A new approach for recommending healthy diet using predictive data mining algorithm

Mr. Vijay Jaiswal – M.Sc. II(Computer Science) – Mithibai Autonomous College
Email-id: - jaiswalvijay2197@gmail.com Contact No:- 9867992177
Dr. Amol Joglekar - Guide

Abstract

In this fast and busy schedule life, people are not giving importance to the quality of food they are eating. They tend to neglect their eating patterns and habits. The fast-food consumption rate is alarmingly high and this consequently has led to the intake of unhealthy food. This leads to various health issues such as obesity, diabetes, an increase in blood pressure etc. Hence it has become very essential for people to have a good balanced nutritional healthy diet. There are many applications which are booming to help people so that they can have control over their diet and hence can reduce weight or they can help them to keep them fit and healthy. The research paper is proposing healthy food habits and eating patterns so that anyone can know the number of calories burned, the intake of macronutrients and so on using on data mining tools. This tool is used for discovering hidden patterns and customer eating habits from different types of data sources. This system will help in tracking and improving the individual's health and the type of food which they can avoid leading towards the risk of illness. A balanced diet means that the intake of each necessary nutrient meets its adequate demand and actual caloric intake balances with calories burned. Additionally, making a diversity of choices from various types of food is also essential to reduce the risk of developing chronic diseases. This diet recommender system focuses on every individual based on their eating habits and body statistics. This research helps in the prediction of a healthy diet for any individual and nutritionist or doctor to design a diet plan as per patient's need.

Keywords: - BMR, Healthy Diet, Recommender System, Harris Benedict equation, Nutrition, Calories, Data mining.

1) INTRODUCTION

1.1 Data mining

Several applications have sprung up for Machine Learning (ML), out of which data mining is the most effective one. Data mining involves intelligent methods, which contributes as an essential step in the process of knowledge discovery in databases. These intelligent methods are applied in order to extract hidden patterns at the intersection of machine learning, statistics, and database systems. Data Mining extracts hidden knowledge from large volumes of raw data. KDD is an iterative process in which Data mining is the analysis step of the "knowledge discovery in databases" process. Six common classes of tasks contribute to data mining which includes Association rule learning, Anomaly detection, Classification, Clustering, Regression, and Summarization. Today the availability of digital data mostly contributes to the use of Data mining to gain knowledge and extract hidden patterns. Data mining systems may integrate techniques from the following spatial data analysis, web technology, information retrieval, pattern recognition, signal processing, computer graphics. The Classification of Data mining system is performed according to the following criteria which include database technology, statistics, machine learning, information science, visualization, other disciplines.

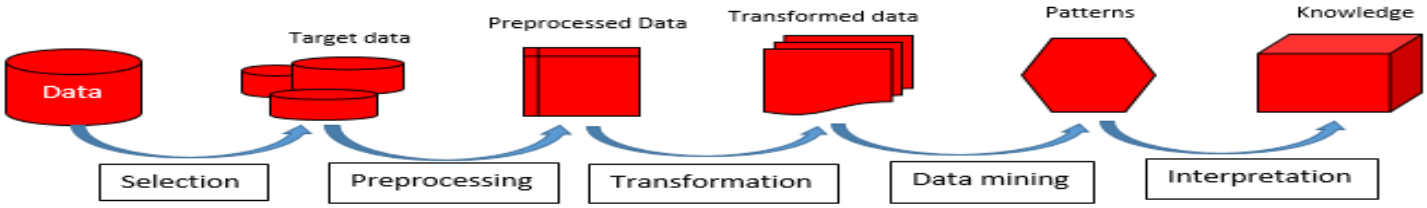

Figure 1: - The KDD Model

The objective of this study is to consider various important aspects of the user's lifestyle and make sure that these factors are incorporated while the system works on a solution to build and recommend a healthy and nutritious diet for the user. A good nutritious healthy diet and a moderate amount of physical activity can help in maintaining a healthy weight. But the benefits of good nutrition have a lot more to do than just managing the weight.

