VOL. 12, NO. 19, OCTOBER 2017 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

PRODUCT RECOMMENDATIONS USING DATA MINING AND MACHINE LEARNING ALGORITHMS

Kaveri Roy, Aditi Choudhary and J. Jayapradha
Department of Computer Science and Engineering, SRM University, Chennai, India
E-Mail: roy.kaveri94@gmail.com

ABSTRACT

Data Mining is a cross-disciplinary field that concentrates on discovering properties of data sets. There are different approaches to discovering properties of data sets and Machine Learning is one of them. Machine Learning is a sub-field of data science that focuses on designing algorithms that can learn from and make predictions on the data. With the increase in the demand for the e-commerce websites, lots of information arises due to which the users face difficulty in finding the relevant information matching their preferences. Thus, we represent a system which will recommend similar food products to the user based on his purchase. The Food Product will be recommended based on the day to day health diseases of the user. The user profile is formed in which health complication of the user is there. The dataset for Recommendation System comprises of 2075 food items. We will apply K-nutrient algorithm to realize the Recommendation System. We will also implement Machine Learning algorithms such as Support Vector Machine (SVM) and Random Forest. In addition to this, the comparison between SVM and Random Forest is performed and SVM outperforms Random Forest algorithm as it shows an increase in the performance.

Keywords: recommendation system, user profile, support vector machine, random forest, collaborative filtering, health hazard.

1. INTRODUCTION

The internet is a powerful tool that has boosted the digital and online applications such as an e-commerce website. The prosperity of online shopping has changed the traditional trading behaviour and thus giving rise to the internet shopping.

There exists a class of Web applications that predict the user preferences. Such a facility is known as a Recommendation System [1]. It includes offering customers an online retailer suggestions about what they might like to buy, based on the product purchased. Our paper builds System for the online food product e-commerce websites. In this paper, we propose the K-nutrient algorithm. The algorithm works on the available nutrient datasets in collaboration with the user specified health issue and recommends the food products. Support Vector Machine (SVM) and Random Forest also have been applied for constructing the Food Product Recommendation System. With the help of these algorithms, the health diseases maps to the ingredient content of the food products worldwide. By analyzing this information, the system will recommend the food products of different varieties based on their health issues. The K-nutrient algorithm when applied gives a satisfactory result to the customer. The Machine learning algorithms such as Support Vector Machine (SVM) and Random Forest is implemented to make the Recommendation System more accurate [2]. The Support Vector Machines is used in applications such as face recognition [3], text classification [4], image classification [5] and much more.

2. RELATED STUDIES

The earlier Recommendation Systems comprised of the Collaborative Filtering (CF) technique. It aims to predict the interests of the users and the recommendation uses a particular query context. The Collaborative Filtering (CF) algorithms find similarities among the items based on the user's feedbacks. Nevertheless, there is a problem related to the Collaborative Filtering (CF) as it suffers from data sparsity problem [6] due to the large availability of items in the system that have compatibility issues with the relatively small data set available. It also has a range of queries that increases the folds of the number of judgments to be made. The questions asked by the new user also depend on the other user's choices, but this may not lead to the accurate result. Our system uses Support Vector Machine (SVM) along with a fuzzy decision support system which is more effectual than the Collaborative Filtering approach. The fuzzy systems deal with imprecise and uncertainty terms based on the fuzziness measurement of set members [7].

3. PROPOSED SYSTEM

The Recommendation System that we propose includes K-nutrient algorithm along with the implementation of Support Vector Machine (SVM) algorithm and Random Forest to increase the efficiency of our Recommendation System. The user will specify his health disease by filling the user profile form in the given e-commerce website. K-nutrient algorithm works on the nutrient database built from the datasets available online [8] and accordingly the recommendation is made on the basis of the food product bought. Support Vector Machines (SVM) is supervised learning models with associated learning algorithms that are for the classification analysis [9]. It works on smaller datasets and eliminates the data sparsity problem .Random Forest is an ensemble learning method for classification that operates by building a multitude of decision trees at training time and outputting the class that is the mode of the classes or means prediction of the individual trees [10]. In the paper, a comparison between Support Vector Machine (SVM) and Random Forest by efficiency, accuracy and time complexity is made. The observation

shows that Support Vector Machine (SVM) outperforms Random Forest. Support Vector Machine (SVM) is much stronger and powerful in building models. According to the results, it is also accurate and less time consuming than Random Forest.

