Nutrient Facts Analysis using Supervised Learning Approaches

J. Aravind, J. Dhalia Sweetlin
Department of Information Technology
Anna University
Chennai, India - 600044

Abstract — A healthy lifestyle in people is achieved by having balanced and nutritional food. In today's world we do not know with absolute certainty what foods can be consumed and what cannot be consumed, that is, we do not know for sure what foods have good nutritional value and what foods do not. The nutritional facts label is printed on food products all over the world and they are represented using a similar structure. These nutritional facts give data about some of the major nutrients present in the food product such as carbohydrates, protein and so on. These nutrition fact labels are not easily understood by common people. People who are careful about their diet such as those who exercise and diet regularly, trainers, and nutritionists may understand these nutritional facts, but not the common people. To make this information accessible in an easier way by classifying these food products into five levels of healthiness ranging from very healthy to very dangerous is the aim of this project work. This is done by a sequential process of data retrieval, data cleaning, data labeling and supervised learning.

Keywords—Nutrition analysis, Supervised classification, Data cleaning, Nutrition facts, Nutrients, Food Classification, Machine Learning.

I. INTRODUCTION

The nutritional facts label also known as nutrition information panel is a label which is present on most food products packaged in most of the countries around the world. These nutritional facts can be found on the back of these packaged food products. For common people who want to know whether the food product is healthy or not, these facts are hard to infer.

In India, on 19th September 2008, the ministry of health and family welfare notified the prevention of food adulteration (5th Amendment). The amendment forced packaged food manufacturers to declare the nutritional information on their product labels and a mark from the F.P.O or Agmark (Companies that are responsible for checking food products) to enable consumers make informed choices while purchasing. Prior to this amendment, disclosure of nutritional information was largely voluntary though many manufacturers tend to adopt the international.

This manuscript provides a system based on nutritional data to classify the food products into five classes such that each class corresponds to a different level of quality of the food with respect to its healthiness. The range of output classes are from 0 till 4, 0 being the healthiest and 4 being the unhealthiest food product. This manuscript uses data retrieval, data cleaning, data labeling, classification of the data using supervised machine learning algorithms and finally visualizing the result. This system aims to provide a machine learning solution to classify products based on how healthy they are.

The rest of the paper is organized as follows: Section II presents the existing works in the domain and section III deals with the proposed system framework. Experimental results are discussed in section IV. Section V presents the conclusion and scope for future work.

II. LITERATURE REVIEW

Some of the articles by eminent scholars on nutritional diet and facts are studied and discussed below:

Baghurst et al. designed a computerized dietary analysis system which can be used with diet diaries and food frequency questionnaires [1]. The results obtained are available to research workers in the area of human nutrition at a running cost and helps them in further analysis.

Cowburn et al. published a research paper which explores the understanding of nutrition fact labels in Europe. The results showed that only 9% of nutrition fact labels were judged to be of high or medium high quality. This supports the premise of this manuscript [2]. The reported use of nutrition labeling was found to be high, but the objective usage of nutrition fact labels while buying food products were found to be low. It is reported that while users understood some aspects of the food labeling, the information in the food labels made them confused. Most users were able to decipher simple information from the food labels but the ability to interpret the nutrition labels decreases as the complexity of the task increases.

A comparison of five health and diet surveys done by Bender et al. provides estimates of numbers of consumers who have paid attention to ingredient lists and nutrition labels and identify the trends based on replicated measures [3]. It reports

978-1-5386-1866-0/17/$31.00 ©2017 IEEE

that more than four out of five consumers pay attention to one or both types of label information. Consumers who read both types of labeling information are likely to be young, better educated, and follow a self-initiated or doctor-prescribed low sodium or low cholesterol diet.

The food labels are viewed as an important source of information to consumers according to Margareta Wandel [4]. The result from the study suggests that many consumers have difficulties when trying to understand this information. The consumers felt that the terminology on the food labels appeared technical and very advanced to them [4]. They have dealt with these problems in various ways. Some made use of the labels to a limited extent whereas some did not even use the labels. Some thought that reading the label consumed too much time or that it was very difficult.

