseases, non-communicable diseases caused by under nutrition and malnutrition has been detected for each district in this research. Again, we also use the news in The Daily Star which is a renowned Newspaper in Bangladesh. There is a health bulletin section of Daily Star which publishes different health and disease-based reports every day for different regions of Bangladesh as well as worldwide.

4. Predictive Model Design

The presence of huge number of irrelevant and unrelated features degrades the value of the evaluation of disease patterns. As a result, it is important to analyze the BIHS dataset on Hadoop from different appropriate perspectives.

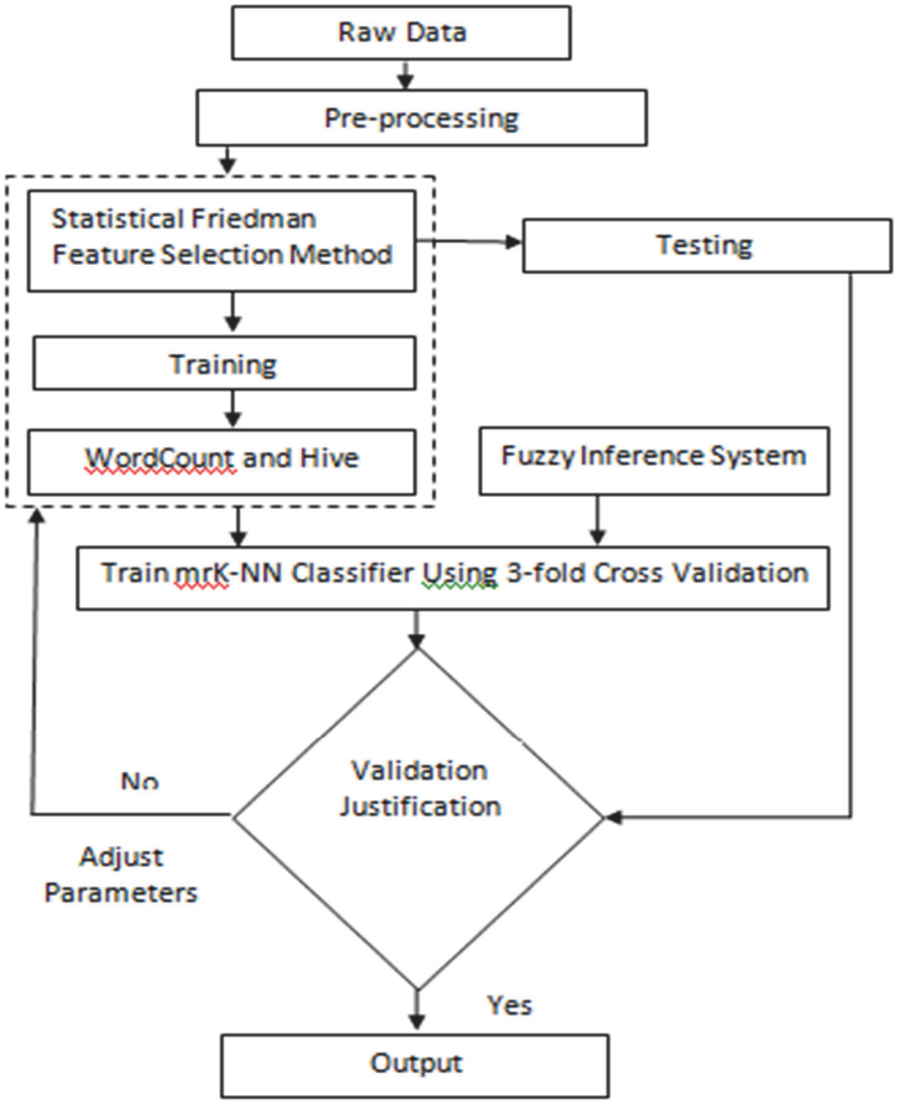

Figure 1. System Overview

Figure 1 represents the system overview of all the algorithms that are used in Hadoop and Fuzzy Inference System. Our system works in the following stages:

i) The input data is preprocessed using four pre-processing techniques for reducing runtime and disk I/O bandwidth. MapReduce-based statistical Friedman test is applied to select relevant features to reduce the dimensionality curse of the BIHS dataset. Then it is divided into two parts – training set to develop the model and testing set to calculate the accuracy.

ii) Word Count and Hive queries are used to calculate the food habit of households and Fuzzy Inference System is used to calculate the recommended intake of 20 nutrients for a household per week. This obtains particular diseases for different groups of nutrition deficiencies.

iii) MapReduce-based K-NN classifier is applied to classify the dataset. The training is done using 3-fold cross validation to obtain K parameter. Different K values are used to test the classifier using the testing dataset. The whole performance is evaluated by analyzing accuracy, precision and recall.

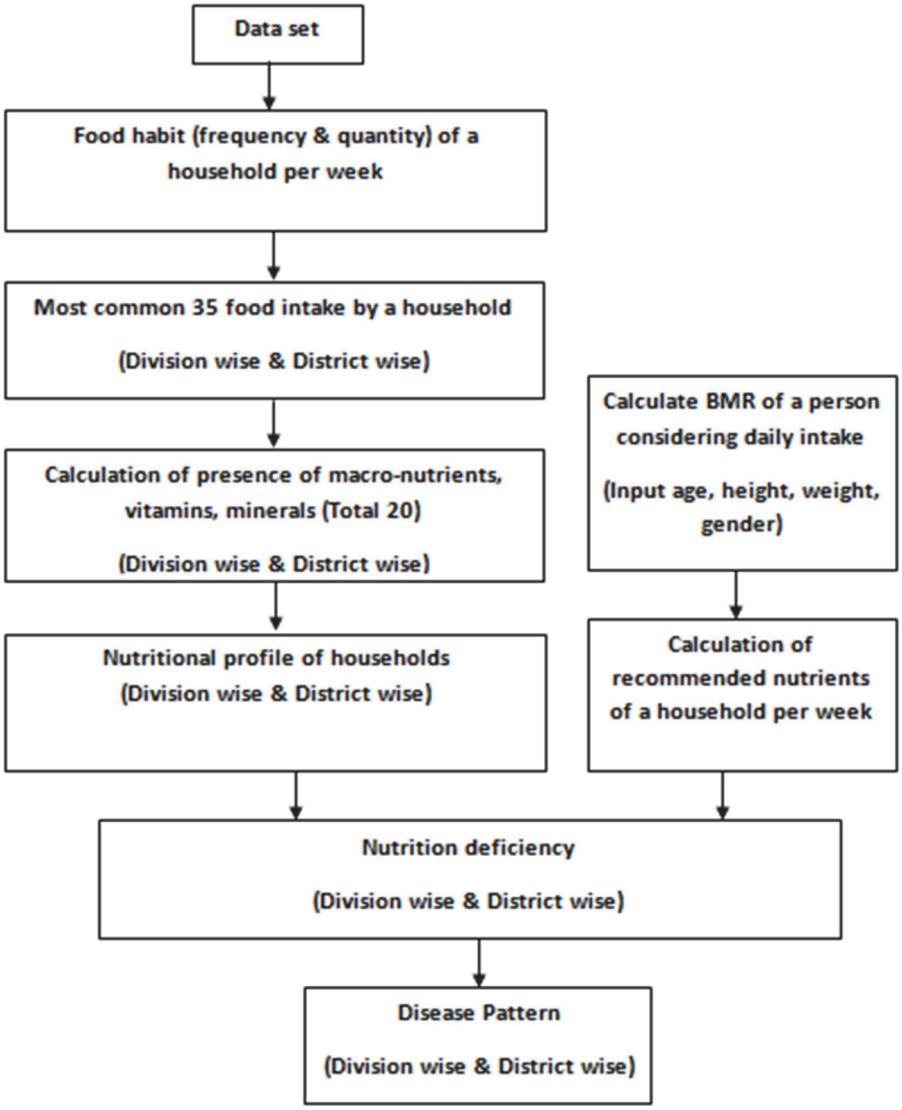

Figure 2 Block diagram of system workflow

Figure 2 shows the summary of the whole method work flow step by step.

4.1 Pre-Processing Techniques

To analyze the food habit of rural people and to evaluate the disease patterns, we need to load the whole dataset of 6500 samples with 125 modules of households in the HDFS. We need to do the whole process by Map Reducing on Hadoop, then after collecting data, Hadoop stores them in distributed storage systems. These are storage nodes in clusters. Then, map operations are performed on the data of storage nodes. The compute nodes perform reduction operations and obtain computation results. In this process, to collect and store data, the computation resources are not entirely utilized as the storage nodes mostly perform IO operations. Therefore the data collection and storage stage will start computation operations earlier to utilize the idle resources in a way that IO performance will not be affected. Hence network communication and the data size of disk transfer can be reduced and the runtime of applications can be minimized (Dai et.al, 2016).