Good nutrition has a direct impact on our health and can help to reduce the risk of some diseases including heart disease, diabetes, stroke, some cancers, and osteoporosis. Combined with physical activity, diet can help you to reach and maintain a healthy weight, reduce the risk of chronic diseases (like heart disease and cancer), and promote overall health. Being fit is all about the 70/30 rule. Here's how it goes, for a person to stay healthy he/she must focus 70% on his dietary intake and 30% on his physical activity/exercise.

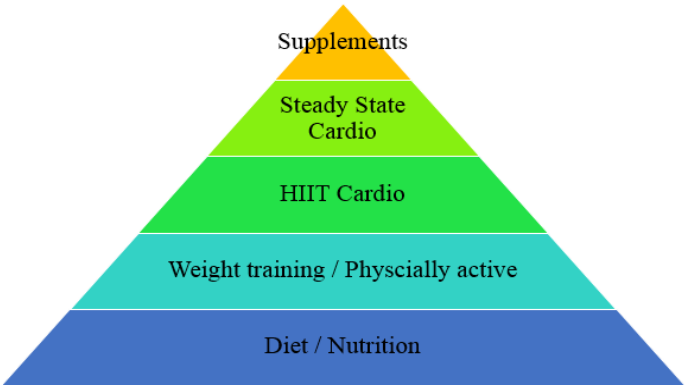

Figure 2: - Hierarchy to stay fit and healthy

1.2 Data mining algorithms

The main task of the research paper is to consume data (height, age, weight, physical activities, fitness goals etc.) from patients and analyze the data to design a healthy diet plan according to individual goals. Data sets are analyzed and the data mining algorithms are applied as the information filtering tools to generate and discover any useful and interesting recommended outputs. We use data mining algorithms like classification, clustering, association rules, and RandomTree decision tree algorithm etc. in the data mining process to extract the useful information of people's eating habit and to recommend healthy diet plan. After comparison of J48, ID3, and RandomTree Decision tree algorithm it was found that RandomTree algorithm worked well for classification technique. First, the nutritive structure of each kind of food is analyzed and how much fat, energy, protein, carbohydrates would be required in an individual's daily diet is calculated. Based on the number of calories and macro nutrition required by each individual the healthy diet report is generated.

Classification is a major technique in data mining that predicts categorical class labels. Accurately predicting the target class for each case in the data is the ultimate goal of classification. Building the Classifier or Model and Using Classifier for Classification are the two steps of the Data Classification process. Classification technique comprises various algorithms. Classification is also sometimes called as Supervised Learning where training dataset is used to learn a classifier. To begin the classification task with a data set the class assignments must be known in advance.

Clustering analysis clusters the data objects that are similar in some sense to one another. Entities in each group are comparatively more similar to entities of that group than those of other groups. The goal of the clustering analysis is to find high-quality clusters such that the inter-cluster similarity is low and the intra-cluster similarity is high. Clustering is also used to segment the data like classification. But unlike classification, clustering is called an unsupervised learning method in which references are drawn from unlabelled, unclassified and uncategorized test dataset.

A rule-based machine learning method for discovering interesting relations between variables in large databases is known as Association rule learning. It comprises of if/then statements that help in analyzing and predicting customer's behaviour. Strong rules discovered in databases are identified using Association rule mining using some measures of interestingness. This rule-based approach also generates new rules as it analyses more data. It is basically a pattern that states when an event occurs, another event occurs with a certain probability. Helping a machine to mimic the human brain's feature extraction and abstract association capabilities from new uncategorized data by assuming a large dataset is the ultimate goal of Association rule mining. An association mining problem can be decomposed into two sub problems.