4. IMPLEMENTATION

4.1 WORKING OF THE RECOMMENDATION SYSTEM ARCHITECTURE

The recommendation system architecture is given below:

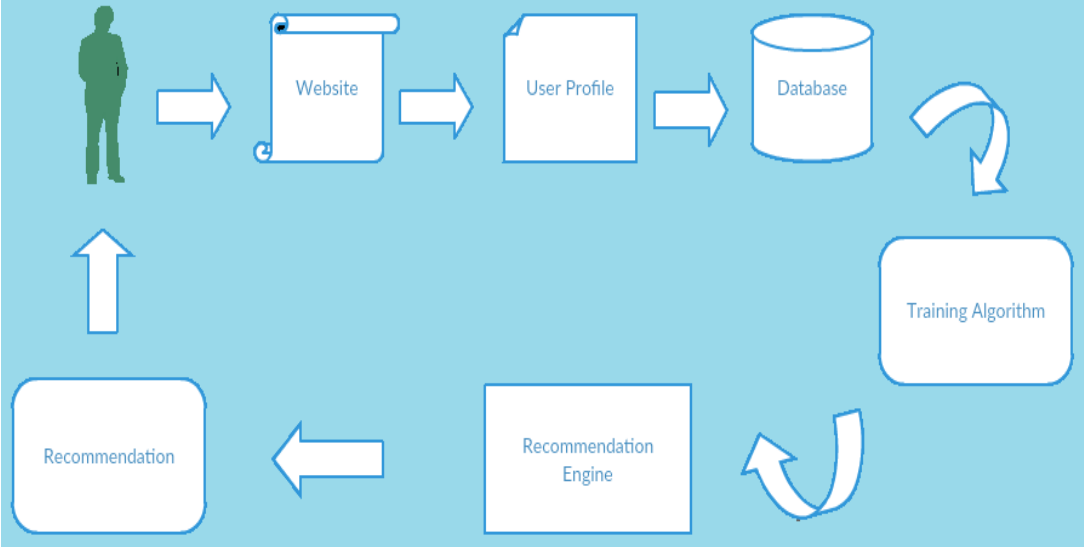

Figure-1. Recommendation system architecture.

In Figure-1, the Recommendation System Architecture obeys the following steps. The user visits the e-commerce website to purchase the food product. If he visits the website for the first time, he has to make the user profile specifying the health issue. The information provided by the user gets stored in the database. The ingredient dataset is also stored in the database. The training algorithm works on the available dataset and is passed to the recommendation engine. The recommendation engine then predicts and recommends the similar varieties of food products purchased by the user based on his health disease.

The training algorithm proposed is the K-nutrient algorithm. Also the system involves usage of Machine Learning algorithms such as Support Vector Machine (SVM) and Random Forest for training algorithm.

4.1.1 K-NUTRIENT ALGORITHM

In the e-commerce website, the user has to create an account before logging into the website. After the user is logged in, he has to fill the form which will, in turn, create the user profile with the unique ID specifying the user's day to day health diseases [11]. The food product dataset [12]is stored in the database. The maximum intake of the nutrients [13] for the particular disease is gathered and stored in the database. The algorithm named K-nutrient is designed to work on the available data as follows:

The user makes the purchase of the food product by specifying the unique ID allotted to him and selecting the food product available on the e-commerce website. The health disease of the user is of major concern while recommending the food product to the user. Two of the specific nutrients whose intake majorly affects the health are taken for the health diseases. The similar food products of different varieties from the datasets in database along with the nutrient values mapped with the user's health is stored in the database. The calculation of the nutrient values is as follows:

Let, the nutrient values of the product bought by the user be i_0 and j_0 for the two nutrients and the nutrient values of the other similar products be i_n and j_n. The calculation of maximum difference p_1 and p_2 of the nutrient values is as follows:

p_1 = $\max(i_0.i_n)$,where n=1,2,.. of the similar food product.

p_2 = $\max(j_0.j_n)$,where n=1,2,.. of the similar food product.