Rules for novel foods and genetically modified foods are analyzed by Cheftel [5]. The nutritional labeling directive is detailed, together with foods for particular nutritional uses and food supplement rulings. Critical surveys on implementation of nutrition and food labeling are summarized, with corresponding recommendations for improvement.

Instead of looking at separate nutrients or foods, pattern analysis examines the effect of overall diet of a consumer. Dietary patterns show a broader picture of nutrient and food consumption, and hence may be useful to predict disease risk [6]. Studies by Frank B. Hu have suggested that dietary patterns from cluster or factor analysis predict disease, risk or mortality.

The association between soft drink consumption and nutrition and health outcomes and health outcomes is analyzed by Vartanian et al [7]. A clear association of soft drink intake with increased energy intake was noted. Soft drink intake was also associated with lower intakes of milk, calcium and other nutrients with an increased risk of several medical problems such as diabetes [8].

Chowdhury et al. summarize evidence about associations between fatty acids and cardiovascular health using observational studies and randomized, controlled trials [9]. Seventy-two unique studies were identified over various continents. There were 45 prospective, observational studies and 27 randomized studies. A clear correlation between fatty acids and cardiovascular health was found out.

The article by Variam et al. provides a model that measures how much factors such as nutrition knowledge and diet awareness influence an individual’s Health Eating Index (HEI) [10]. The results show that one’s level of nutrition information has an influence on one’s HEI, it also shows that higher education leads to healthful choices.

From the above literature survey it is clear that there is a strong correlation between the kind of nutrients which the consumers’ intake and their health. It is also seen that there is a lot of difficulty involved in interpreting nutrient fact labels by the consumers. Dietary patterns are important and influence how healthy a person is and will be. This manuscript is based on the above research papers and implements a system for predicting the healthiness of a particular food item based on the nutrients available in common nutritional labels.

The manuscript follows a sequential process of data collection by retrieving nutrition data using web scraping. The data is collected from the Canadian Government's heath website. The health website contains nutrient data for about 1200 commonly used foods. The system uses these nutrients as features for training and testing using supervised machine learning algorithms. The food products are labeled from 0 to 4. The class label 0 representing the healthiest and 4 representing the unhealthiest food product. To develop a classification system, the class labels of the food products 0 through 4 are assigned under the guidance of nutritionists. A performance analysis of algorithms is done using the accuracy parameter. From the literature survey it is seen that there is no similar analysis of nutrient facts of food using supervised learning techniques and hence this article aims to help consumers choose food products more wisely.

III. SYSTEM FRAMEWORK

The framework of the proposed classification system is presented in Figure 1. The major modules in this framework are data retrieval, data cleaning, class labeling, classification and accuracy prediction.

A. DATA RETRIEVAL

The dataset required to perform training and testing has been retrieved using the process of web scraping. The Canadian health website contains the details of around 1200 food products [11]. Each of these food products has about 25 features which are the various kinds of nutrients present in that food item and its quantity with respective unit.

The data have been split into various types of food products such as meat, biscuits, vitamins and constituents such as sodium, potassium and magnesium. The dataset also included features such as thiamine, niacin, lycopene and essential nutrient details such as carbohydrates, proteins, fats, energy and so forth [12]. Duplicate features such as energy in both Kcal and KJ have been consolidated into just Kcal. Each food product holds a quantity of measure and its weight in grams.

The dataset has been scraped from the website using the Pandas package in python [13]. The scraped data is stored in an appropriate database. The database file contained inconsistent data, as there are missing values. Table 1 shows samples of retrieved data, two of their features and the respective values in the dataset. Table 2 shows some of the inconsistencies present in the dataset which has been retrieved.