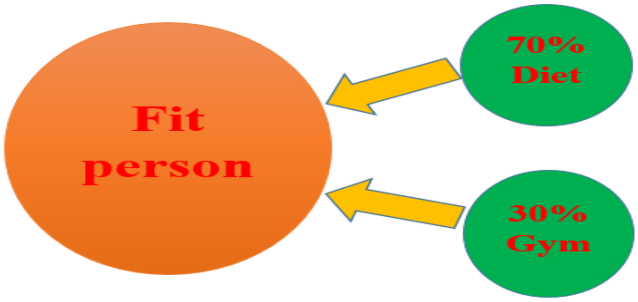

Figure 3: - The Fitness Mantra

2) LITERATURE REVIEW

The authors INGMAR WEBER and PALAKORN ACHANANUPARP [1] made an attempt to gain insights from machine leaned - diet success prediction which would help people trying to stay fit and healthy by keeping a track on their dietary intake. The authors used public food diaries of more than 4,000 long-term active MyFitnessPal users to study the characteristics of an unsuccessful diet. Concretely, authors trained a machine learning model to predict repeatedly being over or under self-set daily calorie goals and then look at which features contribute to the model's prediction, where research was centered around “quantified self " data. The authors observed that classification performance was sufficient and the token-based model performed better than the category-based model and used such data feasibly for more in-depth data mining.

NANDISH SHAH and ISHANI SHAH [2] presented a proposal of healthy food habits and eating system based on web data mining, to discover hidden patterns and business strategies from their customer and web data, which would track eating habits and recommend the types of food that will improve the health and avoid the types of food that raise the risk of illness. The authors used data mining algorithms like classification, clustering, association rules, etc. in the data mining process to extract useful information about people’s eating habit. The nutritive structure of each kind of food was analyzed and the fat, energy, vitamin percentage in the recipe was calculated. Then they used the classification mining algorithm to process the composition data and give out the result whether the diet is healthy or not. As a result, personalized recommendations were suggested for each person.

How a coding system at the meal level might be analyzed by using data mining techniques was demonstrated by the authors AINE P. HEARTY AND MICHAEL J. GIBNEY [3] through this article. They evaluated the usability of supervised data mining methods to predict an aspect of dietary quality based on dietary intake with a food-based coding system and a novel meal-based coding system. The authors used Food consumption databases from the North-South Ireland Food Consumption Survey 1997–1999. A healthy eating index (HEI) score was developed. Quintiles of the HEI based on combinations of foods were predicted by Artificial neural networks (ANNs) and decision trees. As a result, the ANN had a slightly higher accuracy than did the decision tree in relation to its ability to predict HEI. However, on the basis of the meal coding system, the decision tree had higher accuracies than did the ANN.

Data mining was used by CHRISTY SAMUEL RAJU, SANCHIT V CHAVAN, KARAN PITHADIA, SHRADDHA SANKHE, PROF. SACHIN GAVHANE [4] to develop a Fitness Advisor System. “Fitness Advisor” developed by authors was a desktop application that advised the user according to his/her problem associated with body weight by an efficient diagnosis of the same and spreading proper awareness about the health hazards. The authors considered different factors in the system such as height, weight, body type, sex, smoking, drinking, health condition, physical activity, sleeping hours etc. A combination of clustering, association and classification algorithms to effectively deliver the best possible expert advice to the user's problem was used by authors. Apriori algorithm was used by authors for generating association rules. The final output of the system was expert’s advice in terms of diet and exercise.

The authors Lydia MANIKONDA, RAGHVENDRA MALLY, VIKRAM PUDIZ and RAGHUNATHA RAO [5] focused on an application of mining questionnaires to determine the current knowledge of child participants and how this knowledge improves after the training session. The authors divided the paper into nine sections. The major areas on which training was given were 1) Adolescent Phase 2) Breast Feeding 3) Food Groups 4) Anaemia and Folates 5) Family Life Education. Pre-processing of data was applied by authors to avoid any discrepancies. The main algorithms which were used for this purpose were a few basic algorithms which were available in WEKA. Three classification algorithms Naive Bayes, Bayes network, Decision Trees or C4.5 were applied on the dataset. Two clustering algorithms K-Means and DBSCAN were mainly used by authors in the analysis. As a result, it was found that data mining was able to infer the results much better than the results obtained from the traditional statistical analysis.

