# A Comparative Study on Mining the Healthy Food Preferences of Women Clusters

K. Vijayalakshmi, Dr. M. Vinayakamurthy, Dr. V. Anuradha

**Abstract** — A lot of studies have been conducted on behavior of customers generally, but very fewer studies have been conducted specifically on women customers' behavior on healthy food products. In the modern days, where the life is at fast pace, time has become very valuable to every woman and hence importance of spend ing quality time on food choices becomes very critical. At the same time, it is highly crucial to manage the healthy food habits, as preferences and change in lifestyle always play a vital role. The paper highlights the significance in building a healthy society through a survey on socioeconomic factors on food behavior of women who are in domestic, profession or in education using data mining techniques. The work in the paper is carried as a kind of cluster analysis on women dataset using WEKA tool. It consists of a comparative study on the results of various clustering techniques derived using WEKA and determines the influencing factors with various levels of awareness on healthy food products for women clusters. The survey is been conducted to different age groups of women who lives in the gated community of Bangalore, Karnataka state in India.

**Index Terms** — Data Mining, Cluster Analysis, WEKA, Consumer Behavior, Preferences

IJSER

## 1 Introduction

According to the report, the largest retail consumption in India is food category in account to 33 per cent of overall consumption in expenditure. From the IMAGES Research findings, the Indian food market is the sixth-largest food market in the world which is valued at INR 25,13,00 crore and is expected to cross INR 61,00,000 crore by 2020[3].

Data mining is a practice that examines the huge set of existing databases to generate patterns or models that helps the human for decision making. It finds a useful application in CRM where large amount of customer data are dealt. Weka tool is a machine learning tool which provides the solution for all data mining techniques like clustering, classification, preprocessing, visualization etc. The proposed work deals mainly with the clustering techniques as it has been applied in promotion strategies. Clustering is the process of assembling the data objects or records with similar characteristics. This paperwork is focus on analyzing the clustering algorithms: K-Means, Hierarchical clustering and Density based clustering algorithm and find out the most influencing attributes for the healthy food products. Due to ease in implementation and time efficiency, the K-means one of a partitional clustering algorithm is widely used in most of the data mining applications. In the hierarchical clustering, different levels of clusters is built as a tree of clusters which otherwise referred as a dendrogram. In density –based clustering, on the basis of density of data points in a region, it tries to find the clusters.

The only toolkit in data mining and machine learning that has widespread in adoption and withstanding for long period of time is WEKA, a free and open source software that uses GNU General Public License(GPL). It's workbench consists of tools for visualization and algorithms to analyze the data to predict the model. It handles two file formats of type ARFF(attribute relation file format) and CSV(comma separated values).

This study uses cluster analysis techniques to explore the preferences of women consumers who responded to a survey to determine the various dimensions on their purchasing decisions. The term food consumption creates an impact to the environment with different measures. The survey provided with both qualitative and quantitative attributes. The qualitative attributes are represented in terms of preferences, habits, fresh products and perceptions about food consumption whereas the quantitative attributes on amount of food purchased and spent.

## 2 Literature Review

From the study on working women, the results show that they have a greater dislike towards food shopping and cooking. When compared with non-working women, the study further exhibited a tendency is decreasing in concern with the impact of their food shopping and preparation activities on other family members [1]. Over the last two decades, the global increase in obesity associated with childhood, adolescent and adult obesity is due to increase intake of ready-to-eat and snacking [4].

According to World Health Organization's report, 67% of all deaths in India by 2030, will be due to people suffering from non-communicable diseases like diabetes and cardio-vascular diseases. Data mining has been considered as one of the best supporting tools for studying customer preference and behavior, due to its excellent ability in extracting and identifying

- *K. Vijayalakshmi is currently pursuing Ph.D in Computer Science in REVA University,Bangalore, INDIA, E-mail: kvijayalakshmi@reva.edu.in*
- *Dr. M. Vinayakamurhty is currently working as professor in REVA University,Bangalore, INDIA.*
- *Dr. Anuradha is currently working as HOD- PG dept in STC, Pollachi, Tamilnadu*

useful information from large customer databases [5].