In this case, four steps of pre-processing technique are used before using the modules of households' information collection procedure on Hadoop to improve the overall runtime and memory allocation. Those four steps are: i) Resource Monitoring that organizes the desired information by calculating the set rules and comparing the results with the given information before pre-processing. When the rate of transformation threshold is within the fixed limit, it sets rules to start pre-processing tasks in the local storages nodes. ii) Task Distributing Module organizes the pre-processing task of the nodes which are enlisted by resource monitoring module. Then it sorts the data based

on processing rules by choosing the nearby valid nodes within threshold limit. iii) Task Processing Module allows all the steps of pre-processing tasks for generating the processed data. Then it deploys the output in the storage nodes to generate new data input files iv) Input Analysis Model processes the pre-processed data and implements the result into the Map process where Map Reduce is used to divide the data into two systems- map and reduce to finally generate the standard file which we can use as per our desired input file.

Table 1. Comparison of Experiment Results

	Without pre-processing	With pre-processing
File Bytes Read	31893	5272
File Bytes Written	1274383	78901
Physical Memory Bytes	2810617856	2104272601
Merged Map Outputs	10	5
Failed Shuffle	0	0
GC Time Elapsed (ms)	2999	989
CPU Time Spent (ms)	8360	5190
Shuffled Maps	10	5
Total Committed Heap Usage	214176480	214176480
Spilled Records	240	97
Virtual Memory Bytes	21014081536	2101101612
Reduce Input Groups	622	419
Reduce Input Records	1890	1157
Reduce Output Records	870	714
Reduce Shuffle Bytes	3490	1084
Time (min)	1’ 58”	45.502”

Table 1 shows the comparison results of the Hadoop applications with and without the preprocessing system; each form ran MapReduce for the same input dataset. After preprocessing the CPU time indicates that the computational cost of the computation phase declines significantly. It shows that the MapReduce is IO-intensive and computational cost is comparatively low. Hence, in the running process the computing time of the program is negligible. With preprocessing, a noticeable decrease is witnessed for both the file bytes read and written, which indicates that IO data size of the disk file evidently declines after preprocessing, thus the total run time reduces significantly.

4.2 Feature Selection Approach

In the BIHS dataset there are total 126 modules which carry different information about 6500 households. The information includes education, employment, agricultural & nonagricultural foods& assets, usage of agricultural pesticides& chemicals, labor cost, summary of agricultural production, livestock and poultry, anthropometry, health, illness, household food consumption etc. We want to find out the nutritional based food habits of people, therefore, we need the relevant features only. For that we need to group the features which could be found in attributes like household id, food id, quantity consumption, quantity unit, own production of food, food cost, food preference as daily meal by different ages of people and intra household food distribution. Table 2 presents the features that are grouped based on different factors.

Table 2. Feature Classifier Parameters

Food Group Factors	Total consumption quantity	
	Food items’ unit price	
	Food intake choice of a household	
Household Member Factors	Household size	Adult
		Non-adult
		Infant
		Pregnant
	Age, height, weight, sex of household members	
Food Groups by Nutrient Elements	Macro-nutrient	

	Vitamin
	Mineral

MapReduce-based statistical test is applied to select the features. The input for the algorithm is an M x N matrix, where M is the total number of features and N is the sample size of the BIHS dataset. As discussed, the whole process is divided into two phases – map and reduce. In the map phase, every mapper reads a line (fn) which is running on the data node and calculates the necessary test statistic (si). By calculating the feature id (fi) and p-value (the probability of a given outcome under the null hypothesis) as the key-value pair (<fi,(si, pi)>), it sends the pair to an intermediary file. Then the reducer based on the p-value decides which features should be selected and which are not. After that it sends only the selected feature ids (< (fs1, fs2,……)>).

Friedman's nonparametric test is used in this research. It uses g-dependent groups with equal sample size (Kumar et.al , 2016). The null hypothesis is compared to the alternative hypothesis where at least one group is from different class. All the data vectors are then ranked (r) from ascending to descending orders for each of the classes. Here, the range of r is $\in[1, m]$ where m is the number of samples in each class. After that Friedman test is evaluated by the following equation (Kumar et.al, 2016):

$$F = \frac{12}{Mg(g+1)}\left(\sum_{k=1}^{g} R_k^2\right) - 3M(g+1) \quad (1)$$

Where, Rk = rank sum of Kth group, N =$\sum_{k=1}^{g} m_k$, total sample size and g = total no. of classes.

The selected relevant features give us the overview of households' information regarding all food habit related modules. After the feature selection method from 126 modules of BIHS dataset, we have gained relevant 6 module sets that contain all food items related information of different households. Table 3 contains the attributes that are used to calculate relevant features.

Table 3. Attributes of relevant modules

Module name	Attributes
Agricultural Production	Household id
	Quantity Harvested
	Quantity Consumed
Non-plot Food Consumption	Household id
	Produced
	Consumed
Food Inventory(7 Days)	Household id
	Quantity consumed
	Food cost
Household Meal Preference	Household id
	Time of meal (breakfast/lunch/dinner)
	Menu
Intra Household Food Distribution	Household id
	Member age
	Menu
Household Id	Division id
	District id

In this study, the data we have used is module driven. Each module consists of several attributes. Feature relevancy is selected according to our project goal and dimension reduction is done based on that at each of the selected modules. For example, Household id, Member age and Menu (Food Code) is selected as relevant attributes under 'Intra Household Food Distribution' module. Now from here food preference of infant, non-adult, adult and pregnant women can be extracted. Now in a particular area, one age group meal preference is considered as a feature. So for every district we can get a total of number of districts*number of age groups*number of meals per day times feature. Similarly 'Household Meal Preference' has the additional information about time of meal taken by every household. This information is related with our project goal as we are interested about the relationship between delayed meal intake behavior and nutritional impact due to it.

Again other modules also have distinct attributes like crop id, harvested consumption, own consumption etc. We are focusing on bring out as many features as possible, that can be related with total food consumption and disease relation. Consumption from harvested crops and market-bought items are featured in order to find out most intake crops and food items in a particular area. In order to find out nutritional gap, attributes like 'meal intake or not' (e.g., breakfast, lunch or dinner) also contribute to our selected feature list.

To avoid the "curse of dimensionality" this process is based on two hypothesis: null hypothesis (no significant difference between the properties of the classes) which is the discarded features and alternative hypothesis (at least one significant difference exits between the properties of classes) which is the accepted features based on the properties such as mean, median and variance. By considering the confidence interval (~99%) if p-value is <0.001 then it is called the null hypothesis (rejected) otherwise alternative hypothesis (accepted). After that categorization these features based on their p-values identifies the features with strong exemplifications. Table 4 describes the number of relevant features on which we are going to work for the proposed model. The relevant 721 Friedman features are extracted from the 4185 features.

Basically these relevant features carry information of the areas (divisions and districts) that are producing agricultural and non-plot foods the most. Based on this information and incomes of households, the types of foods they prefer are sorted out. After that household's daily meal preference for breakfast, lunch and dinner, food inventory for last 7 seven days and intra household food distribution depending on different age group of family members are extracted. Finally all these features are used to find out the food habit patterns of divisions and districts.

Table 4. Number of relevant features

Dataset	Friedman Features	Number of modules
BIHS: PSU#325 (4185)	721	6

After selecting suitable relevant features, Hadoop sequentially applies MapReduce in all data points in the testing dataset for classification. To ensure that the model is not over or under trained, every third sample is kept from the 6 modules dataset for testing and the remaining samples are used for training set. Table 5 records the distribution of training and testing records.

Table 5. Data Distribution of BIHS Dataset for training and testing

Dataset	Total Sample Size	Training Sample Size	Testing Sample Size
BIHS: PSU#325	6500	4333	2167

4.3 Finding Most Frequent 35 Food Items among Households

To analyze the food habit of households, these 6 module sets are ready to run the Word Count Map Reducing process to find the food frequency of each food item and adhoc Hive (built on top of Hadoop) queries to find the 35 frequent food items that are taken most among all the foods. We also count the amount of those 35 foods taken by each household per week.

From the training dataset 'Word Count' automatically monitors the resource monitoring nodes to deploy the monitoring module using frequency count technique. In this process, at first, it operates function in data file and then count the frequency output to give the result. Hadoop does this job into two steps: Map and Reduce. For mapping it splits data sets into every single data nodes using row formatting which is delimited by coma. Then each node builds sets of key (each word)/value (frequency of each word) of pairs and stores them in a new provisional file. Then the reduce part collects the partial output from each nodes (Dai et.al, 2016). In this phase, Word Count sums up the partial results getting from each node of map phase and computes the total frequency of all words. These new files replace the previous input data files. Without Map processing as shown in Figure 3 Hadoop module recognizes the pre-processed results directly through input analysis. The Map process portion only counts the data which are not pre-processed and gives the same result as standard process. Figure 3 shows clearly that the data size read from map process is hugely reduced after pre-processing. So it reduces the overall usage of disk I/O bandwidth and computational expense is relatively low. Therefore the running time of the whole process becomes very less that it can be negligible and the data transmission rate decreases to improve the runtime of whole application.