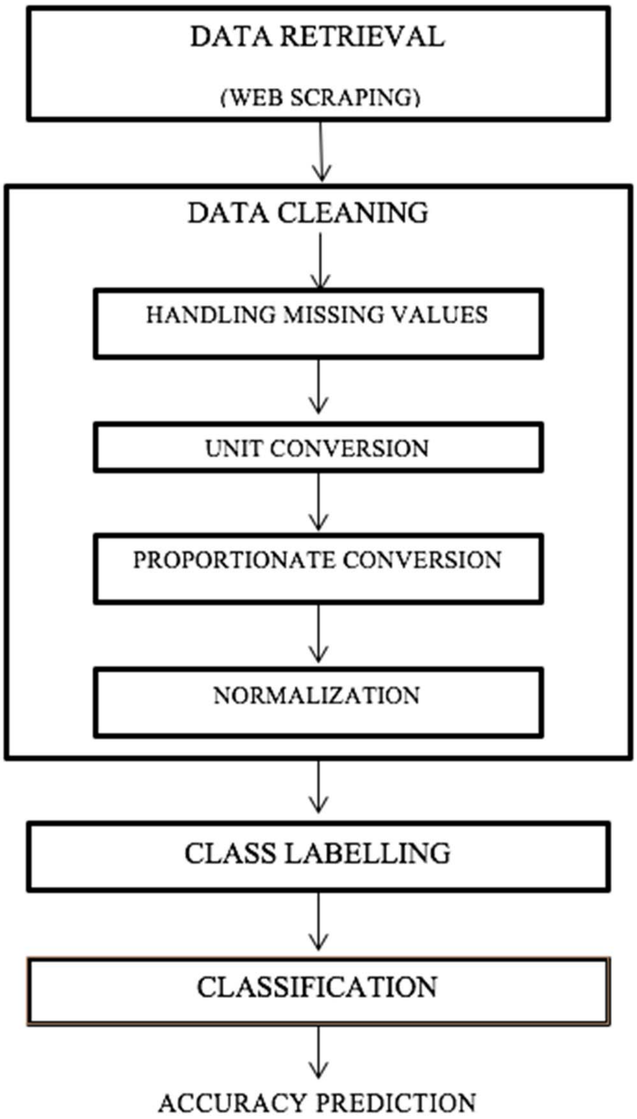

Fig.1. System Framework

TABLE 1. RETRIEVED DATA

Food Product	Weight(g)	Protein(g)
Oat Bran, Dry	50	9
Ketchup	15	0.03
Bread, Italian	35	3

TABLE 2. DATA WITH MISSING VALUE

Food Product	Weight(g)	Total Fat(g)
Bread, French	35	18
Bread, Italian	35	
Bread, Oatmeal	35	17

B. DATA CLEANING

Data cleaning is a pre-processing activity in data-mining which is carried out before training a classifier [14]. Incomplete, noisy and inconsistent data are usually present in real world data. Hence to improve the quality of data and thus the quality of classification, imputation of missing values, unit conversion, proportionate conversion which is specific to this work and normalization activities are carried out in this module.

a) Handling missing values

The retrieved data contains missing values for some features in some records. These missing values can be filled using methods such as replacing with mean, median, constant or with interpolated estimates [14]. In this work, the missing values are replaced using median of the particular feature. The scraped dataset has values for some features as ‘TR’ representing traces in amount. It is replaced with an appropriate value and varied depending upon the type of food product. If more than half of the features of a record are missing due to web scraping, then that record is completely deleted. Table 3 demonstrates handling of missing values.

TABLE 3. HANDLING MISSING VALUES

Food Product	Weight(g)	Total Fat(g)
Bread, French	35	18
Bread, Italian	35	18
Bread, Oatmeal	35	17

b) Unit conversion

The various features (nutrients) in the datasets had different units. Energy had units such as Kcal / KJ, Folate had DFE as its unit and a majority of the features had g, mcg, mg as their units. Some of the features and respective units are shown in table 4.