The more the difference, the minimum the value. Accordingly, p_1 and p_2 are mapped with the similar food products and stored in the database with their respective nutrient values to perform the further calculation on them. Now, the final calculation basis is on the nutrient values of the products selected. Let, the nutrient values of the selected product based on p_1 be i_1 and j_1 and forp_2 be i_2 and j_2 respectively.

$$K = [(i_0-i_1)+(j_0-j_1)]/(i_0+j_0)$$

K1 = $[(i_0 - i_2)+(j_0 - j_2)]/(i_0+j_0)$,based on the individual nutrient value

h = max (K,K1)

The best recommendation is based on the already recommended product by mapping the product with the maximum value of 'h.' The time complexity of the K-Nutrient algorithm is O(n) and the execution time is 0.33 seconds approximately.

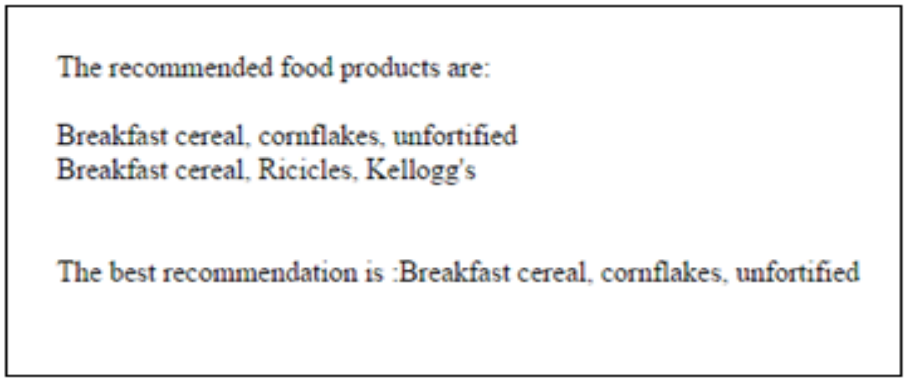

Figure-2. Recommendations with K-nutrient algorithm.

4.1.2 SUPPORT VECTOR MACHINE (SVM)

In our paper, we have applied Support Vector Machine (SVM) algorithm to predict the similar food products based on the items bought and recommended it to the user. Support Vector Machines involves supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis [9]. The classification taken into consideration is "Highly Recommended,""Recommended,""Least Recommended" and "Avoid."The classification applies on the products according to the health-disease of the user. The Support Vector Machine generates the optimal hyperplane that segregates the various products by the classification taking into consideration the health disease. Some of the health diseases include Diabetes, Stroke, Gall Bladder, Arthritis, and Bronchiectasis. The SVM is implemented using R-Studio and R programming.

The training data was provided to the SVM module that is in the ratio of 6:4, and for the 60% of the trained model, the algorithm was written in python to classify the products by fuzzy set approach. Based on the acquired information, the algorithm converts the information into fuzzy variables regarding its membership. For example, let us consider the carbohydrate content of a product is 7.2 g. If we prescript the content >=7.2 g, then it is high. A product with 7.2 g will be low. Such harsh content threshold may seem unfair [14]. So, fuzzy logic is introduced into our system to allow "fuzzy" thresholds or boundaries to be defined. Fuzzy logic uses truth value in a given category. Hence with fuzzy logic, we can capture the notion that the content of 7.15 g is somewhat high, but not as high as 7.2 g. The formula given below involves the fuzzy set membership function for a product recommendation for a person having the health complication of Diabetes. The Figure-3 formula involves the usage of fuzzy logic to segregate the various food items according to the nutrient content. For a Diabetic person, the maximum sugar intake is 7g, and maximum fat content is 4g [13].

recommendation={HighlyRecommended,Recommended,LeastRecommended,Avoid|∀μrecommendation(HighlyRecommended),∀μrecommendation(Recommended),∀μrecommendation(LeastRecommended),∀μrecommendation(Avoid)∈[0,1]}

Degree(Fat)

$$\begin{cases} 1 & 0 < fat < 1.33, HighlyRecommended \\ (2.66 - fat)/(2.66 - 1.33) & 1.33 < fat < 2.66, Recommended \\ 0 & 2.66 < fat < 4, LeastRecommended \\ -1 & fat > 4, Avoid \end{cases}$$

Degree(Sugar)

$$\begin{cases} 1 & 0 < sugar < 2.33, HighlyRecommended \\ (4.66 - sugar)/(4.66 - 2.33) & 2.33 < sugar < 4.66, Recommended \\ 0 & 4.66 < sugar < 7, LeastRecommended \\ -1 & sugar > 7, Avoid \end{cases}$$

Figure-3. Fuzzy logic along with support vector machine (SVM).