The key to effective marketing planning is to understand the customers' preference and behavior which is always crucial for the success of businesses [6]. The analysis of customer preference is necessary for an effective CRM [7]. However, customers' purchase behavior can be influenced by both external environmental factors [8] [9] and internal mental status [10] causing difficulty in evaluation. Estimation of the degree of loyalty of a customer is needed for the better business prospectus [13]. A probabilistic co-clustering approach to pattern discovery in preference data as an extension of the block mixture model can be used both for rating prediction and pattern discovery tasks[11].

## 3 Objectives of the Study

The scope and importance of this survey on women dataset using Weka tool is carried to analyze their food consumption preferences on various attributes like spending pattern, purchasing frequency, convenience, shopping interest, awareness level on various food products. This study will provide the details to :

- Study the food consumption behaviour women based on their categories as a student, working women and housewife through the survey.
- Compare the performance of clustering algorithms using WEKA tool.
- Measure the clusters with different levels of awareness on healthy food products.
- to identify the factors that influence women consumer's awareness on healthy food products.

## 4 Health factors in Food Products

Today the food products in Indian food market are available under different categories as Ready-To-Eat(RTE), Ready-to-Cook(RTC) and Ready-to-Mix(RTM).

- Ready-to-Eat: As per the survey, homemakers are not completely satisfied in consuming the RTE products due to the following reasons: a) preservatives, b) less in nutrition when compared to fresh food, c) mostly a routine food without any novelty.
- Ready-to-Cook: With a growing population of men and women who are always busy to prove them at the workplace finds managing their time became very crucial. Finally the term convenience of every working member is been driven by the RTC industry in India. According to the report, the RTC's market size is 600 – 700 crore out of the total market size of processed food is Rs 1500 crore. Over the next five years, it is expected that RTC industry to grow around 20-25 percent.
- Ready-to-Mix: According to India's RTM Market Outlook, 2021, the overall market for ready-to-mix is growing with a CAGR of 13.22% from last five years. The market is divided into four segments: snack mix, curry mix, dessert mix and others (rice & meal). By the end of 2020, snacks mix is expected to have the highest CAGR of 15.9% followed with 15.8% of curry mix based on application type.

According to World Health Organization, RTC processed food especially meat causes cancer. The usual recommendation for lower acidic canned foods like vegetables and meats is 2 to 4 years whereas higher acidic canned foods like fruits, fruit juices and tomatoes is 1 to 2 years. Refrigeration is very important and must for processed food. If the refrigerator is not maintained properly, then bacteria and fungi may grow inside it and hence more risk factors are associated based on health. Those risk factors are:

- High in calories, fat, sodium, refined carbohydrates and preservatives
- Loaded with sugar
- Contains artificial ingredients
- low in nutrients and Fiber

In one of the websites that provides health tips reported as, cancer is not a disease, but it is due to the deficiency of B17 which can be cured through food without taking medicines and painful radiotherapy treatment. So healthy food plays a very important role to build our nation strong. But India stands a last in the consumption of calorie and protein intake when compare to other countries. To overcome this, awareness on healthy food products is essential to everyone, especially women who takes care of the family.

IJSER

## 5. Methodology

Research design of this study is descriptive in nature. A questionnaire is developed and data are collected using the survey method. The sample of this study is taken through a survey which has been conducted on different age groups of women who lives in the gated community, Bangalore, Karnataka state in India. The set of questionnaires is messaged to a women WhatsApp group on convenience sampling basis and their responses been recorded for the study. The sampling Size is of total 25 respondents who registered their values on demographic variables based on their food preferences. The sources of data collection are primary data from survey and others from journal, papers and internet. Further, to analyze the data, WEKA data mining tool is used.