TABLE 4. NUTRIENTS WITH RESPECTIVE UNITS

Nutrient	Unit
Sodium	Mg
Iron	Mg
Vitamin B12	Mcg
Energy	Kcal
Total Dietary Fiber	G
Protein	G
Trans Fat	G

These different weights of gram were converted into g. This is done in order to ensure uniformity and to make the dataset better for the machine learning algorithms [15].

c) Proportionate conversion

In the retrieved dataset, each food item has a different weight in g. As this will affect the training process, all the features are proportionately converted with respect to weight as 100 g per serving[16]. The proportionate conversion is done by the mathematical formula given in Equation 1. The proportional conversion is done to all the features across the data set.

$$N = \frac{Mn}{m} \tag{1}$$

where n is 100, M is the particular feature value, m is the weight value in g and N is the proportionate feature value for 100g per serving. Table 5 shows the data after performing proportionate conversion.

TABLE 5. PROPORTIONATE CONVERSION

Food Product	Weight(g)	Protein(g)
Oat Bran, Dry	100	18
Ketchup	100	0.20
Bread, Italian	100	8.57

d) Normalization

The data is normalized in the range [0,1]. Normalization is done as it makes computation faster. It also makes each feature contribute equally and proportionately in achieving the result. Normalization of data is done using Equation 2 [17].

$$x' = \left(\frac{x - xmin}{xmax - xmin}\right)\left(x'_{max} - x'_{min}\right) + \quad x'_{min} \tag{2}$$

where x' is the normalized value, x is the current feature value, $xmin$, $xmax$ are the minimum and maximum values in the dataset. The maximum value of the range [0,1] is assigned to x'_{max} and the minimum value to x'_{min}. The data after normalization is shown in table 6.

TABLE 6. DATA AFTER NORMALIZATION

Food Product	Energy(kcal)	Protein(g)
Bread,French	0.2961	0.0093
Bread, Italian	0.2742	0.0124
Bread, Raisin	0.2836	0.0093
Bread, Rye	0.2241	0.0094
Bread, Naan	0.2972	0.0093

C. CLASS LABELING

The records representing the food products are assigned class labels as 0,1,2,3,4 with the help of nutritionists at a hospital at Chennai, Tamilnadu, India who have expertise in this domain. Each record was labeled manually. They were requested to label each food item into one of the 5 classes as shown in table 7. Each class has an associated meaning or description such as healthiest, unhealthiest and so on which makes it easy for the consumer to come to a decision, whether to consume the food product or not to consume that food product [18, 19]. Table 7 gives a clear view of how the classes are assigned to each of the record in the dataset.

TABLE 7. CLASS LABEL DESCRIPTION

Class Label	Description
0	Healthiest
1	Healthy
2	Moderately Healthy
3	Unhealthy in large quantities
4	Unhealthiest

D. CLASSIFICATION

The normalized dataset in the range [0,1] contains 1200 records each having about 26 features (nutrients). The dataset is used to train six classifiers namely Gaussian Naive Bayes, Logistic regression, Linear discriminant analysis (LDA), Classification and regression trees (CART), Support vector machines (SVM) and k Nearest Neighbors (k-NN) in ten-fold cross validation. The trained classifiers can be used to classify new food products.

IV. RESULTS AND DISCUSSION

The pre-processing and classification systems are implemented using python and scikit-learn framework [20, 21]. The performance metric accuracy is computed using Equation 3 for each machine learning algorithm considered and the results obtained are given in Table 8.

$$Accuracy = \left(\frac{TP+TN}{TP+TN+FP+FN}\right) \tag{3}$$

where TP is the number of actual true positives identified by the system, FP is the number of actual negatives labeled as positives, TN is the number of actual true negatives and FN is the number of actual positives missed and labeled by the system as negative. A predictive model has been created for classifying the healthiness of food products using features of these food products. The classification results shown in table 8 determine the machine learning algorithm that is best suited for this application.