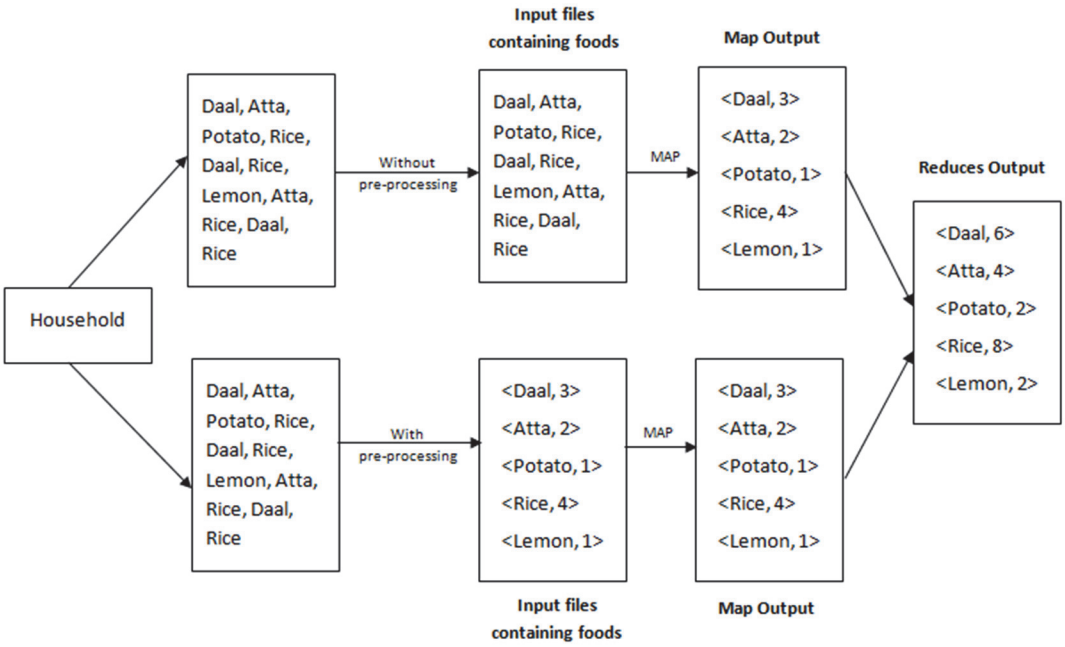

Figure 3. Analysis of all foods frequency per week using Word Count

After calculating all the food items frequencies we have selected the 35 most common food items that are taken by every household per week. Hive query is done as following to partition these 35 food items from all food items.

| IF FOOD_ID= GIVEN FOOD_CODE

| | IF FOOD_FREQUENCY=NOT NULL THEN

PARTITION FOOD_FREQUENCY BY HOUSEHOLD_ID

CLUSTER BY DIVISION_CODE AND DISTRICT_CODE INTO 7 BUCKETS

GROUP RESULT BY FOOD_FREQUENCY FROM ASCENDING TO DESCENDING ORDER

| | IF FOOD_FREQUENCY=NULL THEN RESULT=0

| IF FOOD_ID! = GIVEN FOOD_CODE THEN REPEAT THE FIRST STEP

The projection is done for getting quantity consumption of the 35 food items per week of the households of by using the unique household id and quantity consumption of foods. Relation algebra query is given in (2)-(3) to produce the results.

$$R1 = \left(\rho(Household_{Id})\right) \bowtie Household_{Id} = O1.A01(O1) \quad (2)$$

$$FC = \pi_{(Household_{Id}, Food_{Code}, Food_{Quantity}, Quantity_{unit})}(R1) \quad (3)$$

Here, A01 = Unique codes for every division, O1 = Quantity consumption, R1= Households under particular division, FC =Total amount of Food Consumption table.

So, the final projection contains each and every household's food habit per week which is shown in Table A1 (Frequency of 35 food items) and Table AII (Total amount of 35 foods taken per week, eg., Khulna Division) in the Appendix.

4.4 Selection of 20 Nutrient Elements in 35 Foods

20 major nutrition elements in food items are Energy, Protein, Carbohydrate, Fat, Fiber, Ca, Fe, Mg, P, Na, Zn, Cu, Vit-A, Vit-D, Vit-E, Thiamin Vit-B1, Riboflavin Vit-B2, Niacin Vitamin B3, Vitamin B6 Pyridoxine and Vit-C. Institute of Nutrition and Food Science, University of Dhaka have listed 381 foods and calculated the available nutrition present in per 100g of those foods (Shaheen et al., 2013). We have also used their table for nutrient calculation. The required food and nutritional values are extracted for the selected 35 food items in (Table A3) in the Appendix. Relation Algebra query is provided in equation (4) for this.

$$NutritionalFood, NF = \pi\_((Energy, Protein, \dots\dots\dots Vit - C)\ )\ \sigma\_((Food\_name = Rice, Atta, \dots, Tobacco)\ )\ (FCT)) \quad (4)$$

4.5 Fuzzy Inference System for Family Consumption Recommendation

This method approximates the amount of foods a family should consume every week based on standard average nutrient intakes. Firstly, the average estimation of daily individual intake is based on age and nutrient from 20 elements is calculated. A person needs daily a fixed amount of macro nutrients and vitamins &minerals according to his/her age and gender. The calculation of daily (Lenntech, 2017) recommendation for 20 nutrient (Australian food and Grocery Council , 2011) intakes per household is followed by equation 5:

$$Xi = \lim_{n-\infty} \int \left(\sum_{n=1}^{20} \left(\sum_{fm=1}^{i} \left(\frac{ar * a + nr * b + pr * c + ir * d}{fm_a + fm_b + fm_c + fm_d} \right) \right) \right) \quad (5)$$

where, Xi=daily recommended intake of nutrition, fm = total family member, ar =adult recommended intake, a = number of adult family members, nr = non- adult recommended intake, b = number of non-adult family members, pr = pregnancy period recommended intake, c = number of pregnant family members, ir = infant recommended intake and d= number of infant family members.

Using equation (6) weekly recommendation of 20 nutrients is estimated by:

$$"Wr" \cup\_"n = 1" \hat{}"20" \textstyle\sum "Xi \times Ai \times " " Wd " \quad (6)$$

Here, Wr = weekly recommendation, Xi = daily recommended intake of nutrition, Ai = average family members in rural areas (ESRI , 2017), Wd = total days of week.

Age, height, weight, gender, income range, no. of family members, no. of adults in family and area of living of an individual has taken as input. Height, weight and gender are used to calculate BMR (Basal Metabolic Rate). One of the most popular ST. Mifflin's formula (Orlov, 2017) has been used for calculation as follows:

$$\text{BMR\_Male} = (10 * \text{weight}) + (6.25 * \text{height}) - (5 * \text{age}) + 5 \quad (7)$$

$$\text{BMR\_Female} = (10 * \text{weight}) + (6.25 * \text{height}) \ ? \ (5 * \text{age}) \ ? \ 161 \quad (8)$$

In the Fuzzification process, BMR and age are inputs to predict the recommended nutrient for that individual. So, membership functions are being designed for each nutrient and set rules to get the amount of that nutrient which is recommended for that person based on his/her BMR and age.

For rule validation IOM's (Shaheen et al., 2013) findings are being used on average intake of every person based on age and gender as source. For example, suppose a person's age is low (0-18), BMR is average (1400-1800) with our calculation (eq. 7-8). Now according to IOM's report (Shaheen et al., 2013) the age ranging from 4-18 should have 45-65% of carbohydrate of total intake which is (770-990) in calorie. BMR gives total calorie need for individual's body and age which sets the percentage of nutrition should be taken according to his age. So, calculating all the cases, a rough measurement has been set so that the recommended carbohydrate should return a scale between the average limit (825-1200 calorie) for the above case. Figure 4 gives the snapshot of fuzzy rules that are generated in MATLAB.

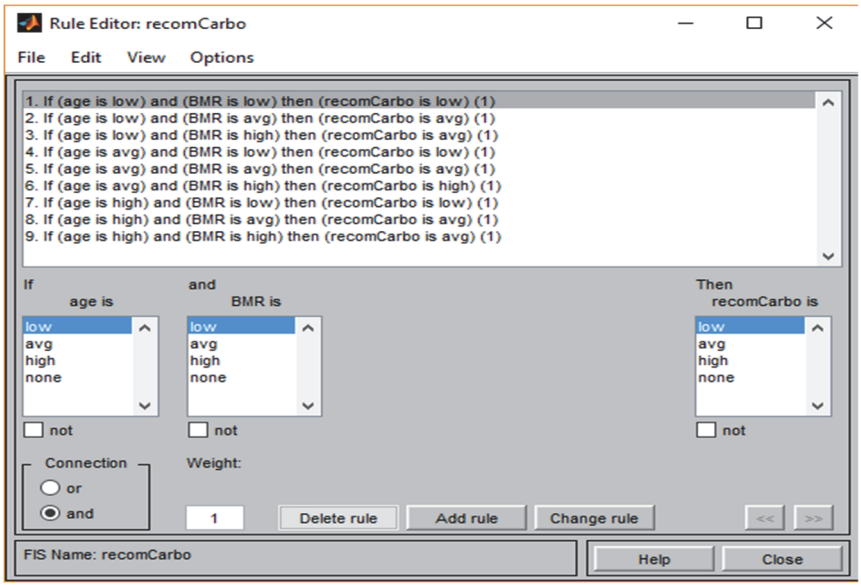

Figure 4. Rule for generating recommended carbohydrate value

In this process, the recommended value for carbohydrate for the individual is being calculated. Calculation of all other nutrient elements has been done by similar process. By equation (5) and equation (6) we know the proper amount of nutrients that should be taken by a household per week. Table 6 shows a sample weekly recommendation of nutrients for a household per week.

