INGREDIENT CONSUMPTION TRACKER

APPLICATION

Enrol. No. (s) - 9917103009, 9917103204, 9917103227

Name of Student (s) - Anshika Bajpai, Vishal Kumar, Rohan Kumar

Name of Supervisor (s) - Dr. Himanshu Mittal

May 2021

Submitted in partial fulfillment of the Degree of

Bachelor of Technology

in

Computer Science Engineering

DEPARTMENT OF COMPUTER SCIENCE ENGINEERING & INFORMATION TECHNOLOGY

JAYPEE INSTITUTE OF INFORMATION TECHNOLOGY, NOIDA

Table of Contents

DECLARATION

I/We hereby declare that this submission is my/our own work and that, to the best of my knowledge and belief, it contains no material previously published or written by another person nor material which has been accepted for the award of any other degree or diploma of the university or other institute of higher learning, except where due acknowledgment has been made in the text.

Place: JIIT, Noida
Date: 15-05-2021

Signature: Anshika Bajpai
Name: Anshika Bajpai
Enrollment No: 9917103009

Signature: Vishal Kumar
Name: Vishal Kumar
Enrollment No: 9917103204

Signature: Rohan Kumar
Name: Rohan Kumar
Enrollment No: 9917103227

CERTIFICATE

This is to certify that the work titled **Ingredient Consumption Tracker Application** submitted by **Anshika Bajpai, Vishal Kumar and Rohan Kumar** in partial fulfillment for the award of degree of **B Tech.** of Jaypee Institute of Information Technology, Noida has been carried out under my supervision. This work has not been submitted partially or wholly to any other University or Institute for the award of this or any other degree or diploma.

Signature of Supervisor	Dr. Himanshu Mittal
Name of Supervisor	Dr. Himanshu Mittal
Designation	Dept. of CSE, Jaypee Institute of Information Technology
Date	15 May 2021

ACKNOWLEDGMENT

I would like to place on record my deep sense of gratitude to Dr. Himanshu Mittal, Dept. of CSE, Jaypee Institute of Information Technology, India for his generous guidance, help, and useful suggestions.

I also wish to extend my thanks to Anushka Bajpai and Anushka Bhargava for their insightful comments and constructive suggestions to improve the quality of this project work.

Signature of Supervisor	Dr. Himanshu Mittal
Name of Supervisor	Dr. Himanshu Mittal
Designation	Dept. of CSE, Jaypee Institute of Information Technology
Date	15 May 2021

SUMMARY

In this project we present an approach towards studying the pattern of consumption of packaged food ingredients. The research consists of logging data in an efficient manner as well as analyzing the logged ingredients data using multiple algorithms like Apriori, ECLAT & FP Growth. We proposed an approach towards improvising the Apriori Algorithm in a manner so that the resources and data can be best utilized. Different datasets were analyzed in a way so that the pattern among packaged food consumption can be noticed. The algorithms successfully ran on various datasets and marginal differences were observed.

This research provided the authenticity that packaged food items can be traced along with ingredient consumption among users. The study opened the approach towards recording and analyzing datasets of daily consumption tracking to find respective patterns.

Anshika Bajpai
Name of student
Vishal Kumar
Name of student
Rohan Kumar
Name of student

15 May 2021
Date

Dr. Himanshu Mittal
Name of the supervisor
15 May 2021
Date

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS AND ACRONYMS

CO2	Carbon Dioxide
FP	Frequent Pattern
ECLAT	Equivalent Class Clustering and bottom up Lattice Traversal
Secs	Seconds
MSG	Monosodium Glutamate
PUFA	Poly-Unsaturated Fatty Acid
NHANES	National Health and Nutrition Examination Survey

1. Introduction

1.1 General Introduction

Traditionally Indian population has consumed a diet based upon fruits, vegetables and unprocessed cereals. National Nutrition Surveys done over the past 25 years show that there is a significant change in the consumption patterns particularly in urban and rural areas. Increasing per capita disposable income, alterations to lifestyle and changes in food environment are mainly due to increase in the consumption of processed, ready-to-eat and fast food items.