The products are classified as soon as the algorithm implemented in Python executes. The formula in Figure-3 depicts a fuzzy variable function which sets the degree to *1* if the margin between the maximum limit and the fat or sugar content is very high and falls under the category "Highly Recommended."The formula then sets the degree to a factor between*0* and *1* that falls under the

category of “Recommended.” The degree is set to 0 if the margin between the maximum limit and fat or sugar is low and the category given is “Least Recommended.”Finally, if the degree is -1 then the value of fat or sugar in the food product is very high and falls under the class “Avoid.”Based on the degrees of fat and sugar and the fuzzy logic we have trained the training set. Then the training set is given to the SVM module to classify the remaining data. As the SVM module makes the prediction, the data is stored in the database. The algorithm devised in Python will work when the user gives a particular product as his preference. It will take the similar products from the database, and according to the classification, it will display the best-suited product for the user. For example, if the user chooses Yogurt then all the Yogurt products will be selected and the product which is under "Highly Recommended" will be displayed to the user. If the products are not under "Highly Recommended" then the next best classification is chosen until "Avoid" is encountered. The database schema in Figure 4 includes the table “FoodItems” that has the food product ID and the food product name. The table “FoodDetails” contains the nutrient value of the food products. The “RequiredNutrients” table consists of the nutrients that will be taken into account while we provide the data in the “Testing_Data” table. From the “SVM_Prediction” table, the user is given the choices of the food products which is beneficial to his/her health and is best suited.

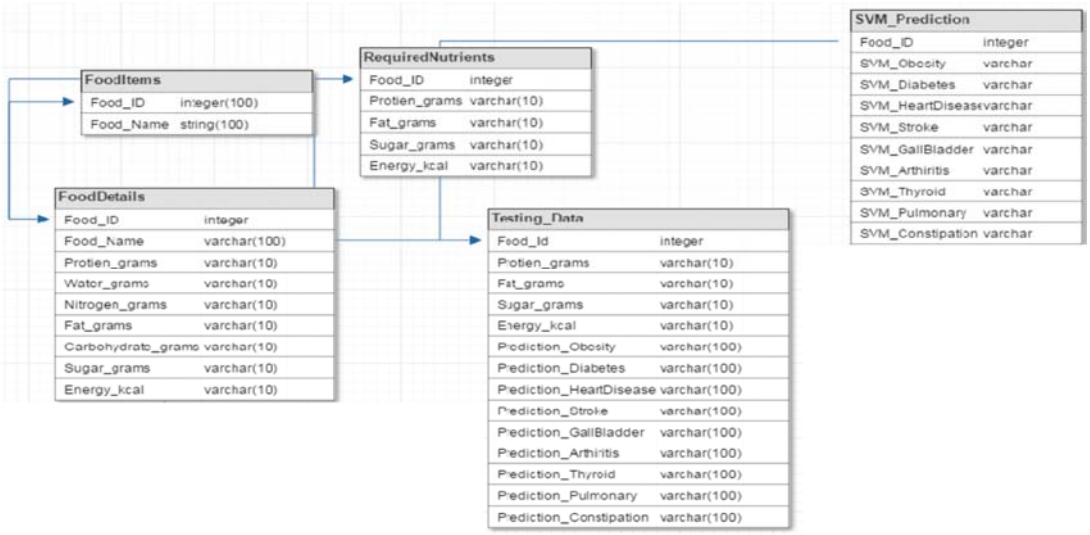

Figure-4. Database design for applying support vector machine (SVM).