Table 6. Weekly recommendations

Nutrition Class	Recommended amount
Protein	1540000 mg
Energy	267960(KJ)
Calcium(Ca)	30800 mg
Zinc	462 mg
Magnesium(Mg)	10780 mg
Vitamin-C	2310 mg
Vitamin-D	0.45 mg
Iron(Fe)	4.62 mg
Vitamin-B12	.1848 mg
Phosphorous	19250 mg
Carbohydrate	9548000 mg

4.6 Prediction of Nutrition Deficiency

In the previous section we find out the amount of nutrient in each of those 35 food items based on the nutrient content report of (Shaheen et.al, 2013). Then, based on the number of family members we find the amount of nutrients per week a household should intake. We analyze and compare it to the amount of nutrients that a household currently takes per week. By this we are be able to find out the households that are suffered from nutrients deficiency and predict the possible diseases that the household could suffer in future. Table 7 shows the most common diseases of rural areas of Bangladesh which are caused by different nutritional deficiencies. By comparing between equation (9) and equation (10) we are getting the nutritional deficiency profiles of each household. By using this information we can evaluate the households that have higher probability of suffering from particular disease (s). It ultimately triggers the diseases information for all the districts and divisions.

Table 7. Diseases and causes

ICD 10 codes	Class Labels	Disease	Nutrition deficiency (Goenka, 2017)
E46	1	Protein-energy malnutrition	Protein , Energy
O150	2	Eclampsia	Protein , Calcium
D50.9	3	Iron deficiency anemia	Iron , Vitamin B-12
I219	4	Acute myocardial infarction	Vitamin D , Magnesium
D53.2	5	Scurvy	Vitamin C
M81	6	Osteoporosis	Calcium , Vitamin D
E55	7	Rickets	Vitamin D, Phosphorous , Calcium
P05.9	8	Growth retardation	Zinc

Table 7 depicts the diseases and possible cause of those diseases. ICD 10 is the International Classification of Diseases, Tenth Edition (ICD-10), a clinical cataloguing system. By analyzing this part, the amount of nutrient deficiency that can cause diseases is found. By using Hadoop the divisions and districts that have the most nutrient deficiencies and are affected by the diseases are found out. Algorithm 1 depicts the techniques to find disease pattern in districts and divisions of Bangladesh. To acquire a particular nutrient profile, projection is done in that nutrient from this table.

$$NC\ \prod_{1}^{20} J_i = \min_{Xi}\left(\left(W_{\emptyset}^{T} \triangleq \max_{Xi}(Xi - S)\right)\right) \tag{9}$$

Where, NC = Non-Communicable diseases based on nutrition deficiency, $\emptyset$ = Co-efficient vector id of foods, Xi = Recommended nutrition per week, T = 35 food items, S = intake amount of nutrition per 100gm of 35 food, W = Presence amount of nutrients, Ji = Less amount of nutrition then per week recommendation.

Algorithm to find disease patterns based on grouping of divisions and districts on Hive

|IF TOTAL INTAKE = RECOMMENDEDTHEN HOUSEHOLD = NO DEFICIENCY
PARTITION BY HOUSEHOLD ID
GROUP BY (DIVISION ID, DISTRICT ID)
|THEN GO TO NEXT

|IF TOTAL INTAKE < RECOMMENDED THEN HOUSEHOLD=DEFICIENCY
GET THE AMOUNT OF DEFICIENCY OF DIFFERENT NUTRITION'S
FIND THE NUTRIENT NAME THEN FIND DEFICIENCY BASED DISEASE
||IF DISEASE FOUND THEN
PARTITION IT BY DISEASE ID AND HOUSEHOLD ID
CLUSTERED BY (DIVISION ID AND DISTRICT ID) INTO 7 BUCKETS.
||IF DISEASE NOT FOUND THEN GO TO NEXT STEP

| IF TOTAL INTAKE > RECOMMENDED THEN HOUSEHOLD = OVERDOSE
GET THE AMOUNT OF OVERDOSE OF DIFFERENT NUTRITION'S
FIND THE NUTRIENT NAME THEN FIND OVERDOSE BASED DISEASE
| | IF DISEASE FOUND THEN
PARTITION IT BY DISEASE ID AND HOUSEHOLD ID
CLUSTERED BY (DIVISION ID AND DISTRICT ID) INTO 7 BUCKETS.
| |IF DISEASE NOT FOUND THEN GO TO FIRST STEP
BREAK

4.7 MRK-NN Classification

MapReduce-Based K-NN (mrK-NN) is built to create the model of the classified nutritional training datasets that

are obtained from feature selection paradigm (nutritional profiles of households). The training of mrK-NN is done here by using 3-fold cross validation to obtain the parameter k.

To run the mrK-NN algorithm in Hadoop Framework the whole training dataset is now divided into two phases – map and reduce. In the map phase, mappers read every single line sequentially from the file to process input and calculate the output (e.g. distances in mrK-NN). The values with their equivalent keys (e.g. nutrition deficiency classes, class labels) are stored to the intermediate file which is sent to the reducers for processing the input and writing the result in HDFS.

In this section, each mapper reads a sample datum from the testing set to calculate the Euclidean distance between training and testing samples. Thus it stores distances from all the training samples through their class labels (Table 4). The mapper then sends the testing samples id and distances to the file system. Then the reducer phase classes the distances in ascending order to select the K-nearest training samples. The testing sample is allocated to a unique class that corresponds to the modal class of those K-training samples as shown in mr-KNN algorithm. After that the reducer phase produces the instance id to assign classes to the testing sample. Finally, the whole process will evaluate majority of the K-training samples as usual in the K-NN process and the reducer carry out the instance id to assign class labels of the testing sample. As the MapReduce paradigm is used here so we have to make sure that the input form is <key, value> pair where key is the class labels and value is the nutrition classes.

Let, training Dataset (TRD) and Testing Dataset (TSD) are stored in the HDFS. The Training data points are stored in the TrainFile (TRF) and Testing data points are stored in the TestFile (TSF) that holds the data vectors as vectors. After that, K-NN is applied in the distributed environment to design the map and reduce methods in it (Anchalia & Roy et al., 2014). Here training dataset (TRD) is divided into small training points to classify the nutritional deficiency for different class labels for every household. Such as, if a household is taking fewer amounts of protein and calcium than recommended amount per week, it is classified as class label 1 and so on. TRD is carrying all the 20 nutrient elements classified by different labels as per disease and training data points are carrying each class labels separately to monitor the model. The diseases of the households are classified by the class labels according to the distance of 'K' neighbors in increasing which is selected by majority vote. After that it sets the training data point's label as the label for a particular disease class with that majority vote count as shown in following algorithm.

Mapper for K-NN

Start

- Create list for storing data points in TSD
- Load TSF
- Update testList comparing with TSF
- Open TRF
- Calculate Euclidean distance(TRD, TSD)
- Write distance of test data points from all training data points
- Labelling every class in ascending order
- Check TSF<= TSD(distance ,class label)
- Call Reducer for K-NN

End ()

Reducer for K-NN

Start

- Load value of K, TSF
- Open TSF to read Testing data points
- Initialize Count = 0 for all class labels
- From 0 to K
- Count = Count + 1
- Find Highest Counts for Testing data points by assigning labels
- Add classified Testing data points with Output File
- Update The System

End ()

Implementation of Mapper and Reducer Functions in K-NN Algorithm

Start

- Read value of K
- Set paths for TRD and TSD directories by adding TRF and TSF
- Create New Class
- Set defined Mapper to map class and Reducer to reduce class
- Set paths for output directory
- Submit Class

End ()

5. Results and Analysis of Results

We have divided this section into different subsections. Those are described below

5.1 Experimental Setup

The whole model is executed on Hadoop cluster with 3 commodity PCs which are connected by Linksys wireless router (WRT54GSUK) to share data between them. Details of the configuration process is given below:

5.1.1 Software Requirements

- Ubuntu 16.04: User-friendly for Hadoop
- JDK 1.8 : Java is used as soul programming language of Hadoop
- Hadoop 2.7.1 : MapReduce implementations and Hive queries
- Maximum mapper tasks: 10
- Maximum reducer tasks: 1
- MATLAB Fuzzy Tool Box: designing and analyzing Fuzzy Inference System

5.1.2 Hardware Requirements

All the experiment is executed on a cluster which is designed of – one master node and two slave nodes.

- Master node : Name Node 1, Intel(R) Core(TM) i5 – 4200CPU , RAM 8 GB, Hard disk 250 GB
- Slave node 1 &2 : Data Nodes ,Intel (R) Core(TM) 2 DUO CPU, T5870 , Ram 4 GB , Hard disk 250 GB

5.2 Class wise Data Distribution

There are 6500 samples (households) in the BIHS Dataset. We have used 2167 samples as testing set and 4333 as training set. The data is divided into 8 classes. Table 8 shows the BIHS dataset class label and number of test samples in each class. The classes are categorized based on the nutrition deficiencies that together trigger one particular disease. For example, if a household intakes less amount of Vitamin – D and Calcium than required then the family members of that household will have higher chance of suffering by Osteoporosis. Again if a household intakes Calcium almost near to the amount as they need but lacks of Magnesium and Vitamin-D then it will have higher chance of suffering from acute myocardial infarction than Osteoporosis. So all the nutrition deficiency classes are defined by indicating the particular diseases that can be happened due to deficiency of one significant nutrient or many nutrient together using Table 8.