Performance analysis of a Healthy diet recommendation system using web data mining was performed by SHILPA DHARKAR, ANAND RAJAVAT [6] where the Design and implementation of a healthy diet recommendation system was based on web data mining. On analysis the authors found out that the Healthy Diet Recommendation System consisted of Data collection, Data pre-processing, Information Filtering, Data Base Design and Implementation with Web-Based User Interface and The User and Recommender system interaction via HTTP. The performance of a healthy diet recommendation system used the ID3 and C4.5 decision tree classification algorithm to classify the healthy diet data set. In the implementation phase, the authors first selected the data set then the generated rule. Then these rules were applied to the healthy diet recommendation data set. As compared to C4.5, ID3 worked with each instance and provided more accurate results after bagging was applied. As a result, it was found that a Performance analysis of a healthy diet recommendation system recommended the food that was beneficial for an individual's health.

SHILPA DHARKAR and ANAND RAJAVAT [7] presented a proposal of a healthy eating system based on web data mining, which would track the eating habits and recommend the foods that improve an individual's health and avoid the types of foods that raise your risk for illness. The authors introduced a web data mining solution to e-commerce to discover hidden patterns and business strategies from their customer and web data, propose a new framework based on data mining technology for building a Web-page recommender system, which would be used as the basic framework for the healthy eating system. Two information filtering methods for providing the recommended information were considered by authors. By using the data mining algorithms, the information filtering processes could be performed prior to the actual recommending process which resulted in improving the system response time and making the framework scalable.

3) PROPOSED METHODOLOGY

3.1 System Methodology

The proposed system of food recommendation for a particular client is based on factors such as food preferences, availability of food, medical information, disease information, caloric information for a food item, personal information, the activity level of each individual, for a given food database. The important task in implementation is to recommend a particular food item from the food database based on certain constraints such as the likeliness of a particular food item, availability of that food, allergy towards a food item, its nutritional contents such as protein, carbohydrates, fats in that food. This recommendation helps to select the food from the database such that the nutrition deficiencies will not occur in the near future and proper diet plan will be given to each individual while fulfilling the daily calorie intake.

The main objective of the presented work is to construct the decision tree until the appropriate classification is reached to select the proper food item based on food availability, Category of user (Fat, healthy, lean etc.), Likeness Factor, user fitness goals, Overall content of Nutrients in that food, decision rules and constraints on it are defined to design a healthy diet plan for each individual.

3.2 System Architecture

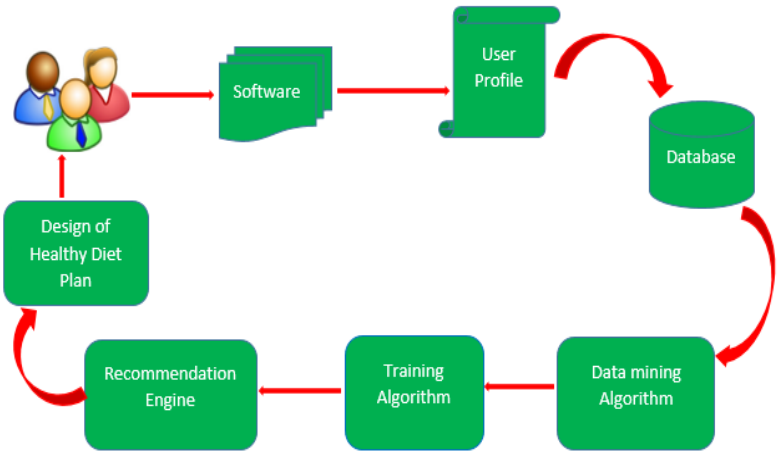

Figure 4: - Diet Recommendation system architecture

The overall process of developing Healthy Diet Recommender System involves mainly seven steps which includes collecting user data, database designing and implementation, data acquisition, pre-processing the data to filter and transform the food datasets based on user profiles, performing classification and clustering, applying the rules on the algorithm based on each individual profile, and recommending a healthy diet plan.