## 6 Demographic attributes of Women Behavior on Food Products

The strongly influenced factors of the consumer behavior and the purchase decision are cultural, social, personal and psychological characteristics such as age, marital status, education, occupation, income, life-style (activities, interests, opinions) and personality etc. In order to develop suitable marketing strategies, better understanding and observation on those influence factors is very essential for the long-run success of any marketing program. Hence to do so, the details through a survey which has been conducted on different age groups of women who lives in the gated community, Bangalore, Karnataka state in India is consolidated on various dimensions and given in the below table:

TABLE 1
CATEGORIES OF WOMEN CUSTOMER'S (AGE> =25 YEARS) DETAILS

| S. No | | Working Women | | | House-wife | Student |
|---|---|---|---|---|---|---|
| | | Full time | Part time | Work from home | | |
| 1 | Availability of Time for shopping | Less | Average | More | More | Average |
| 2 | % of interest towards shopping | 60 | 65 | 65 | 75 | 80 |
| 3 | % of buying healthy food products | 40 | 55 | 60 | 80 | 50 |
| 4 | % of buying instant food products(RTE) | 30 | 20 | 20 | 10 | 30 |
| 5 | % of buying packaged processed food products(RTM & RTC) | 30 | 25 | 20 | 10 | 20 |
| 6 | % of awareness on food products | 80 | 70 | 75 | 60 | 65 |
| 7 | Amount spent towards food products over a month | 1000 | 800 | 700 | 600 | 1200 |
| 8 | Number of times taken food from outside in a month. | 4 | 1 | 3 | 1 | 5 |
| 9. | Frequency of eating fast food | Once in a week | Twice in a week | Once in a week | Occasionally | 3-4 times in a week. |

The survey is taken from around 25 -30 women of age 25 years and above, from upper middle class society. It is been observed during the survey that, various factors force them to consume or prefer the processed foods and from outside. Those factors are:

- Affordability(Especially fast food) and sharing time with friends and family
- Service, quality and Satisfaction.
- Get together, parties and functions
- Taste and location orientation on its ambience
- Due to children's interest and fun.
- Convenience and time factor
- Relaxation and wide varieties of food stuffs
- When no choice due to sick.

## 7. RESULTS

### i. Comparison and Evaluation of clusters to measure levels of awareness on healthy food products as class attributes using WEKA tool

With the survey details of women, the dataset womendetails that consists of 25 instances and 15 attributes is created in .CSV format (excel sheet). Using the WEKA tool, dataset is tested with the three clustering techniques: K-Means, Density Based cluster and Hierarchical clustering and results been compared and given below:

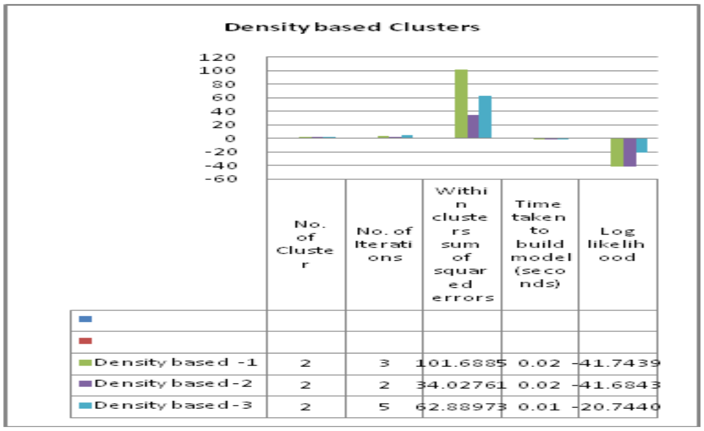
TABLE 2
K- MEANS

| K-Means under different constraints | No. of Cluster | No. of Iterations | Within clusters SSE | Time taken to build model(seconds) |
|---|---|---|---|---|
| K-Means - 1 | 2 | 3 | 101.6885551 | 0.01 |
| K-Means - 2 | 2 | 2 | 34.02761283 | 0 |
| K-Means -3 | 2 | 5 | 62.8897368 | 0.01 |

TABLE 3
HIERARCHICAL

| Hierarchical under different constraints | No. of Cluster | Time taken to build model(seconds) |
|---|---|---|
| Hierarchical -1 | 2 | 0.01 |
| Hierarchical -2 | 2 | 0.02 |
| Hierarchical -3 | 2 | 0 |

TABLE 4
DENSITY- BASED



| | No. of Cluster | No. of Iterations | Within clusters sum of squared errors | Time taken to build model (seconds) | Log likelihood |
|---|---|---|---|---|---|
| 🟦 | | | | | |
| 🟥 | | | | | |
| 🟩 Density based -1 | 2 | 3 | 101.6885 | 0.02 | -41.7439 |
| 🟪 Density based -2 | 2 | 2 | 34.02761 | 0.02 | -41.6843 |
| 🟦 Density based -3 | 2 | 5 | 62.88973 | 0.01 | -20.7440 |