Table 8. Class label and number of test samples in each class

Nutrition deficiency Classes	Class Label	Sample numbers
Protein, Energy	1	180
Protein, Calcium	2	368
Iron, Vitamin B-12	3	270
Vitamin-D, Magnesium	4	418
Vitamin C	5	155
Calcium, Vitamin-D	6	331
Vitamin-D, Phosphorous, Calcium	7	289
Zinc	8	156

5.3 Performance of mrK-NN Classifier

Our proposed mrK-NN algorithm is applied to classify the testing dataset with the reduced features and the training data from Friedman method. It is trained by 3 fold cross validation by varying the K parameter between the range [1, 21]. After getting the (median) values of K from each fold, the average training accuracy is estimated. Here from the three properties of feature selection (mean, median and variance) we have used the median values as they are more robust to outliers compared to rest of properties. The average accuracy is the training accuracy value of the proposed model with the median K values for each fold. The testing accuracy of the model is gained by using the optimal K values on the test data set of the model.

Here the Span of 2 within the given range of K[1,21] means that at first the training accuracies are gained using Friedman test as feature selection method. After that corresponding to these training accuracies, by using the

optimal values of K and the testing accuracies gained using Friedman test as feature selection method are used to test the model. This whole process represents the performance for a multi-class confusion matrix M with N classes, such as, when K=19, the training accuracy is 79.49% which is obtained from the Friedman Feature selection method. Then the model is evaluated for obtaining the testing accuracy which is 78.10% by using the corresponding training accuracy and optimal value of K. Accuracy is calculated by using the following formula:

$$Accuracy = \sum_{k=1}^{21} \frac{\sum mf - \sum af}{\sum af} x100 \quad (10)$$

Where, mf = measured features, af = accepted features, K = parameter of mrK-NN

The performance of the parameters of ith class is calculated by the summation of rows and columns of M matrix with n classes (Kumar et.al (2016)):

Table 9. Confusion Matrix

Output Class	Target Class	
	negative	positive
Classifier for negative	tn	fn
Classifier for positive	fp	tp

Note: tn= true negative, fn=false negative, fp= false positive, tp=true positive

Here,

$$\text{Recall} = \frac{M_{ii}}{\sum_j^n M_{ji}} \quad (11) \text{ and Precision} = \frac{M_{ii}}{\sum_j^n M_{ij}} \quad (12)$$

Figure 5 represents the confusion matrix for testing the dataset reduced by Friedman test. The values of precision and recall for every class are shown in the last row and last column of the matrix.

Output Class \ Target Class	1	2	3	4	5	6	7	8	9	
1	155 1.7%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	1 0.1%	8 1.1%	0 0.0%	68.1% 31.9%
2	0 0.0%	289 10.3%	6 0.9%	0 0.0%	0 0.0%	1 0.1%	0 0.0%	0 0.0%	1 0.1%	54.3% 45.7%
3	0 0.0%	0 0.0%	180 7.8%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
4	6 0.9%	0 0.0%	0 0.0%	331 11.1%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	78.9% 21.1%
5	0 0.0%	0 0.0%	1 0.1%	0 0.0%	418 18.2	0 0.0%	1 0.1%	0 0.0%	0 0.0%	61.7% 38.3%
6	0 0.0%	2 0.3%	0 0.0%	0 0.0%	0 0.0%	155 1.7%	0 0.0%	7 1.0%	0 0.0%	81.2% 18.8%
7	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	368 16.3%	0 0.0%	0 0.0%	100% 0.0%
8	4 0.6%	0 0.0%	0 0.0%	0 0.0%	2 0.3%	0 0.0%	0 0.0%	180 7.8%	1 0.1%	74.7% 25.3%
9	0 0.0%	0 0.0%	0 0.0%	3 0.4%	1 0.1%	0 0.0%	0 0.0%	0 0.0%	156 3.2%	86.2% 13.8%
	76.2% 23.8%	68.9% 31.1%	100% 0.0%	67.3% 32.7%	75.1% 24.9%	55.9% 44.1%	100% 0.0%	88.3% 11.7%	74.3% 25.7%	78.1% 21.9%

Figure 5. Confusion matrix for mrK-NN classifier with Friedman using BIHS: PSU#325 dataset (f = 2180, K = 19)

Table 10 represents a view of accuracies for different values of K parameter by using the confusion matrix and equation (10)-(12).

Table 10. Training and Testing accuracy (%) for various ranges of parameter K

Dataset	Freidman	
BIHS: PSU#325	Training Accuracy	Testing Accuracy
K=1	63.84	61.60
K=7	70.33	69.50
K=14	77.05	75.20
K=19	79.49	78.10
K=21	85.71	84.20

5.4 Run Time Calculation

By implementing MapReduce in K-NN method we have found a good reduction of computation time which was previously slower in the standard K-NN version when the mapper's numbers are increased. First we have performed the whole analysis based on the original standard version of K-NN algorithm to set our baseline. Since our dataset size is not that much large, our block size is reduced so that all the data can be distributed equally to all the nodes for utilizing the resource and all data nodes process equivalently. Here 2MB (defaults are 64MB or 128MB) block size (Kumar et al., 2016) is considered to run an individual mapper. After that, the total time taken by the Hadoop cluster using mrK-NN is compared to the original K-NN algorithm. From this it shows that the total time taken by mrK-NN on the Hadoop cluster consumes much less than the K-NN system as data size increases. Hence the higher values of parameter k gives better timing for the same accuracy that original K-NN gives us.

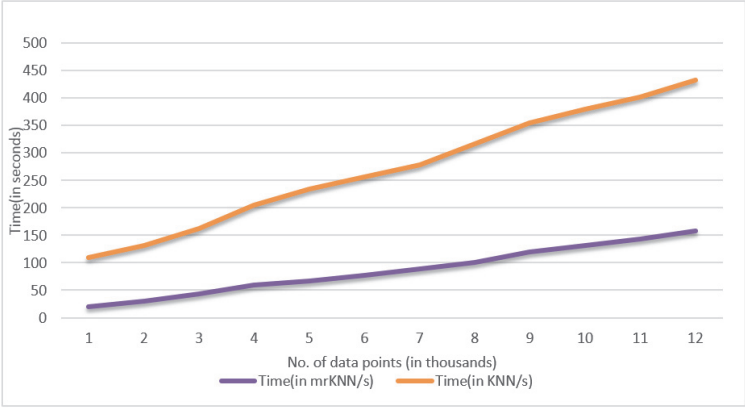

Figure 6. Graph of standard K-NN vs. mrK-NN

5.5 Disease Prevalence in Different Areas of Bangladesh

By analyzing the food habit patterns of different rural areas of Bangladesh we have tried to find out the disease patterns that are more likely to occur in different parts of Bangladesh. For this we have generated every household's nutritional profile per week and compared them with the recommended nutritional profile per week. Based on that we have tried to find out different diseases that could occur due to nutritional deficiency.

Though in rural areas of Bangladesh the death rate caused by various diseases have been reduced but still every year many people die because of malnutrition. Most people in rural areas are not fully aware of food nutrition. The income of rural people is not high so they cannot buy the foods to get rid of the diseases that are occurred due to nutrition deficiency. We have evaluated the 64 districts' disease patterns and probable occurrence rates of the diseases such as- Protein-Energy Malnutrition, Eclampsia, Scurvy, Acute Myocardial Infarction, Iron deficiency Anemia, Osteoporosis, Rickets, Growth Retardation etc. We have tested our results by using the reports of 65 Civil Surgeon Offices - Completed (Local Health Bulletin – 2016) which is conducted by Ministry of Health and Family Welfare (MOHFW) and data belong from January to December, 2015 (DGHS, 2016).

For instance, by analyzing the nutritional profiles of Khulna division Table 11 and its districts we have found out that in Khulna Sadar, there is mainly lack of Vitamin – D, Calcium, Magnesium which estimates that Khulna Sadar will face diseases more such as, Osteoporosis , Acute Myocardial Infarction, Eclampsia etc.

Table 11. Percentage (%) of diseases occurrence in Khulna Sadar

Nutrition Deficiency	Disease Name	Disease Pattern	Percentage
Protein, Energy	Protein - Energy Malnutrition	Infant & Toddler	12.56%
Zinc	Growth Retardation		3.05%
Vitamin-D, Phosphorous, Calcium	Rickets	Toddler & Children Mainly	6.90%
Protein, Calcium	Eclampsia	Pregnancy Death	22.61%
Iron, Vitamin-B12	Iron Deficiency Anemia		2.20%
Calcium, Vitamin-D	Osteoporosis	30+ Women Mainly	18.52%
Vitamin-D, Magnesium	Acute Myocardial Infarction	30+ Age People	22.62%
Vitamin-C	Scurvy	All	4.13%

Again, in another district of Khulna division like Bagerhat we have found out by comparing their nutritional profile Table 12 with Table 8 that children are more likely to be suffered from deficiencies of protein, energy and zinc than the adult members of a family. It estimates that here infants, children will face a high rate of protein-energy malnutrition and growth retardation problem than any other diseases.