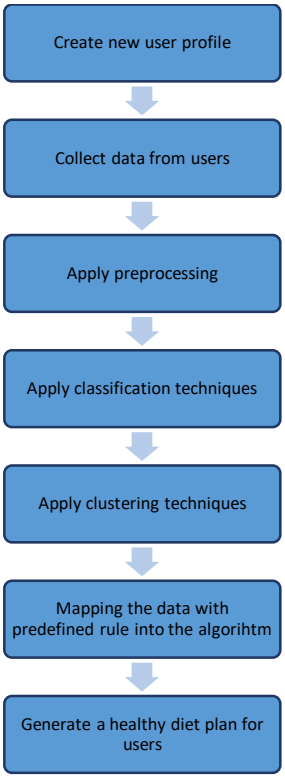

Figure 5: - Process flow of methodology

a) Data Acquisition and Collection

People login into the system and directly enter their health-related information and working lifestyle which is directly stored into the database. This information is acquired to track people's health conditions on regular basis. This health and lifestyle-related information acquired through the software is the first-hand material for this system e.g. the height, weight, diseases, allergy conditions, physical activity of users, their habits etc. The data regarding food items and its nutritional value is collected from the dataset provided by USDA agricultural research service food datasets. Generally, the data acquisition module selectively obtains data from the outside web environment, in terms of functionality to provide material and resources for the latter data mining.

b) Data Pre-processing

Data pre-processing mainly processes and reconstructs the source data acquired in the data acquisition phase and builds the data warehouse of related themes to create a basic platform for the data mining process. It is an important data mining technique used to transform the raw data (inconsistent, incomplete etc.) into an understandable format. Data pre-processing is preparation for data mining and it mainly resolves issues of raw data by methods such as data scrubbing, data integration, data conversion, data reduction etc.

c) Information Filtering

The information filtering step is the core process of the recommender system framework to remove unwanted or redundant data, where the data sets are analyzed and the data mining algorithms are applied as the information filtering tools using (semi)automated or computerized methods to generate and discover any useful and interesting recommended outputs. We use data mining algorithms like classification techniques, clustering etc. in the data mining process to extract useful information about people's eating habits and to predict the food items to be suggested based on these habits and their health status. First, we analyze the nutritive structure of each kind of food and calculate how much fats, energy, proteins, carbohydrates an individual would require in their daily diet. Then we apply the predefined rules into the algorithms to process the composition data and give out the healthy diet reports.

d) Recommending Healthy Diet

After the information mining method in the last step, we tend to might get a piece of abundant helpful information. Then we could advocate a healthy diet arranged consistent with the private condition of the users. The recommendations by our recommender system would improve your organic process structure and lift your health standards. On the opposite hand, we tend to conjointly track user's individual preferences. This method might advocate the related diet attempt to fulfill the personalized needs by the exploitation of association rule mining. So it'd give higher service and knowledge to users.

3.3 Implementation

Decision tree learning is a decision support tool commonly used in data mining that uses a tree-like model for decision making and for calculating the target value having a distinct function. It uses a branching method to display

every possible outcome of a decision. The decision tree symbolizes the function that has been learned. It is a simple representation for classifying examples. The decision tree is a widely used non-parametric effective machine learning technique to build classification or regression models in the form of a tree-like structure.
The following three reasons for which the decision tree learning algorithms are widely used are:
1. The decision tree can widely be used to conclude results from the unobserved instances of particular cases.
2. In these methods, efficient calculations are performed that are proportional to the observed instances.
3. Ultimately after doing all calculations, the produced decision tree is easily understood by the human.
The proposed system demonstrates the use of the RandomTree algorithm for decision tree learning for making decisions such as which proper food item should be assigned while planning the menu. To make this decision, the training dataset is provided to classify the decision tree. Training data helps in calculation of Entropy and Information gain factors. Negative as well as positive outcomes are derived to a certain condition. The proposed system uses the decision tree to determine whether a particular food item should be given to an individual or not by considering the factors such as the Category of user fitness goals, the Likeness Factor, Allergic towards food items or not. Based on the positive or negative outcome, one can suggest the foods to be assigned to an individual efficiently. The RandomTree algorithm is used to take proper decisions among available foods. User choices can be taken into consideration when designing a proper healthy diet plan. Larger training set leads to more accurate outcomes. Training set increases as and when the number of possible cases in attributes increases. Here, the training data set is taken into account as all the attainable cases of every user so that each time accurate results are obtained.
A proper set of attributes and an output decision is to be provided to train RandomTree. For calculations, the assumed sample dataset is as shown below in Table 1. First, create a training data set using the following attributes shown in table 1. It classifies database depending on below attributes.
(i) Food item {milk}
(ii) Likeness {high, medium, low}
(iii) Category {‘maintain’, ‘lose weight’, ‘gain weight’}
(iv) Allergic/not recommended {no, yes}
(v) The outcome class is: decision {yes, no}