Food_ID	SVM_Diabetes	SVM_Stroke	SVM_GallBladder	SVM_Arthiritis	SVM_Bronchiectasis
101	Highly recommended	Highly recommended	Highly recommended	Highly recommended	Highly recommended
102	Highly recommended	Highly recommended	Highly recommended	Highly recommended	Highly recommended
103	Highly recommended	Highly recommended	Highly recommended	Highly recommended	Highly recommended
104	Highly recommended	Highly recommended	Highly recommended	Highly recommended	Highly recommended
105	Highly recommended	Highly recommended	Highly recommended	Highly recommended	Highly recommended
106	Highly recommended	Highly recommended	Highly recommended	Highly recommended	Highly recommended
107	Highly recommended	Highly recommended	Highly recommended	Highly recommended	Highly recommended
108	Recommended	Highly recommended	Highly recommended	Recommended	Highly recommended
109	Highly recommended	Highly recommended	Highly recommended	Highly recommended	Highly recommended
110	Recommended	Highly recommended	Highly recommended	Recommended	Highly recommended
111	Avoid	Recommended	Avoid	Avoid	Avoid
112	Avoid	Avoid	Avoid	Avoid	Avoid
113	Highly recommended	Highly recommended	Highly recommended	Highly recommended	Highly recommended
114	Highly recommended	Highly recommended	Highly recommended	Highly recommended	Highly recommended
115	Avoid	Highly recommended	Avoid	Avoid	Avoid
116	Avoid	Highly recommended	Avoid	Avoid	Avoid
117	Avoid	Highly recommended	Avoid	Avoid	Recommended
118	Highly recommended	Highly recommended	Highly recommended	Highly recommended	Highly recommended

Figure-5. The SVM prediction for diseases on the food product ID.

4.1.2.1 WORKING OF THE RECOMMENDATION SYSTEM ARCHITECTURE USING SVM

In Figure-6, the system architecture describes the whole process of recommending the products. The user is originally the list of products in the database; the list includes a variety of products along with a unique identification of the product. The user then selects the food item that he/she wants to buy according to his/her choice. With the help of the SVM predicted categories for the various diseases that is "Highly Recommended,""Recommended,""Least Recommended" and "Avoid," the user receives output of the recommendation of the best similar product suited to his/her health. The user can look to those recommended products and select the one which will be beneficial to his health.

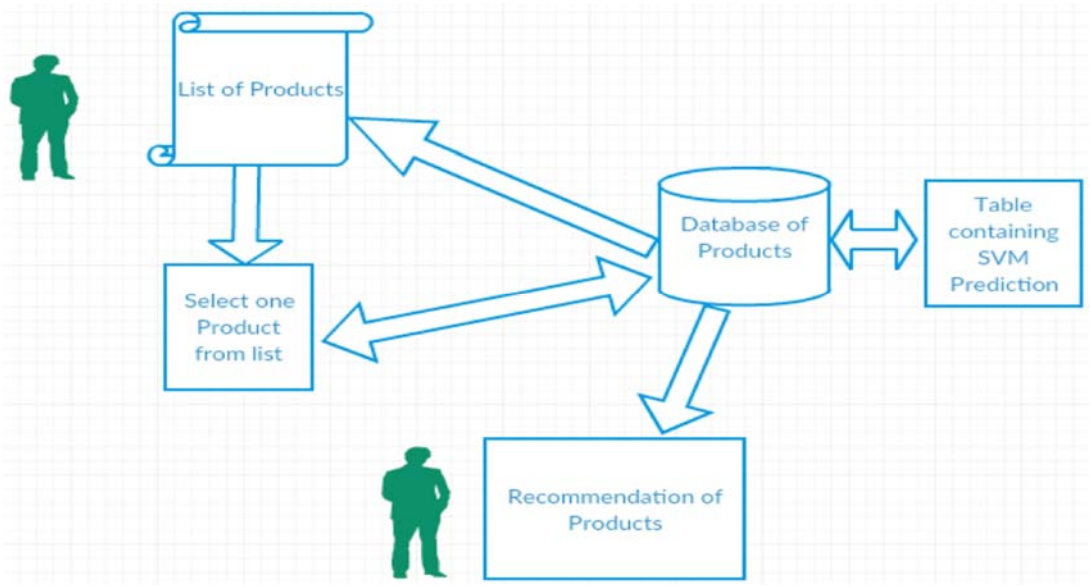

Figure-6. Architecture diagram for food product recommendation using SVM.

4.1.3 RANDOM FOREST

The Random Forest algorithm involves the usage of multitudes of decision trees to classify the data given to it [10]. This takes excess time (approximately twice) to execute than the SVM because multiple trees are created to acquire the conclusion by the training set provided. The fuzzy set approach trains the dataset. Random Forest module in R Programming uses the fuzzy set approach and with the help of it, it classifies the other data under "Highly Recommended,""Recommended,""Least Recommended" and "Avoid."After the classification, in the table, the predicted values are stored. From this table, the values are used for the recommendation of the products to the user. As the user enters the product he/she wants to purchase, a list of similar products is fetched. In our prediction approach, we have chosen two factors to determine the classification for the Bronchiectasis disease that is the amount of sugar content in the food and also the fat content. The Random Forest approach provides the following plot for the same.