In the above table, k-means-1, hierarchical-1 and density based-1 are the clustering algorithms experimented in WEKA with no filters(preprocessing), where as k-means-2,3 hierarchical-2,3 and density based-2,3 are with filters on attribute removal and instance reduction(principal component). The

IJSER

results of evaluation of clusters to measure levels of awareness on healthy food products as class attributes using WEKA tool is given in the table below:

TABLE 5
MEASURE OF CLASS ATTRIBUTES

| Name | No. of Cluster | Cluster Instances | | No.of Iterations | Within clusters SSE | Time ta to b model (seconds | Log likelihood | Uncluste Instances | Remarks |
|---|---|---|---|---|---|---|---|---|---|
| | | Cluster 0 | Cluster 1 | | | | | | |
| K-means (No filters) | 2 | 15 ( 60%) | 10 ( 40% | 2 | 95.412081475 76589 | 0 | | 7.0 28 % | No filters |
| K-means (With filters) | 2 | 15 ( 60%) | 10 ( 40% | 2 | 58.396471356 71828 | 0 | | 7.0 28 % | attribute. Remove |
| Hierarchical (No filters) | 2 | 24 ( 96%) | 1 ( 4%) | | | 0 | | 12.0 48 % | No filters |
| Hierarchical (With filters) | 2 | 13 ( 52%) | 12 ( 48% | | | 0 | | 5.0 20 % | attribute. Remove |
| Density based (No filters) | 2 | 14 ( 56%) | 11 ( 44% | 2 | 95.412081475 76589 | 0.01 | -41.362 6 | 6.0 24 % | No filters |
| Density based (With filters) | 2 | 15 ( 60%) | 10 ( 40% | 2 | 58.396471356 71828 | 0 | -20.778 84 | 7.0 28 % | attribute. Remove |

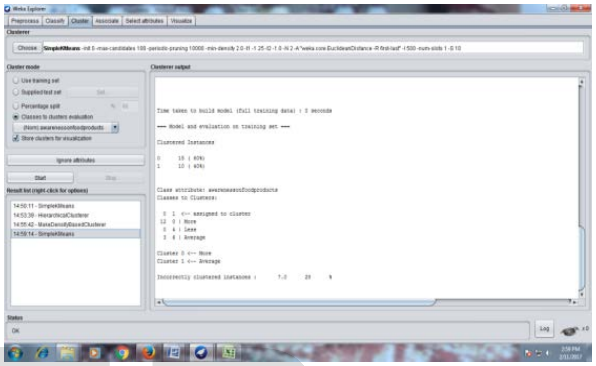
IJSER

Based on time, iterations, SSE and number of instances, the clustering algorithm been tested for the given dataset women-details and listed the observations as follows:

- On preprocessing the dataset with instance and attribute filters (using principle component), the result shows that the value of Sum of squared Error (SSE) and log likelihood varied.
- Also based on the number of clusters and iterations, the value of SSE will be reduced.
- On comparing the performance of Hierarchical Clustering algorithm and density based clusters with K-Means algorithm, it is found to K-means as better one due to:
  - ✓ Density based clusters take more time even though unclustered instances are only 24%
  - ✓ Hierarchical clustering doesn't show any discrimination between clusters and also percentage of unclustered instances is 48% which means nearly 50% of the instances are uncovered and it is more than K-means and density based clusters.

In all the above observations, any clustering technique works well when applied filters (preprocessed) on data. All the clustering algorithms with filters and no filters didn't show any difference in number of clusters and iterations under class attributes whereas the differences is shown in the SSE , log likelihood and uncovered instances.