Table 1: - Sample Attributes for training algorithm

food item	likeness	category	allergic/not recommended	decision
milk	high	maintain	no	yes
milk	medium	maintain	no	yes
milk	low	maintain	no	no
milk	high	lose weight	no	yes
milk	medium	lose weight	no	no
milk	low	lose weight	no	no
milk	high	gain weight	no	yes
milk	medium	gain weight	no	yes
milk	low	gain weight	no	no
milk	high	maintain	yes	no
milk	medium	maintain	yes	no
milk	low	maintain	yes	no
milk	high	lose weight	yes	no
milk	medium	lose weight	yes	no
milk	low	lose weight	yes	no
milk	high	gain weight	yes	no
milk	medium	gain weight	yes	no
milk	low	gain weight	yes	no

3.4 Proposed Mathematical Constraints and Algorithms

The Harris-Benedict equation or the Harris-Benedict principle is used to calculate an individual's basal metabolic rate (BMR). The total daily energy expenditure (calories)is determined by the Harris Benedict Equation. Calories are burned even when the person is idle or not doing any work, Basal Metabolic Rate (BMR) calculates the number of calories a person would burn if they were asleep all day. Variables such as height, weight, age, and gender are used in BMR formula to calculate the Basal Metabolic Rate (BMR). Harris Benedict Equation is then used to determine a person's total daily energy expenditure (calories). An activity factor is applied on BMR to calculate the Harris Benedict Equation.
To determine the number of calories to be taken in each meal of the day we derive simple mathematical equations. Consider Calorie intake for a particular individual as ‘CAL’ depending upon individual’s Activity level and BMR. These mathematical equations help to assign the numbers of calories for different timings of the day such as morning breakfast, snacks, lunch, evening snacks, dinner, etc.
Equations are defined as,
Morning Breakfast: a1 + a2 + a3 +...+an1 = CAL/5
Snacks: b1 + b2 + b3 + ...+bn2 = CAL / 10
Lunch: c1 + c2 + c3 + ...+cn3= 4*CAL / 10

Evening snacks: d1 + d2 + d3 + ...+dn4 = CAL/ 20
Dinner: e1 + e2 + e3 + ...+en5= CAL / 4
Where n1, n2, n3, n4, n5 denotes number of food items eaten in different timings such as in morning beverages, breakfast, lunch, evening snacks, dinner, etc. ai, bi, ci, di, ei denotes the calories eaten from different food items 'i' in morning breakfast, snacks, lunch, evening snacks, dinner, etc.
An algorithm is designed to determine which food items must be added in the diet plan so that nutritional value is met according to the user's requirements. An upper bound and a lower bound is decided by default for each macronutrient that must be present in the diet plan. Suppose adding paneer may cause fat to exceed its upper bound that food item must be removed and must be replaced by another food item meeting the requirements. During this looping process, our prototype keeps a record of the number of times a macronutrient causes a food rejection. If a macronutrient causes a food rejection three times we remove the food item from the meal plan with the highest amount of that macronutrient. Considering the example above, once paneer is rejected, our algorithm may add whole fat milk, which may also increase fat above its UB. Assume two more foods are added that cause fat to exceed its UB. Our assumption is that the meal plan is already high in fat, and our algorithm is having trouble adding foods that maintain fat below its UB. Therefore, the algorithm removes the food currently in the meal plan that has the highest fat content. For eg: - adding a food item (milk) meets the requirement of protein in diet plan but exceeds the upper bound for requirement of fat then milk must be replaced by another food item such as soybean which will satisfy the requirement of protein as well as fat within the range of lower as well as upper bound.