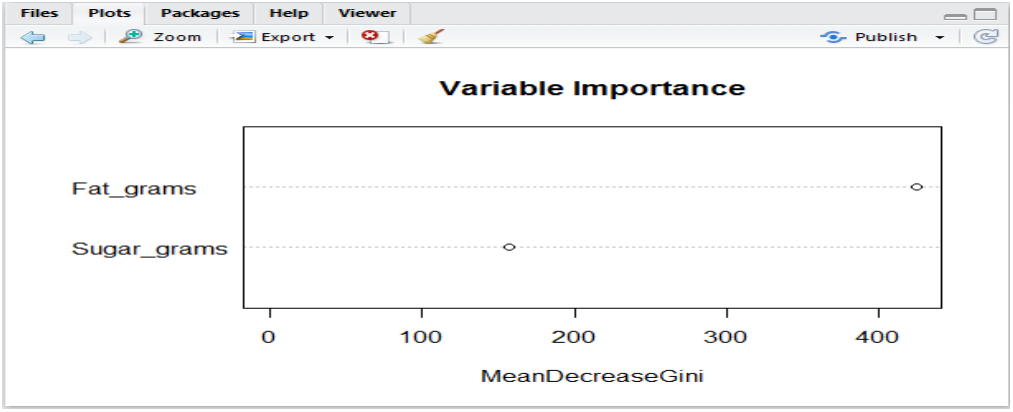

Figure-7. Random Forest plot with classifications for people suffering from Bronchiectasis.

It is comprehensible from Figure-7 that the importance to the fat content is more than the sugar content for prediction. If the products are not classified under "Highly Recommended," then the next best classification is chosen until the "Avoid" category comes.

The database design for it is the same as that for the SVM algorithm, but the only difference is that instead of "SVM_Prediction" table we will be using the table "RandomForest_Prediction" with the same attributes.

Food_ID	RR_Stroke	RR_GallBladder	RR_Bronchiectasis
1	Avoid	Avoid	Avoid
2	Highly recommended	Highly recommended	Highly recommended
3	Recommended	Recommended	Avoid
4	Avoid	Avoid	Avoid
5	Avoid	Avoid	Avoid
6	Avoid	Avoid	Avoid
7	Recommended	Recommended	Avoid
8	Recommended	Recommended	Avoid
9	Recommended	Recommended	Recommended
10	Recommended	Recommended	Recommended
11	Recommended	Recommended	Recommended
12	Avoid	Avoid	Avoid
13	Recommended	Recommended	Recommended
14	Avoid	Avoid	Avoid
15	Recommended	Recommended	Avoid
16	Recommended	Recommended	Avoid
17	Avoid	Avoid	Avoid

Figure-8. The Random Forest prediction for diseases on the food product ID.

5. COMPARISON BETWEEN SUPPORT VECTORS MACHINE (SVM) AND RANDOM FOREST

The SVM approach is better as compared to Random Forest due to the accuracy factor that is calculated by applying the fuzzy set variable logic on the different testing model and comparing the results with the trained model. The number of food products in a different testing model is 875 and in trained model is 1200 food products. The total records sums up to 2075 food products.

The time taken for the Random Forest algorithm to execute was higher than the SVM approach [15]. The time complexity of SVM approach is O(max(n,d) min (n,d)^2), where n is the number of points and d is the number of dimensions [16], and that of the Random Forest is O(v * n log(n)), where n is the number of records and v is the number of variables/attributes [17]. The accuracy factor comes by the predicted data and the trained data.

Food_ID	Protien_grams	Fat_grams	Sugar_grams	Energy_kcal	Prediction_Diabetes	Prediction_Stroke	Prediction_GallBladder
1	2.9	15.2	0.8	151	Avoid	Avoid	Avoid
2	4.0	0.7	0.3	24	Highly recommended	Highly recommended	Highly recommended
3	25.2	10.0	0.0	191	Avoid	Recommended	Avoid
4	0.5	0.6	57.6	223	Avoid	Avoid	Avoid
5	0.3	0.2	20.2	79	Avoid	Avoid	Avoid
6	0.2	0.4	17.1	69	Avoid	Avoid	Avoid
7	0.2	0.3	14.3	57	Avoid	Recommended	Avoid
8	0.2	0.4	11.4	47	Avoid	Recommended	Avoid
9	0.2	0.3	9.6	40	Avoid	Recommended	Recommended
10	0.3	0.3	8.9	37	Avoid	Recommended	Recommended
11	0.2	0.2	6.4	27	Recommended	Recommended	Recommended
12	0.2	0.3	20.8	81	Avoid	Avoid	Avoid
13	0.2	0.3	9.7	40	Avoid	Recommended	Recommended

Figure-9. The table in the database for achieving the prediction result set for the diseases.