TABLE 8. ACCURACY PERFORMANCE METRIC

Algorithm	Accuracy
Logistic Regression	44%
Gaussian Naïve Bayes	50%
LDA	51%
CART	76%
Support Vector Machines	64%
K Nearest Neighbors	73%

The table 9 shows the average standard deviation with respect to each supervised learning algorithm.

TABLE 9. STANDARD DEVIATIONS

Algorithm	Standard Deviation
Logistic Regression	0.2629
Gaussian Naïve Bayes	0.2806
LDA	0.2569
CART	0.2037
Support Vector Machines	0.2598
K Nearest Neighbors	0.2287

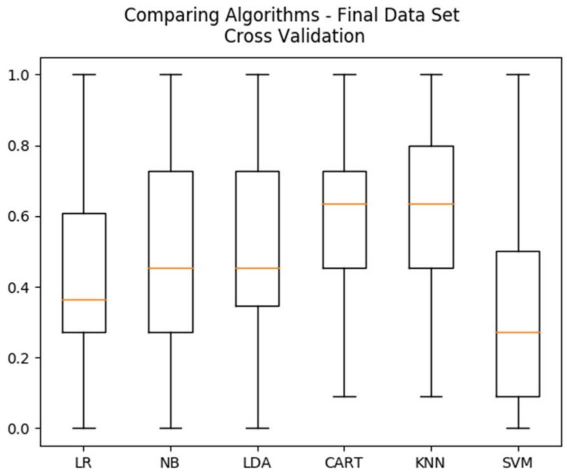

Fig. 2. Box Plot

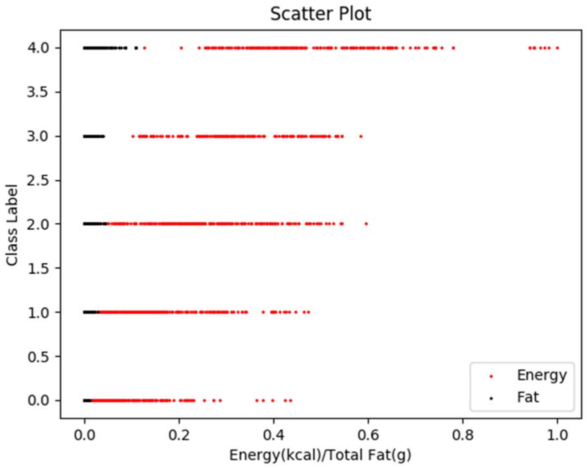

Fig. 3. Scatter plot representing Energy/Total fat

From the results, it can be observed that CART and k-NN algorithms perform well than the other algorithms. The results can be visualized using various data visualization techniques [19] such as box plot in figure 2, scatter plot in figure 3. The accuracy of this model can be improved by adding more records (food products) for training the classifiers.

The confusion matrix in figure 4 provides a clearer understanding of class label prediction in CART Algorithm. It can be inferred that the CART algorithm classifies the food products with maximum accuracy.