Using WEKA, the various clustering algorithms been experimented with the dataset for the class attributes awarenessonfoodproducts. The screenshot of k-means is given below:



SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 2 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10

● Classes to clusters evaluation
(Nom) awarenessonfoodproducts
☑ Store clusters for visualization

14:50:11 - SimpleKMeans
14:53:38 - HierarchicalClusterer
14:55:42 - MakeDensityBasedClusterer
14:59:14 - SimpleKMeans

```
Time taken to build model (full training data) : 0 seconds

=== Model and evaluation on training set ===

Clustered Instances

0      15 ( 60%)
1      10 ( 40%)


Class attribute: awarenessonfoodproducts
Classes to Clusters:

  0  1  <-- assigned to cluster
 12  0 | More
  0  4 | Less
  3  6 | Average

Cluster 0 <-- More
Cluster 1 <-- Average

Incorrectly clustered instances :     7.0     28     %
```

OK

FIG 1 : K- MEANS

The result of all the three clustering is been consolidated in the below table:

TABLE 6
MEASURE OF CLASS ATTRIBUTES

| S. No | Clustering techniques | Cluster 0 | | | Cluster 1 | | |
|---|---|---|---|---|---|---|---|
| | | More | Less | average | More | Less | average |
| 1 | K- Means | 12 | 0 | 3 | 0 | 4 | 6 |
| 2 | Hierarchical | 12 | 4 | 8 | 0 | 0 | 1 |
| 3 | Density based | 12 | 0 | 2 | 0 | 4 | 7 |

From the above table, it is very clear that women in cluster 0 have maximum awareness on healthy food products whereas women in cluster 1 have average awareness. By refining and focusing on the instances which has less and average awareness, maximum level of awareness can be obtained.

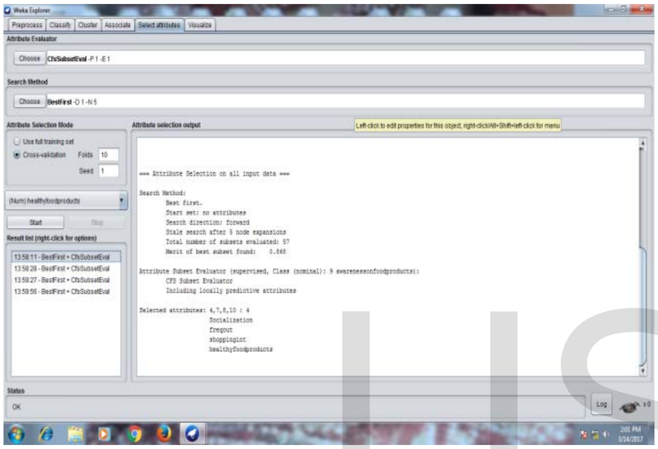
### ii. Identify the factors that influence women consumer's awareness on healthy food products

The influencing factors of women to prefer processed foods depend on the nature of their work and their responsibilities they perform at home as well as at office. The categories of the women based on their roles are as follows:

1. if women works , she has an option of choosing her nature of job that can be of full time, part time and work from home.

2. else if women does business, she do it at home or else outside.
3. else if women is doing higher education which is away from her hometown or in abroad etc..
4. else a great homemaker.

The change from traditional to modern orientations has ushered due to an increase in educational and awareness levels, and contribution to the family income by the women sometimes force to utter and accept the changes in food culture. The attribute selection using CFS- subset evaluator with BestFirst search is applied on womendetails dataset preprocessed by remove attributes filters with 25 Instances and 12 attributes using WEKA is given below:



Attribute Evaluator
CfsSubsetEval -P 1 -E 1

Search Method
BestFirst -D 1 -N 5

Attribute Selection Mode
○ Use full training set
● Cross-validation Folds 10
Seed 1

(Num) healthyfoodproducts

Result list (right-click for options)
13:58:11 - BestFirst + CfsSubsetEval
13:58:28 - BestFirst + CfsSubsetEval
13:59:27 - BestFirst + CfsSubsetEval
13:59:56 - BestFirst + CfsSubsetEval

Attribute selection output

```
=== Attribute Selection on all input data ===

Search Method:
        Best first.
        Start set: no attributes
        Search direction: forward
        Stale search after 5 node expansions
        Total number of subsets evaluated: 57
        Merit of best subset found:    0.868

Attribute Subset Evaluator (supervised, Class (nominal): 9 awarenessonfoodproducts):
        CFS Subset Evaluator
        Including locally predictive attributes

Selected attributes: 4,7,8,10 : 4
                     Socialization
                     freqout
                     shoppingint
                     healthyfoodproducts
```

Status
OK

Fig 2: Select Attributes – CFS –BFS

IJSER

The attribute Subset Evaluator on the nominal attribute awarenessonfoodproducts predicts the following attributes as Socialization, freqout, shoppingint and healthyfoodproducts. The evaluation is carried on all training data in a forward direction providing 57 evaluated subsets with the merit of finding the subset which is best as 0.868.