The trained data is in Figure-9, and the data predicted by the SVM function in R Programming is in the other tables (Figure-5 and Figure-8). In the Figure-10 the accuracy values are depicted on the basis of the untrained

data and the trained data. It is clear from the Figure-10 that the accuracy percentage for Diabetes, Stroke, Gallbladder disease and Bronchiectasis is 88.45%, 89.37%, 85.71% and 81.74% respectively via SVM. Whereas with Random Forest the accuracy for the diseases that are Stroke, Gallbladder disease and Bronchiectasis is 74.97%, 50.51% and 81.71% respectively. The overall correctness provided by SVM is 88.77% whereas by Random Forest is 69.07% for prediction of complete untrained and different model. Time taken for execution of the SVM approach model is 1.38 seconds whereas for Random Forest approach it is 2.388 seconds. The conclusion achieved by this research is that for prediction over testing dataset the Random Forest fails drastically over SVM. So the SVM module can be used to provide the recommendation to the user with greater accuracy.

```
=== RESTART: C:\Users\kaveri\Desktop\major project\8th sem\svm\Accuracy.py ===
Accuracy for Diabetes via SVM  0.884571428571
Total Correct Data774
Total Data 875


Accuracy for Stroke via SVM  0.893714285714
Total Correct Data782
Total Data 875


Accuracy for Stroke via RR 0.749714285714
Total Correct Data656
Total Data 875


Accuracy for gallbladder via SVM  0.857142857143
Total Correct Data750
Total Data 875


Accuracy for gallbladder via RR 0.505142857143
Total Correct Data442
Total Data 875


Accuracy for Bronchiectasis via SVM  0.918857142857
Total Correct Data804
Total Data 875


Accuracy for Bronchiectasis via RR 0.817142857143
Total Correct Data715
Total Data 875
```

Figure-10. Accuracy values.

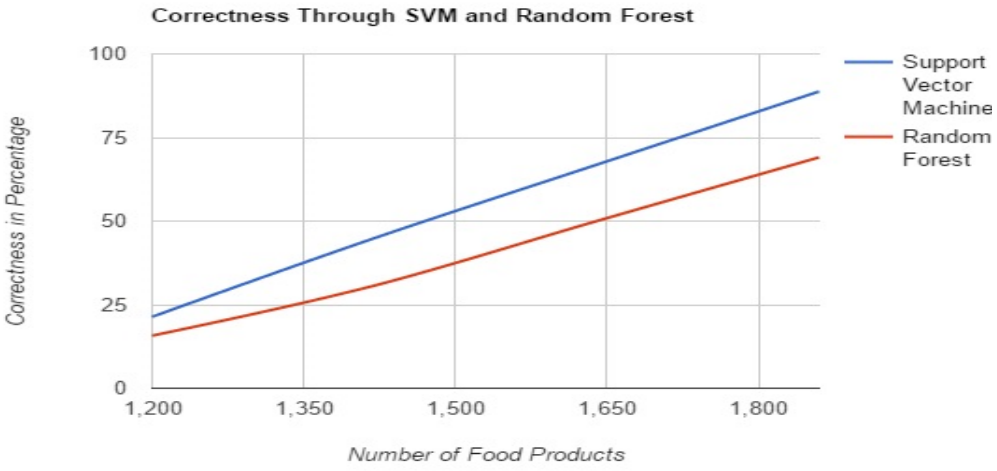

Figure-11. Comparison between SVM and random forest with respect to 'correctness'.