Table 12. Percentage (%) of diseases occurrence in Bagerhat District

Nutrition Deficiency	Disease Name	Disease Pattern	Percentage
Protein, Energy	Protein - Energy Malnutrition	Infant & Toddler	20.76%
Zinc	Growth Retardation		16.05%
Vitamin-D, Phosphorous, Calcium	Rickets	Toddler & Children Mainly	4.91%
Protein, Calcium	Eclampsia	Pregnancy Death	0.61%
Iron, Vitamin-B12	Iron Deficiency Anemia		0.40%
Calcium, Vitamin-D	Osteoporosis	30+ Women Mainly	12.52%
Vitamin-D, Magnesium	Acute Myocardial Infarction	30+ Age People	4.30%
Vitamin-C	Scurvy	All	8.94%

From the two districts of same division (Khulna) we can clearly estimate that only because of different food habits of different districts how disease patterns differ in a very significant way.

Again, by analyzing the nutritional profiles of Dhaka division and all the districts under it , we have evaluated that almost all the districts have a common pattern of facing Iron-deficiency Anemia related Pregnancy deaths cases where the rate is at least up to 20% in almost every district which is a very serious issue. It gives us a strong recommendation that in Dhaka division people are more likely to be suffered from iron deficiency based diseases as their daily meal preferences lack of iron.

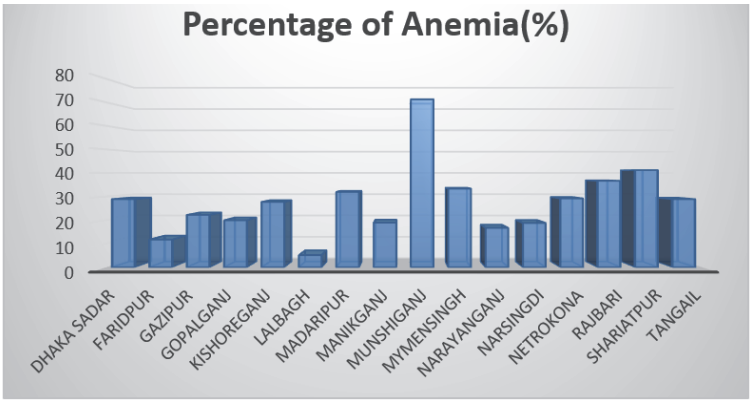

Figure 7. Anemia percentage of all districts in Dhaka Division

Table 13. Result of deficiency of Calcium, Vitamin-D and Phosphorous for Chittagong division per week

Nutrient	Recommended intake	Consumed intake
Calcium(mg)	30800	12050.31
Vitamin-D(mg)	0.45	0.182
Phosphorus(mg)	19250	13800.27

In Sylhet division and Chittagong division, there seems higher deficiency in Calcium, Vitamin D and Phosphorous intakes. Households are taking much less amount of these nutrients per week than recommended intake per week. As a result the areas have higher possibilities to be affected by Rickets. Among all the districts of Chittagong, Cox's Bazar has the highest rate (~7.88) of Rickets affected children. In The daily Star (Zannat,2008) and (Mahmood, 2013) also written in their report that children of Bangladesh particularly of Sylhet and Chittagong divisions suffer more from Rickets. In (Mahmood, 2013) the authors reported the rate of Rickets in Cox' Bazar district is (8.7% at least) which also validates our evaluation nearly. From our analysis a common deficiency of Vitamin D and Calcium has been observed in almost every district especially in adult women. As a result they have a very high possibility to suffer from Osteoporosis (One in three older women suffer from Osteoporosis, 2015).

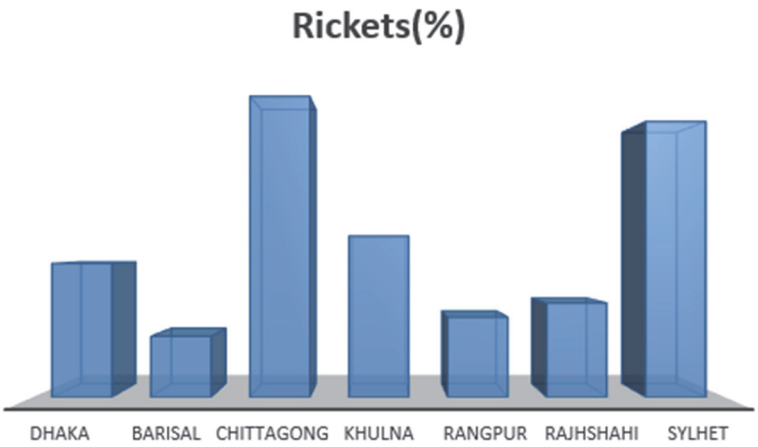

Figure 8. Variations of Rickets among divisions

Again, when we consider about Vitamin-C deficiency based diseases mainly Scurvy we have found in Barisal division (figure 9). This disease rate is higher in Barisal compared to any other divisions. As almost in all districts of Barisal Division by analyzing their households' food preferences we have found that they do not consume foods that generally have Vitamin-C. As a result a significant amount of people there have higher probability of suffering from Scurvy which can cause mouth cancer as well.

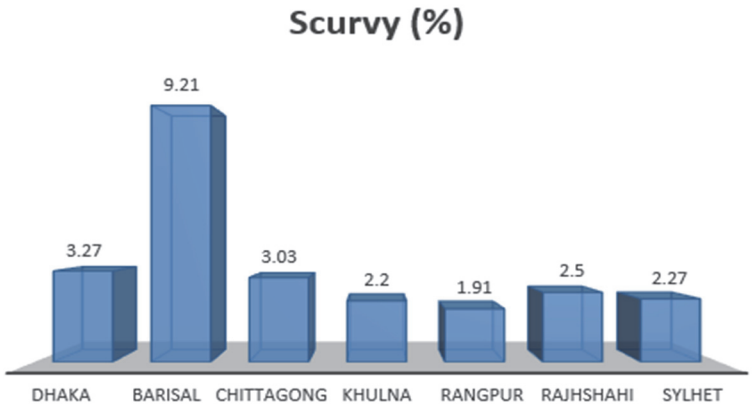

Figure 9. Variation of Scurvy among divisions

Similarly, these types of evaluations for the 8 main diseases that we have targeted have been done for all the 7 divisions and 64 districts to find the diversity of nutrient based diseases so that mass people can have an accurate idea of the foods that should be taken according to avoid the diseases that are generally prevailed in their areas.

From our analysis and calculation we have finally estimated that among the 64 districts and 7 divisions there seems some relevant patterns that can be followed in future for more accurate predictions. Such as in the North side of Bangladesh like - Rangpur and Rajshahi divisions, there seems more deficiency of protein-energy comparing to other nutrients. Therefore in all districts under those two divisions have higher chances of facing Protein-Energy Malnutrition. Again, in Dhaka division there seems more deficiency of iron, calcium and vitamin-D in rural areas. As a result pregnancy deaths because of Eclampsia and Anemia, Osteoporosis in women (30+ age) are more prone here. Moreover, in the South Side such as – Barisal & Chittagong divisions, there people face more problems regarding Scurvy, Growth Retardation (infant & toddlers) and Rickets (mainly toddlers & children) due to deficiency of vitamin-C, zinc, vitamin-D and calcium. So, hopefully all these information can be used as a predictive model for the government of Bangladesh to prepare division and district wise nutritional guidelines for the rural people and thus to proceed towards making a healthy Bangladesh.

6. Conclusion and Future Work

Bangladesh being a developing country is facing against various crises. Food and lack of nutrition falls under these issues, especially for rural people. The main objective of this research is to help the rural people for the betterment of their lives with the correct intake of required nutrients. From the BIHS dataset the most common 35 foods are selected. Map reducing framework on Hadoop, Fuzzification is applied to make the whole model more precise and accurate. Importance has been given on feature selection model based on the statistical test and classification of various diseases. This whole initiative is taken to predict different disease prone rural areas all over Bangladesh which will allow taking early intervention and treatment. This research also tries to help the government of Bangladesh to take necessary steps by predicting their nutritional deficiency and disease pattern.

However there are many other food items beyond the selected 35 items that have not been considered in this research and that is why we could not achieve 100 percent accuracy in nutrition calculation. Again, there are many other factors like physical inactivity, unhygienic food, impure water, smoking or drug addiction, etc. that adds to the reason of a disease. Moreover, from the dataset we have got only 6500 households information which is not enough compared to the huge population of Bangladesh. Though the BIHS data set contains different information about rural households but many of the data fields are empty.

As a future work, we plan to implement a user- friendly software (mobile) application to calculate the amount of nutrients people take and compare that with the dietary reference intakes based on BMR, age etc. The software will have the ability to generate a food list suggestion having proper balance of all nutrients according to our food habit. In this research 35 food intake items are chosen and 20 nutrients based 9 diseases are classified. We plan to extend the list by adding more foods and disease patterns so that the research can get more accuracy and mass people can be benefited from it. Moreover, the whole system can be extended for generating more precise results by using deep learning, ANN, neural networking etc. in Hadoop framework and compare the work in the Spark framework.