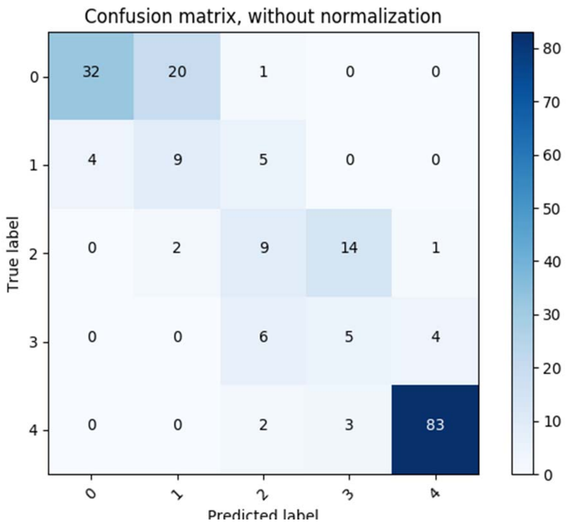

Fig. 4. Confusion Matrix

V. CONCLUSION

Nutrition fact labels have been hard to discern for a long time. This manuscript has provided a basis for classification of food products using the nutrient content present in food products. The contribution of this research paper to existing literature is countable as it has tried to use supervised learning algorithms to interpret nutritional food fact details. The existing literature provided a much needed insight on how dietary analysis helps in predicting diseases and in understanding the fact that nutritional labels are not consumer friendly. This manuscript can be further improved by increasing the quantity and quality of the data set and trying out the other machine learning algorithms in existence. The ability of the computer to make sense of the data is why this paper focuses on machine learning algorithms to solve food classification.

REFERENCES

[1] Katrine I. Baghurst and Sally J. Record, “A computerised dietary analysis system for use with diet diaries or food frequency questionnaires,” in Community Health Studies, vol. 7, pp. 11 - 18, 1984.

[2] Gill Cowburn and Lynn Stockley, “Consumer understanding and use of nutrition labelling: a systematic review,” in Public Health Nutrition, vol. 8, pp. 21-28, 2005.

[3] Mary M. Bender and Brenda M. Derby, “Prevalence of reading nutrition and ingredient on food labels among adult Americans: 1982-1988,” Journal of Nutrition Education and Behavior, vol. 24, pp. 292-297, 1992.

[4] Margareta Wandel, “Food labelling from a consumer perspective,” in British Food Journal, vol. 99, pp. 212-219, 1997.

[5] J. Claude Cheftel, “Food and nutrition labelling in the European union”, Food Chemistry, vol. 93, pp. 531-550, 2005.

[6] Frank B. Hu, “Dietary pattern analysis: a new direction in nutritional epidemiology”, Current Opinion in Lipid ology, vol. 13, pp. 3-9, 2002.

[7] Lenny R. Vartanian, Marlene B. Schwartz, and Kelly D. Brownell, “Effects of soft drink consumption on nutrition and health: a systematic review and meta-analysis”, American Journal of Public Health, Vol. 97, No.4, pp. 667-675, 2007.

[8] M. J. Gibney, H. H. Vorster, and F. J. Kok, “Introduction to human nutrition”, Oxford: Blackwell Publishing, 2nd ed., 2002.

[9] R. Chowdury et al., “Association of dietary, circulating, and supplement fatty acids with coronary risk”, Annals of Internal Medicine, vol. 160, pp. 398-406, 2014.

[10] Jay Variam, James R. Blaylock, and David Smallwood, “USDA’s healthy eating index and nutrition information” in Economic Research Service/USDA, Technical Bulletin No. 1866, pp. 26, 1998.

[11] Canadian Health Website - Data Retrieval: http://www.hc-sc.gc.ca/fn-an/nutrition/fiche-nutri-data/nutrient_value-valeurs_nutritives-tc-tm-eng.php

[12] Canadian website – Nutritional facts: https://www.canada.ca/en/health-canada/services/under standing-food-labels/nutrition-facts-tables.html

[13] Python Toolkit – PANDAS: http://pandas.pydata.org/

[14] P. N Tan, M. Steinbach, and V. Kumar, “Introduction to Data Mining”, Pearson Education India, 2006.

[15] Friedman Jerome, Trevor Hastie, and Robert Tibshirani, “The elements of statistical learning”, Springer: Berlin, Springer series in statistics, 2001.

[16] Proportional Reasoning: http://teachmath.Openschool network.ca/grade-6/proportional-reasoning/

[17] Han J, Kamber M, Data Mining: concepts and techniques, 2nd ed., The Morgan Kauffmann Series, 2006.

[18] Dietary Analysis Reference: https://nutritionfacts.org/

[19] Food Nutrition Labelling: http://www.nutridata.com/

[20] Python Toolkit – MATPLOTLIB: https://matplotlib.org/

[21] Python Toolkit – SCIKIT-LEARN: http://scikit-learn.org/ stable/