In a study it is shown that increase in consumption of processed foods and ready-to-eat foods has gone up together with income. This is more common in urban areas than rural, which consists of about 20 percent of population.Consumption of beverages, biscuits, processed foods, salted snacks, prepared sweets and other purchased foods constitutes 100-427 g/per capita/day, from the lowest to the highest expenditure class. Average consumption is about 167 g/per capita/day.

This is probably an indication of the increasing consumption of snack foods and high-calorie foods such as sweets, which are purchased away from the home. Energy-wise they may contribute to as many as 1000 kcals in the daily diet, or 30 to 40 percent of the required calories in the high-income classes.[2]

Healthy wholesome foods have been replaced by junk food. Junk food is cheap, processed and prepackaged, making it easily available for everyone. Consumption of junk food increases the risk of many chronic diseases which can have a serious impact on the quality of life of consumers like obesity, cardiovascular diseases, nutritional deficiencies, liver harm, mental health problems because of added fats, refined grains, sugars, sweeteners and sodium in it.

1.2 Problem Statement

With rise in globalisation and easy access to resources, consumption of packaged food items has taken a surge among different age groups. Every age group consumes packaged food items based on their requirements or comforts. There are many studies and research going around the globe based on types of items being consumed by the people of different demographic groups, races, countries.

The variety of packaged food items has been increasing due to high demand among the population. Whether it's a rural area or urban area packaged food items are gaining their daily markets day by day. There are several startups and well established organizations who are doing research about how many new products can be launched based on the consumption pattern among the different regions.

There is a need to study the pattern of ingredients that are being consumed by the different age groups and to make people aware about the harmful effects of ingredients which they consume. People of different regions, easiblity, financial status, etc may be consuming the ingredients intentionally or unintentionally. If there could be any platform which can detect, analyze, find patterns among these can be helpful to detect early signs of any sort of uneasiness.

The processed food is high in energy density, saturated fat, sugar and salt which is why consumers are now experiencing a nutrition transition. Local data suggest that these processed food items constitute a similar proportion of the food supply across multiple regions of India. These processed foods can improve nutritional status, while a shift towards high-fat, high-sugar snack foods may lead to obesity and chronic diseases. Another adverse impact may be displacement of the poor by structural reforms.The poor experiences lower affordability and a reduced calorie intake leading to growth disorders, such as stunting. A more severe impact may be on higher infant mortality rates and lower life expectancies.

1.3 Empirical Study

Adolescents are the most vulnerable to these chronic diseases as between 10-19 years the major chronic degenerative diseases start or are reinforced. This is mainly due to their improper food habits. The food intake of adolescents in developed countries such as the USA, UK and Australia is not balanced, it doesn't contain all the nutrients and fibers required for a proper healthy diet. Adolescents from these countries have high rates of consumption of energy-dense, nutrient poor foods because they lack consumption of fruits and Vegetables, this does not meet the normal dietary guidelines of food In addition, adolescents also exhibit unhealthy eating habits such as meal skipping, untimely eating habits and snacking on fast and packaged foods. India also reported similar findings.

Consumption of fast food among youngsters has significantly increased in India. In a study it is revealed that Between 2007 and 2010, the average American adult got approximately 11.3 percent of his calories from processed food, with the result concluding younger people consuming more fast food than older people, according to the Centers for Disease Control and Prevention. The nutrient content of processed food lacks many important nutrients and fibers which people get from vegetables, putting people who consume a lot of fast food at higher risk for nutrient deficiencies.

1.4 Brief Description of Solution Approach

We propose an approach as a solution to this problem summarized as following:

1. A Dataset was generated based on Firebase ML Kit Text Recognition where users can log their packaged food consumption data, with the help of Android Application.
2. A public Dataset was found consisting of ingredients available in packaged food, which can be used for our preprocessing of data.
3. Once the data is gathered, we proceed towards preprocessing of data.
4. Testing the data on multiple algorithms like Apriori, ECLAT, FP Growth.
5. It is found that FP Growth algorithm was the most efficient algorithm to use for the dataset available.
6. Displaying the frequent set data on chart visualization, keeping most frequent consumed data on top.
7. Created an Ingredient-Food dataset for suggesting packaged food items.
8. Created an Ingredients-disease dataset using the research based on the diseases caused by harmful ingredients in packaged food items.
9. Checking the disease vulnerability.
10. Applying and modifying the efficient algorithms to achieve the results.
11. Suitable steps to search and suggest the food items.
12. Created a dataset which categorizes ingredients based on the Food categories of India.
13. Calculated HSR rating for food items and suggested healthier packaged food items to users.