References

Ahmed, A., Ahmad, K., Chou, V., Ricardo, H., Menon, P., Naeem, F., … Yu, B. (2013). The Status of Food Security in the Feed the Future Zone and Other Regions of Bangladesh: Results from the 2011–2012 Bangladesh Integrated Household Survey, International Food Policy Research Institute. Retrieved from http://ebrary.ifpri.org/cdm/ref/collection/p15738coll2/id/127518

Ahmed, A., Tauseef, S., & Ghostlaw, J. (2016, December 19). International Food Policy Research Institute (IFPRI). Bangladesh Integrated Household Survey (BIHS) 2015 [dataset]. https://doi.org/10.7910/DVN/BXSYEL

Anchalia, P. P., & Roy, K. (2014). The k-Nearest Neighbor Algorithm Using MapReduce Paradigm. Fifth International Conference on Intelligent Systems, Modelling and Simulation. https://doi.org/10.1109/ISMS.2014.94

Araújo, P., Ferraz, B., Lima. B., Machado, M., Luz, P., Cruz, M., … Nascimento, N. (2017, November 27). Effect of mineral status and glucocorticoid use on bone mineral density in patients with cohn's disease. Elsevier. Nutrition, 47, pp. A1-A6. https://doi.org/10.1016/j.nut.2017.10.016

Australian Food and Grocery Council. (2011). Daily intake guide : Healthy eating made easy. Retrieved from http://www.mydailyintake.net/daily-intake-levels

Bazzano, L. A., He, J., Ogden, L.G., Loria, C. M., Vupputuri, S., Myers, L., & Whelton, P. K. (2002). Fruit and vegetable intake and risk of cardiovascular disease in US adults: the first National Health and Nutrition Examination Survey Epidemiologic Follow-up Study. *The American Journal of Clinical Nutrition, 76*(1), 93-9. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/12081821

Bharill, N. & Tiwari, A. (2014). Handling Big Data with Fuzzy Based Classification Approach, 312, pp. 219-227. https://doi.org/ 10.1007/978-3-319-03674-8_21

Dai, H., Zhang, S., Wang, L., & Ding, Y. (2016). Research and implementation of big data preprocessing system based on Hadoop. Big Data Analysis (ICBDA), 2016 IEEE International Conference on https://doi.org/10.1109/ICBDA.2016.7509802

Directorate General of Health Services (DGHS). The Ministry of Health & Family Welfare, Bangladesh (MOHFW), Civil Surgeon Office, Local Health Bulletin, 2016. Retrieved from http://app.dghs.gov.bd/localhealthBulletin2016/publish/

ESRI (2017, June 28).Bangladesh Average Household Size. Retrieved Nov. 16, 2017 from https://www.arcgis.com/home/item.html?id=692cee7e5a5e47dd86531ab0c6a00cff

Goenka, S. (2017, September 18). 10 Diseases Caused By Nutritional Deficiencies. Stylecraze. Retrieved Nov. 16, 2017 from http://www.stylecraze.com/articles/diseases-caused-by-nutritional-deficiency/#gref

International Classification of Diseases, Tenth Revision, Clinical Modification (ICD-10-CM) (2016, April 15). Centers for Disease Control and Prevention. Retrieved Nov. 16, 2017 from https://www.cdc.gov/nchs/icd/icd10.htm

Khan, M. K. (2017, October 29). Birth weight and growth of our babies - where are we? The Daily Star. Retrieved from http://www.thedailystar.net/health/birth-weight-and-growth-our-babies-where-are-we-1483042

Krbez, J. M., & Shaout, A. (2013). Fuzzy Nutrition System. *International Journal of Innovative Research in Computer and Communication Engineering, 1*(7), 1360-1371. Retrieved from https://www.ijircce.com/upload/2013/september/1_Fuzzy.pdf

Kumar, M., Rath, N. K., & Rath, S. K. (2016) Analysis of microarray leukemia data using an efficient MapReduce-based K-nearest-neighbor classifier. Journal of Biomedical Informatics, 60, 395-409. https://doi.org/10.1016/j.jbi.2016.03.002.

Kumar, N. M. S., Eswari, T., Sampath, P., & Lavanya, S. (2015). Predictive Methodology for Diabetic Data Analysis in Big Data. Elsevier, Procedia Computer Science, 50, 203-208. https://doi.org/10.1016/j.procs.2015.04.069

Lenntech (2017). Recommended daily intake of vitamins and minerals. Retrieved from https://www.lenntech.com/recommended-daily-intake.htm

Mahmood, I. (2013). Rickets Treatment for Children in Bangladesh.[online] Hope Foundation for Women & Children of Bangladesh. Retrieved Nov. 16, 2017 from https://www.globalgiving.org/projects/rickets-surgery-for-children-in-bangladesh/

Maillo, J., Triguero, I., & Herrera, F. (2015). A Mepreduce-Based K- Nearest Neighbor Approach for Big Data Classification. IEEE Trustcom. https://doi.org/ 10.1109/Trustcom.2015.577

Maitrey, S., & Jha, C. K. (2015). MapReduce: Simplified Data Analysis of Big Data, pp.563-571. https://doi.org/10.1016/j.procs.2015.07.392

Moktadir, S. A. (2015, May 17). If you suffer from scurvy. [online] The Daily Star. Retrieved Nov. 16, 2017 from http://www.thedailystar.net/health/if-you-suffer-scurvy-5685

Nguyen, D. D., Ngo, L. T., & Pham, L.T (2013). Interval Type-2 Fuzzy C-means Clustering using Intuitionistic Fuzzy Sets, pp. 299-303. https://doi.org/10.1109/WICT.2013.7113152,01

One in three older women suffer from Osteoporosis.(2015, October 21).[online] The Daily Star. Retrieved Nov. 16, 2017 from http://www.thedailystar.net/health/one-three-older-women-suffer-osteoporosis-160051

Orlov, A. (2017, September 17). How to Calculate Your BMR (And Why It Matters). Daily Burn Life. Retrieved from http://dailyburn.com/life/health/how-to-calculate-bmr/

Priyono, R. A., & Surendro, K. (2013). Nutritional Needs Recommendation Based on Fuzzy Logic. Proceduia Tecnology, 11, pp. 1244-1251. https://doi.org/10.1016/j.protcy.2013.12.320

Shaheen, N., Rahim, A., Mohiduzzaman, M., Banu, C. Bari, L., Tukun, A., … Stadlmayr, B. (2013). Food

Composition Table for Bangladesh. ISBN: 978984337522-3. Retrieved from http://www.fao.org/fileadmin/templates/food_composition/documents/FCT_10_2_14_final_version.pdf

The Daily Star , Heath news , Retrieved from http://www.thedailystar.net/health

Zannat, M. (2008, June 04). 5 lakh children suffer from rickets. [online] The Daily Star. Retrieved Nov. 16, 2017 from http://www.thedailystar.net/news-detail-39576

Appendix

Table A1. Frequency of 35 food items (All Divisions and Districts)

		Division													
Food Code	Food Name	Khulna	%Khulna	Rajshahi	%Rajshahi	Dhaka	%Dhaka	Rangpur	%Rangpur	Sylhet	%Sylhet	Barishal	%Barishal	Chittagong	%Chittagong
1	Rice	481	51.78	410	44.13	1425	100	412	44.35	753	81.05	706	76	963	100
7	Atta	121	13.02	148	15.93	231	24.87	38	4.09	196	21.1	212	22.82	376	40.47
8	Flour	4	0.43	13	1.39	13	1.39	5	0.54	26	2.81	17	1.83	37	3.98
21	Daal	339	36.49	291	31.32	931	100	152	16.36	464	49.95	653	70.29	634	68.25
31	Soyabean Oil	437	47.04	390	41.98	1042	100	301	32.4	602	64.8	627	67.49	783	84.28
34	Ghee	2	0.22	4	0.43	4	0.43	4	0.43	3	0.32	1	0.11	2	0.22
57	Green chili	442	47.58	420	45.21	1218	100	337	36.28	602	64.8	293	31.54	775	83.42
59	Kochu	32	3.44	10	1.08	18	1.94	12	1.29	7	0.75	78	8.4	18	1.94
61	Potato	443	47.69	490	52.74	1300	100	389	41.87	703	75.67	634	68.25	906	97.52
64	Onion	435	46.82	494	53.17	1325	100	393	42.3	723	77.83	634	68.25	923	99.35
65	Garlic	394	42.41	458	49.3	1343	100	332	35.74	641	67	558	60.06	842	90.64
86	Pui Shak	51	5.49	51	5.49	47	5.06	27	2.91	32	3.44	48	5.17	36	3.88
87	Lal Shak	95	10.23	37	3.98	163	17.55	34	3.66	131	14.1	277	29.82	170	18.3
121	Beef	57	6.14	56	6.03	233	25.08	55	5.92	94	10.12	67	7.21	219	23.57
122	Mutton	13	1.4	20	2.15	14	1.51	10	1.08	6	0.65	3	0.32	11	1.18
123	Chicken	116	12.49	134	14.42	380	40.9	84	9.04	202	21.74	179	19.27	272	29.28
127	Liver	1	0.11	0	0	0	0	1	0.11	3	0.32	3	0.32	5	0.54
130	Egg	290	31.22	310	33.37	805	86.65	217	23.36	343	36.92	440	47.36	516	55.54
132	Milk	95	10.23	195	20.99	527	56.73	127	13.67	156	16.79	131	14.1	258	27.77
142	Banana	65	7	65	7	221	23.79	33	3.55	101	10.87	126	13.56	97	10.44
160	Lemon	9	0.97	4	0.43	33	3.55	1	0.11	23	2.48	19	2.05	13	1.4
161	Dates	1	0.11	1	0.11	4	0.43	2	0.22	2	0.22	7	0.75	2	0.22
216	Mola Fish	3	0.32	9	0.97	37	3.98	5	0.54	28	3.01	1	0.11	28	3.01
217	Dhela Fish	21	2.26	0	0	1	0.11	0	0	1	0.11	15	1.61	1	0.11
219	Kachki	1	0.11	1	0.11	7	0.75	0	0	8	0.86	4	0.43	19	2.05
252	Salt	349	37.57	483	52	1324	100	384	41.33	728	78.36	637	68.57	914	98.39
254	Coriander	49	5.27	93	10.01	606	65.23	12	1.29	450	48.44	18	1.94	195	21
255	Ginger	43	4.63	343	36.92	586	63.08	131	14.1	185	19.91	177	19.05	350	37.67
257	Black Cumin	3	0.32	5	0.54	15	1.61	22	2.37	8	0.86	15	1.61	6	0.65
266	Sugar	242	26.05	229	24.65	695	74.81	162	17.44	581	62.54	471	50.7	652	70.19
269	Tea	36	3.88	63	6.78	103	11.09	69	7.43	489	52.64	334	35.95	431	46.39
270	Badam	17	1.83	45	4.84	110	11.84	5	0.54	4	0.43	10	1.08	10	1.08
271	Honey	3	0.32	5	0.54	5	0.54	1	0.11	1	0.11	4	0.43	1	0.11
309	Gaja	6	0.65	9	0.97	27	2.91	20	2.15	15	1.61	4	0.43	11	1.18
314	Tobacco	86	9.26	217	23.36	543	58.45	171	18.41	384	41.33	141	15.18	275	29.6