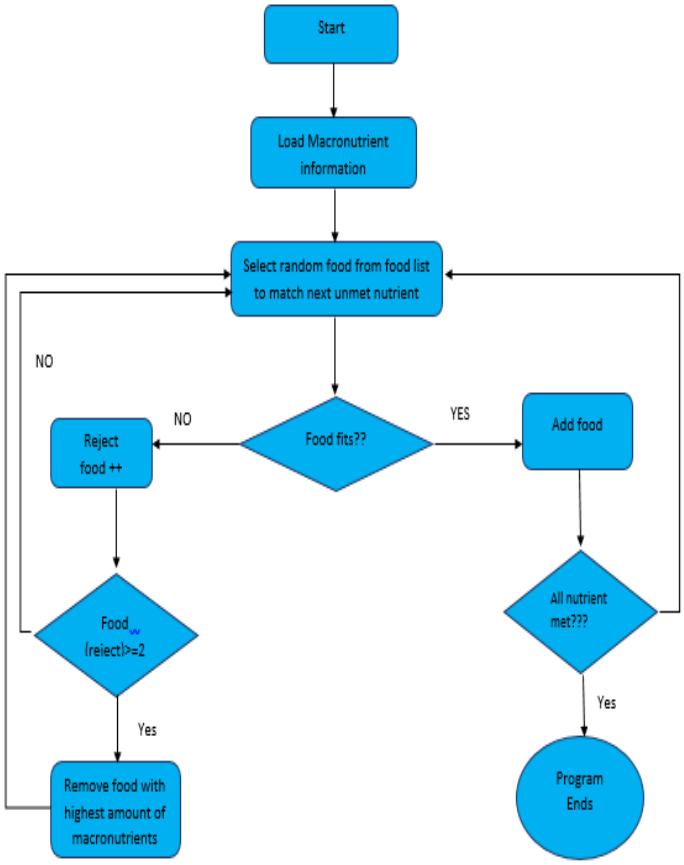

Figure 6: - Graphical representation of algorithm used to generate a food list

3.5 Challenges Faced in Diet Recommendation System

The three main challenges are:
1. Challenges faced with regard to the user.
2. Challenges faced with regard to the algorithm.
3. Challenges to describe relationship between different kinds of food items.

4. RESULTS AND FUTURE SCOPE

To decide whether a particular food item should be given to an individual or not the RandomTree algorithm is used. To use RandomTree, it is important to train it using the training dataset as shown above in Table 1. After implementing the RandomTree algorithm on the training set the decision tree is generated as shown in the below Figure 8. The decision tree describes whether the particular food item must be included in the client's diet plan or not. If a person is allergic to a displayed food item in the system the food is directly removed from his diet plan, else the system further checks the likeness factor of clients towards that food item. If the likeness factor of a client for a

particular food item is high the food will be included in his diet plan regardless of his fitness goal. If the likeness factor is medium for a food item displayed, then the decision, whether that food item should be included in the diet plan or not will be taken by considering the fitness goals of that particular client, else if the likeness factor is low the food item won't be added to the client's healthy diet plan. After the classification technique is used to determine the food items that are suitable for an individual, the predefined rules are applied on the algorithm and a healthy diet plan is generated based on an individual's preferences.

In the future, an algorithm can be generated to suggest the diet plan based on advanced nutrition levels such as sodium content, phosphorous, fibre content, manganese content etc. Along with the food items suggested for each meal the system can also be designed to generate and provide recipes which will include all the food items suggested in the meal plan. More flexibility can be provided to users to add their own food items as per their choice as well as the functionality to add cheat meals into the diet plan can also be introduced.