2. Literature Survey

2.1 Summary of papers studied

While different researches have been done on the vast topic of food processing covering different points and impacts on society. There still remains much to cover. But the collection of some of these individualistic researches can be put together to focus on a particular area, as done by us.

More recent studies explained in the paper have found that, nutritionally speaking, conventionally grown foods and organically grown foods have no significant overall differences in their nutritional quality.

Though the processed food we consume today has become a part of our daily lives and requirements, it is highly affected by global industries.The influence of the link between globalization and and its effect on food intake can be substantially high , although it works mostly through employment, incomes, prices and the market influence on food preferences. The second expected impact is the shift towards fast foods. Third is the market influence of popular processed foods.

And this change in diet patterns and leaning more towards processed and packaged foods has started to show in Indian families. According to the Nutrition Survey of 20 years [3], consumption patterns are now changing, particularly in high- and middle-income groups. Increase in per capita disposable income, alterations to lifestyle and changes in the food environment are driving consumers towards highly processed products.

This paper [5] also discusses the major reasons for the increase in consumption of packaged food which are as follows:

- Choosing Convenience: Busy schedules typically diminish the number of your time individuals ought to prepare healthy, nutrient meals, in order that they take quicker, easier choices. Over time, that convenience becomes a habit and eventually a perceived necessity to stay up with such a fast society. Disrupting that routine needs an investment of your time, and the general public chooses to follow the quicker possibility.
- Easing Anxiety: In the present world, the level of stress and anxiety has increased forcing people to consume more calorie-dense and fatty, sweet food. High levels of hysteria conjointly cause individuals to hunt out food as a method of comfort. Once stressed,

individuals seek ways to calm themselves, and junk food's positive effects on the reward center of the brain create it a comforting go-to alternative.

- Insomnia: There is conjointly proof to counsel that sleep deprivation motivates individuals to decide on junk foods over healthy foods.
- Addiction: People may opt for food just because they need to develop a gentle physical dependence on that. Studies show that binge ingestion foods high in sugar or fat leads to organic compound changes within the brain just like those who develop in addiction.

Also, according to [5] the most common ingredients found in processed/packaged food today are:

- Chicken: The most common meat product
- Xanthan Gum: The most common stabilizer or thickener
- Mono- and Di- glycerides: The most common emulsifiers
- Soybean Oil: The most common oil or fat
- Niacin: The most common nutrient
- Monosodium Glutamate: The most common flavor enhancer
- Salt: The most common flavor or spice
- Caramel Color: The most common color additive
- High-fructose Corn Syrup: The most common sweetener
- Citric Acid: The most common preservative

For data preprocessing, It is seen that several terms are irrelevant, for instance the terms extracted from the statement "100% pure", here "100%" and "pure" are to be ignored. These adjectives are mainly used to describe ingredients and for advertisement purposes. Many such irrelevant words that one can come across are "ounce", "teaspoon", "chopped", etc. Moreover in most of the cases it is seen that a particular term is occurring in both singular as well as in its plural form, for instance "strawberry" and "strawberries".

The main findings of the study are as follows: [36]

- Ingredients are used in recipes. It is a combination of the food substance along with the quantity of the substance used in a recipe.
- It is found that condiments like sugar, oil, pepper and salt are the most commonly occurring ingredients.
- Amongst spices, cloves are most frequently used in a variety of recipes.
- Of all the vegetables onion, garlic and tomatoes are widely used as per the dataset is concerned.

- Considering milk products, butter is the most frequently ingredient in several recipes
- followed by cheese, cream and milk.
- When animal products are considered it is found that eggs are ranked first in the list followed by chicken.

2.2 Integrated summary of the literature studied

With rise in globalisation and easy access to resources, consumption of packaged food items has taken a surge. Now, packaged food items have become an integral part of our lives, with the increase in consumption of packaged food and sedentary lifestyle people are facing health related problems and are linked to environmental risk transition. Due to its availability, affordability, accessibility, acceptability and accomodation there is an increase in massive amounts of processed food intake which can result in obesity and health conditions such as high blood pressure.