## 8. Conclusion and Future Enhancement

The study concludes that the role of women as a professional, student and homemakers need to know the importance of awareness on healthy food products through the influencing factors like socialization, frequently out for shopping and knowledge of shopping towards healthy food products. The sources of energy in the form of refined cereals, sugars, oils may lead to risk factors like overweight and obesity, high blood pressure, etc reported by Public Health Foundation of India, an autonomous foundation in New Delhi.

According to NSSO data analyses on food consumption, as calorie and protein consumption have fallen, there will be an increase in fat intake which leads to heart problems. Also it is reported that pattern of food consumption is moved away from pulses and cereals to more into beverages, edible oil, dry fruits and other processed products.

The scope of the paper highlights on women understanding towards the knowledge on food consumption always makes the healthy society without diabetics and cardiac problems. To make our future India as a healthy nation, the role of women is very important. Finally it is a recommendation for the food administrators and entrepreneurs to build a healthy nation by promoting and improving the food product strategies that helps to attain good health in every consumer.

The future work is to mine the scalability issues between food preferences (consumption) based on the income against food product supply (availability) in urban and rural areas in Bangalore.

## 9. References

1. Mr. Havish Madhvapaty, Ms. Aparajita Dasgupta, Study of Lifestyle Trends on Changing Food Habits of Indian Consumers, IOSR-JESTFTe-ISSN: 2319-2402,p-ISSN: 2319-2399.Volume 9, Issue 1 Ver. II (Jan. 2015), PP 16-22, www.iosrjournals.org.
2. 2017, India Brand Enquiry Foundation (IBEF), www.ibef.org.
3. The India Food Report 2016- The term report on the world's exciting Food Market.
4. Mr. Vijayabaskar. M, Dr. N. Sundaram, A Market Study On Key Determinants Of Ready – To - Eat/Cook Products With Respect To Tier - I Cities In Southern India, International Journal of Multidisciplinary Research, Vol.2, Issue 6, June 2012, ISSN 2231 5780.
5. A. Berson, S. J. Smith, and K. Thearling. Building Data Mining Applications for CRM. McGraw-Hill Osborne, New York: Wiley, 2000.
6. J. Rong, H. Q. Vu, R. Law, and G. Li. A behavioral Analysis of web sharers and browsers in Hong Kong using targeted association rule mining. TourismManagement, 3(4):731–740, 2012.
7. T. Allard, B. Babin, J. C. Chebat, and M. Crispo, "Reinventing the branch: An empirical assessment of banking strategies to environmental differentiation," Journal of Retailing and Consumer Services, vol. 16, no. 6, pp. 442-450, 2009.
8. De Nisco, and G. Warnaby, "Urban design and tenant variety influences on consumers' emotions and approach behavior," Journal of Business Research, In Press, 2012.
9. Brunner-Sperdin, M. Peters, and Strobl, A , "It is all about the emotional state: Managing tourists' experiences," International Journal of Hospitality Management, vol. 31, no. 1, pp. 23-30, 2012.
10. R. Kittler, and W. Wang, "Data mining for yield improvements. Proceedings from MASM, 2000.

11. Nicola Barbieri, Massimo Guarascio, Giuseppe Manco, , A Block Mixture Model for Pattern Discovery in Preference Data, 2010 IEEE International Conference on Data Mining Workshops.
12. Manish Verma, Mauly Srivastava, Neha Chack, Atul Kumar Diswar and Nidhi Gupta, "A Comparative Study of Various Clustering Algorithms in Data Mining", International Journal of Engineering Research and Applications (IJERA) Vol. 2, Issue 3, May-Jun 2012, pp.1379-1384.
13. Chad West, Stephanie MacDonald, Pawan Lingras, and Greg Adams, Relationship between Product Based Loyalty and Clustering based on Supermarket Visit and Spending Patterns, International Journal of Computer Science & Applications. Vol. II, No. II, pp. 85 – 100, 2005.

IJSER