Table A2. Total amount of 35 foods taken per week (Khulna division)

Food code	Food name	Energy(kJ)	Protein(gm)	Carbohydrate(gm)	Ca(mg)	Fe(mg)	Mg(mg)	Zn(mg)	Vitamin-D(mg)	Vitamin-C(mg)	Vitamin-B(mg)
1	Rice	1004451.88	11304.28	87913	49046.793	1690	30845	1891.67	1571.91	2290.21	395.731
7	Atta										
8	Flour										
21	Lentil										
31	Soyabean oil										
57	Green chili										
59	Kochu										
61	Potato										
64	Onion										
65	Garlic										
86	Pui shak										
87	Lal shak										
121	Beef										
123	Chicken										
130	Egg(1=62g)										
132	Milk										
142	Banana(1=120g)										
160	Lemon(1=19g)										
161	Dates(1=24g)										
217	Dhela fish										
254	Coriander										
255	Ginger										
266	Sugar										
269	Tea										
270	Badam										
271	Honey										
2521	Salt										
2522	Salt with iodine										

Table A3. All 20 nutrients calculation for the food items per 100 gm

Food Name	Energy (kJ)	Protein (g)	Fat (g)	Carbohydrate (g)	Fibre (g)	Ca(mg)	Fe (mg)	Mg (mg)	P (mg)	Na (mg)	Zn (mg)	Cu (mg)	Vit-A (mcg)	Vit-D (mg)	Vit-E (mg)	Thiamin Vit-B1(mg)	Riboflavin Vit-B2(mg)	Niacin Vitamin B3 (mg)	Vitamin B6 Pyridoxine(mg)	Folate (mcg)	Vit-C (mg)
Rice (Parboiled)	1460	7	0.3	76	3.4	1	0.7	43	127	2	1.33	0.2	0	0	0.8	0.8	0.03	1	0.16	11	0
Atta	1440	11	1.9	66	8	35	4.2	100	200	13	2.2	0.4	0	0	0.6	0.3	0.1	5	0.25	25	0
Flour	1470	9.8	1	73.1	2.7	13	2.7	58	140	10	1.55	0.19	0	0	0.06	0.12	0.07	4	0.044	20	0
Daal	1200	23.7	1.2	60.9	0.7	69	7.2	147	315	30	2.73	1.66	3	0	1.9	0.36	0.14	6.6	0.5	140	0
Soyabean Oil	3700	0	100	0	0	0	0.1	0	0	0	0.01	0	0	0	16.06	0	0	0	0	0	0
Ghee	3690	0	99.8	0	0	1	0.2	0	0	2	0.01	0.01	642	1.9	3.31	0	0	0	0	0	0
Green Chili	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Kochu	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Potato	282	1.2	0.2	14.1	2.1	12	0.5	21	39	15	0.75	0.41	2	0	0.02	0.07	0.08	0.7	0.25	15	15
Onion	249	1.4	0.1	12.2	1.9	24	0.9	24	29	11	0.41	0.36	2	0	0.02	0.05	0.14	0.3	0.168	19	4.5
Garlic	623	6.9	0.6	27.6	2.1	25	1.6	25	162	11	1.08	0.18	0	0	0.08	0.13	0.12	2.4	1.235	4	24.1
Puishak	105	2.4	0.3	2.1	2.2	111	2.2	179	31	69	0.35	0.06	170	0	0	0.02	0.36	0.5	0.24	140	51.8
Lalshak	131	4.5	0.3	0.5	4.2	287	5.3	129	34	53	0.85	0.25	842	0	0	0.02	0.1	1.2	0.147	50	19.8
Beef	600	18	9	0	0	10	2.1	20	180	80	4	0.1	200	0.8	0.3	0.06	0.3	10.3	0.32	8	0
Mutton	497	21.4	3.6	0	0	12	2.8	27	193	82	4	0.26	0	0	0.18	0.11	0.49	3.8	0.4	5	0
Chicken	500	19.2	4	0.2	0	17	1	26	190	55	2.09	0.22	23	0.1	0.24	0.12	0.12	9.7	0.4	7	0
Liver																					
Egg	600	15	10	0	0	40	1.7	20	230	125	2.36	0.3	178	2	0.94	0.17	0.4	3.6	0.14	48	0
Milk	263	3.1	3.7	4.3	0	103	0.1	13	90	51	0.45	0.05	32	0	0.08	0.06	0.28	0.8	0.053	9	2
Banana	400	1.3	0.8	19.2	2.6	11	0.3	23	36	10	0.24	0.09	2	0	0.75	0.05	0.08	0.9	0.105	20	1
Lemon	234	0.8	1	10.2	1.3	65	0.3	11	10	2	0.07	0.06	4	0	0.8	0.02	0.03	0.2	0.052	17	45.9
Dates	636	1.2	0.4	33.4	4.1	22	1	12	38	7	0.2	0.12	2	0	0.05	0.06	0.07	1.4	0	25	14
Mola fish	452	17.1	4.4	0	0	767	3.8	30	440	43	3.19	0	2680	0	0	0	0	0	0	0	0
Dhela fish	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Kachki fish	420	10	4.5	2.5	0.4	300	1.8	23	300	150	2.5	0.1	25	1.5	0.65	0.03	0.05	2	0.2	7	3
Salt	0	0	0	0	0	0	0	0	0	39340	0	0	0	0	0	0	0	0	0	0	0
Cariander	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Ginger	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Tea	150	0.5	0.4	7.3	0.2	22	0.3	4	15	8	0.08	0.01	3	0	0	0.01	0.02	0.2	0.005	1	0.2
Badam	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Honey	1390	0.3	0	81.1	0.2	5	0.5	2	9	9	0.49	0.04	0	0	0	0	0.06	0.16	0.16	1	1.4
Weed	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Tobacco	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Table A4. Bagerhat District (under Khulna Division) Nutritional Profile

Area	Nutrition Id	Nutrition Name	Nutrition Profile	Class Label
Bagerhat	1	Energy(KJ)	251112.0723	1
	2	Protein(g)	2382.102974	1,2
	3	Fat(g)	1717.194329	
	4	Carbohydrate(g)	8679.832661	
	5	Fibre(g)	613.5312953	
	6	Ca(mg)	7236.709953	2,6
	7	Fe(mg)	316.4579788	3
	8	Mg(mg)	9452.903071	4
	9	P(mg)	39364.67621	7
	10	Na(mg)	1269922.347	
	11	Zn(mg)	401.7432875	8
	12	Cu(mg)	71.19423329	
	13	Vit-A(mcg)	18642.62272	
	14	Vit-D(mg)	179.65838	4,6,7
	15	Vit-E(mg)	327.4207351	
	16	Thiamin Vit-B1(mg)	85.357666	
	17	Riboflavin Vit-B2(mg)	53.93744212	
	18	Niacin Vitamin B3(mg)	659.5383882	
	19	Vitamin B6 Pyridoxine(mg)	51.91737474	
	20	Vit-C(mg)	627.5858165	5

Copyrights

Copyright for this article is retained by the author(s), with first publication rights granted to the journal.

This is an open-access article distributed under the terms and conditions of the Creative Commons Attribution license (http://creativecommons.org/licenses/by/4.0/).