These processed foods, junk food, fast food, ready-to-eat food have low protein content that comes mainly from fruits and vegetables, and are high in fats, sugar, carbohydrates and energy density. Fast food is considered as empty calorie food, that is, it lacks in micro-nutrients such as vitamins, minerals or amino acids and fiber.

People living in urban areas tend to consume more diversified diets, which mostly consists of higher input grains, animal products, more processed foods as well as more food away from home.This hypothesis is tested in India. India, having less than one-third of the total population living in urban areas and still in the early stages of its urban demographic transition.Many urban areas still lack access to basic amenities such as electricity.

India has some large metropolitan areas such as New Delhi, Mumbai, or Bangalore. That is why analyzing the consumption in India especially in these areas will help to inform what can be expected in the future. Furthermore, regarding income, urban households spend more on processed foods compared to metropolitan households.

This relationship is different in the higher income quartiles: in the third, both spend approximately the same on processed foods (about 4%), while in the fourth, metropolitan households spend 0.5 percentage points more (6.3% compared to 5.8%). The results from 2000 confirm the general

picture of increasing consumption with increasing income and variation within income quartiles[14].

Consumption of processed food and mainly fast food has become a global phenomenon and is most common among young adults and adolescents. A study shows that 36% of students eat fast food meals more than three times per day. And 48% of the students who were overweight and 23% of those who were obese percieved themselves as being in the healthy weight category.[5]

The major challenge before starting the analysis was to define the domain of research. As mentioned already the packaged food items are being consumed by a large amount of the population chunk which is increasing day by day. The randomness impacts a lot while analyzing datasets for better results.

We improvised dataset collection in various ways including manual input. Defining age group and domain of research was possible because of the collected data.

Followings are the dataset which were found/analyzed during our research,

1. Packaged food dataset.

   This dataset we found was created by a survey, there was a group of 4-5 people whose 7 days consumption was recorded. All the data recorded was of packaged foods. This dataset was used for our research and analysis of algorithms. It comprises 32 packaged food items which users consume on a daily basis.

2. Ingredient Dataset

   This dataset eased the way of approaching ingredients logging into the database. It contains 10,000 packaged food item names along with the ingredients present in it. There were other data available as well which can play a key role while analyzing the chunk. We will be using this dataset for tracking ingredients.

3. Manually tracking daily consumption

   We expanded our research in all directions. In this dataset users logged everyday food-intake manually according to breakfast, lunch and dinner along with quantity.

4. Restaurant Food Dataset

In this dataset we came to know about all foods that are being consumed and can be traced as well in restaurants. This defined the domain heavily as consumption of food out-from-home. There were a significant amount of inputs which again raised the curiosity of analyzing ingredients consumption.

This dataset was not user specific, it was heavily relying upon the variety of people coming to the restaurant.

5. Food Categories dataset

A dataset of ingredients categorized according to the above categories were made using Food Categorisation of India. We categorized the food ingredients into 12 categories:

- Eggs
- Seafood
- Cereals
- Oil products
- Dairy items
- Non-alcoholic Beverages
- Sauces and Dressings
- Meat products
- Bread and Bakery products
- Sugar and Honey items
- Confectionaries

6. Ingredient Weight dataset

This dataset comprises food items, their ingredients and weight/portion of those ingredients. This dataset can be used for further research to check the quantity of ingredients consumed.

7. Age and Gender - wise dataset

Age-wise the dataset consists of food consumption and their nutrition intake of the food items consumed by the user. Gender-wise data sets consist of the food choices by user and their nutritional intake.These two factors are one of the crucial parts of the research. Our domain was defined majorly as teenagers and college students. Gender too can play a major role. These two datasets can be used in future for further research work.

8. Ingredients - Disease Dataset

The disease dataset eased the way of searching and classifying the ingredients while running algorithms.