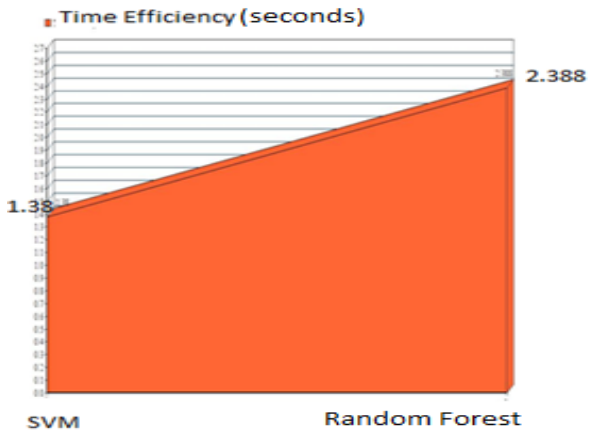

Figure-12. Comparison between SVM and random forest with respect to 'time efficiency'.

6. CONCLUSION AND FUTURE ENHANCEMENTS

The earlier system mostly used the Collaborative Filtering (CF) algorithm, but the algorithm suffers from data sparsity problem. It also has a range of queries that increases the folds of the number of judgments to be made. The system may not lead to the accurate result as the questions asked by the new user also depend on the other users' choices.

In this paper, we have proposed an algorithm named K-nutrient that works well and provides relevant recommendations to the user. We have measured the performance of our Recommendation System on the Support Vector Machine (SVM) and Random Forest algorithms. After the comparison, we see that Support Vector Machine (SVM) outperforms Random Forest on efficiency, accuracy and time complexity.

In our future enhancements, we aim to examine the effect of the size of the training data on the performance of different algorithms applied. We will be comparing the K-nutrient algorithm with the Support Vector Machine (SVM) and Random Forest further.

REFERENCES

[1] http://infolab.stanford.edu/~ullman/mmds/ch9.pdf.

[2] http://sci2s.ugr.es/keel/pdf/specific/congreso/xia_dong_06.pdf.

[3] P. Jonathon Phillips: 1998. Support Vector Machines Applied to Face Recognition. NIPS 1998: 803-809.

[4] T. Joachims. 1998. Text categorization with support vector machines. In European Conference on Machine Learning (ECML).

[5] Olivier Teytaud, David Sarrut: Kernel- Based Image Classification. ICANN 2001: 369-375.

[6] Ko-Jen Hsiao, Alex Kulesza, and Alfred O. Hero. 2014. Social Collaborative. IEEE Journal of Selected Topics in Signal Processing. 8(Disease: 4).

[7] D. Dubois and H.F'rade. 1990. Fuzzy Sets and System; Theory and Applications. New York: Academic.

[8] http://nutritiondata.self.com.

[9] https://en.wikipedia.org/wiki/Support_vector_machine.

[10] https://en.wikipedia.org/wiki/Random_forest.

[11] http://www.cmaj.ca/content/suppl/2013/02/19/cmaj.121349.DC1/physical-payne-1-at.pdf.

[12] http://nutritiondata.self.com/.

[13] Heart Disease-http://pubs.ext.vt.edu/348/348-898/348-898.html Obesity - http://www.maso.org.my/spom/chap9.pdf Diabetes-http://tna.europarchive.org/20120419000433/http://www.food.gov.uk/multimedia/pdfs/publication/whichcard0908.pdf.

http://health-diet.us/diabetes/ Stroke - https://www.stroke.org.uk/sites/default/files/healthy_eating_and_stroke.pdf Gallbladder-https://patient.info/in/health/gallstones-diet-sheet http://www.livestrong.com/article/374074-recommended-diet-after-gallbladder-removal/ Arthritis- http://dietgrail.com/arthritisdiet/ Constipation -http://health-diet.us/constipationdiet/.

[14] Hong-Wei Yang, Zhi-Geng Pan and Bing-Xu, Ming-Min Zhang. 2004. Machine Learning-Based Intelligent Recommendation in Virtual Mall. Proceedings of the Third International Conference on Machine Learning and Cybernetics, Shanghai.

[15] Keita Tsuji, Fuyuki Yoshikane, and Sho Sato. 2014. Book Recommendation Using Machine Learning Methods Based on Library Loan Records and Bibliographic Information. Advanced Applied Informatics (IIAIAAI), 2014 IIAI 3rd International Conference.

[16] https://www.quora.com/What-is-the-computational-complexity-of-an-SVM.

[17] https://www.quora.com/What-is- the-time-complexity-of- Random-Forest-both-building-the-model- and-classification.