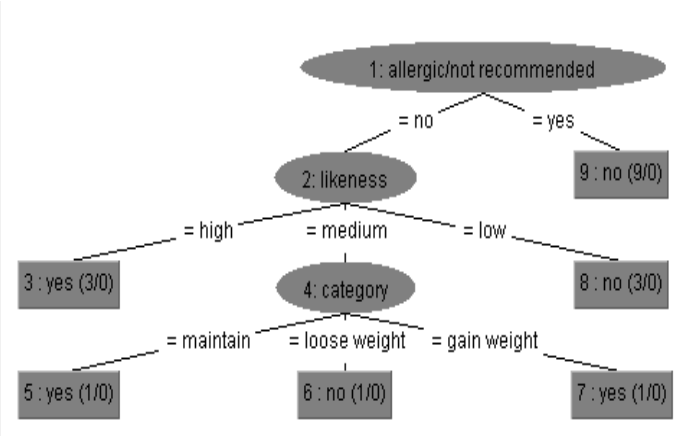

Figure 7: - Generated RandomTree decision tree

CONCLUSION

The importance of nutritional guidance is increasing day by day to lead a healthy and fit life. By accepting the user's food preferences and a user's profile in the system a healthy diet plan can be generated. Typically, balancing the diet and creating a healthy diet plan by tracking the calorie requirements based on an individual's preferences requires expert knowledge on nutrition, as well as it is a time-consuming process.

The studies and their implementation showed that the decision tree learning algorithm, RandomTree works well on any classification problems having a dataset with the non-repeated values. The proper decision regarding which food item should be assigned to a particular individual is done by considering factors such as their likeness for a particular food item, their fitness goals, the content of Nutrients, etc. After that by taking a decision about particular food based on proposed Mathematical Constraints and predefined rules, it is easy to map daily menu planning for an individual.

REFERENCES

1. Ingmar Weber and Palakorn Achananuparp Qatar Computing Research Institute Doha, Qatar, Singapore Management University Singapore, "INSIGHTS FROM MACHINE-LEARNED DIET SUCCESS PREDICTION", Pacific Symposium on Biocomputing 2016, pages 540-551.
2. Nandish Shah, Ishani Shah Department of IT Engineering, Inst. Of Tech, Nirma University, Ahmedabad (India), "DIET RECOMMENDER SYSTEM USING WEB DATA MINING", International Journal of Advanced Technology in Engineering and Science, Volume No.02, Special Issue No. 01, September 2014, ISSN (online): 2348 – 7550, page – 334 to 341.
3. Aine P Hearty and Michael J Gibney American Society for Nutrition," Analysis of meal patterns with the use of supervised data mining techniques—artificial neural networks and decision trees", Am J Clin Nutr 2008;88:1632– 42. Printed in USA. © 2008 American Society for Nutrition.
4. Christy Samuel Raju, Sanchit V Chavan, Karan Pithadia, Shraddha Sankhe, Prof. Sachin Gavhane Information Technology Department, Atharva College of Engineering, Mumbai," Fitness Advisor System Using Data Mining", International Journal of Advanced Research in Computer and Communication Engineering Vol. 5, Issue 4, April 2016, DOI 10.17148/IJARCCE.2016.5451, page-201 to 204.
5. Lydia Manikonda, Raghvendra Mally, Vikram Pudiz and Raghunatha Rao Center of Data Engineering Hyderabad, India National Institute of Nutrition Hyderabad, India, "Mining Nutrition Survey Data".

6. Shilpa Dharkar, Anand Rajavat International Journal of Scientific & Engineering Research Volume 3, "Performance Analysis of Healthy Diet Recommendation System using Web Data Mining", Issue 5, May-2012, ISSN 2229-5518.
7. Shilpa Dharkar, Anand Rajavat, International Journal of Internet Computing, Volume-I, Issue-1, 2011, "WEB DATA MINING FOR DESIGNING OF HEALTHY EATING SYSTEM", Department of Computer Science & Engineering, RGPV University Bhopal, India.
8. Akshay Mahajan, Ankita Dharmale, Ayushi Agarwal, Shriya Pawar, Sneha Sunchu Student, Computer Engineering, PCCOE, Pune, Maharashtra, India," NUTRIEXPERT: A HEALTHY DIET RECOMMENDER SYSTEM", IJARIIE-ISSN(O)-2395-4396, Vol-3 Issue-2 2017.