For example,

1. Alzheimer's Disease, Aluminum-induced bone disease, microcytic anemia, Aluminum toxicity in the premature infant: Aluminium(used in milk, processed cheese, yogurt, cheese, jams and preserves.)
2. Asthma, hypertension, iron deficiency, decrease in liver transaminases: MSG(Corn starch, Corn syrup, Modified food starch, Lipolyzed butter fat, Dextrose, Rice syrup, Brown rice syrup, Milk powder, etc)

Ingredients	obesity, metabolic disorders
agents with minerals	1
Ajinomoto	1
and glutamate act as chelating	1
Autolyzed yeast	1
Brown rice syrup	1
Calcium caseinate	1
Calcium glutamate (E 623)	1
cheese	0
Corn starch	1
Corn syrup	1
Dextrose	1
Fermented foods	1
Fortified protein	1
Gelatin	1
Glutamate (E620)	1
Glutamic acid (E 620)	1
Hydrolysed protein	1
jams and preserves.	0
Lipolyzed butter fat	1
Low fat products	1
Magnesium glutamate (E 625)	1
milk	0
Milk powder	1

Fig1. Disease Dataset

3. Requirement Analysis and Solution Approach

3.1 Overall description of the project

In this project we present an approach where users can track their pattern of consumption of packaged food ingredients. Different datasets were collected and generated in a way so that the pattern among processed food items consumed can be tracked. The research consists of logging data in an efficient manner as well as analyzing the logged ingredients data using multiple mining algorithms for tracking the most frequent items consumed by the user. We also proposed an approach towards improvising the Apriori Algorithm in a manner so that the resources and data can be best utilized. Alternatives for Apriori algorithms such as ECLAT and FP Growth algorithm were also studied in order to reduce the time complexity and space complexity. The algorithms successfully ran on various datasets and marginal differences were observed.

We then used the Fp growth algorithm as it was the most efficient algorithm based on our research. Data preprocessing was done on the Ingredients dataset which was then passed in the most efficient algorithm to find the most frequent ingredients consumed by the user. The results were then used for the suggestion of packaged food items and health tip generation.

This research provided the authenticity that packaged food items can be traced along with ingredient consumption among users. The study opened the approach towards recording and analyzing datasets of daily consumption tracking to find respective patterns.

3.2 Requirement Analysis

3.2.1 Functional Requirement

The application aims at providing an efficient output to the user in the form of a graph for the user to know the most consumed ingredient and manage the future diet according to it. Application also provides an interface to look at the image scanned of the packaged food item. The algorithm used for processing the output is designed to generate the most efficient output and form the graph. The ingredient levels are continuously monitored based on the user's usage and are checked for the threshold levels in the database and accordingly the user is alerted about low levels of certain ingredients. The design is such that the user does not have to manually update the user database every time, the application automatically does if for the user.

3.2.2 Non-Functional Requirements

1. Usability
    - The application must be easy to use by any user such that they do not need to read an extensive number of instructions.
    - The application must be quickly accessible.
    - The output generated should be clearly visible, easy to read and accurate.
    - The application must be intuitive and simple in the way it displays all relevant data.
    - The application must be easily navigable by the users with buttons that are easy to understand.
2. Reliability
    - The application must give accurate output status to the user and continuously alert according to the working of the application. Any inaccuracies are to be taken care of.
    - The application must provide data protection.
    - The application should provide the user updates on completion of requested processes and if the requested processes fail, it should provide the user the reason for the failure.
    - The application should not update the data in any database for any failed processes.

3. Performance
    - The application must not lag, because the user using it doesn't have down-time to wait for it to complete an action.
    - The application must complete updating the databases, adding of ingredients and occasions successfully every time the user requests such a process.
    - All the functions of the application must be available to the user every time the system is turned on.
    - The calculations performed by the system must comply according to the norms set by the user and should not vary unless explicitly changed by the user.
    - The application shall have enough memory space in order to store a high number of data.

4. Supportability
    - The application is designed such that it works even on systems having the minimum configuration.
    - The application should be able to tell the user about the necessary hardware and system requirements and ask for all the permissions needed.
    - The data can be exported to the database so as to make the system more portable.

5. Implementation
    - The application is built on Android Studio and specifically designed for android smartphones.
    - The connection between the database and the application is achieved by using Firebase tools.

6. Interfacing
    - The application must offer an easy and simple way of viewing the current input and output.
    - The application interface must be easy to use, avoiding any hectic environment and able to guide the user to the next step without much explanation.

7. Legality and Ethical Requirements
    - The application shall be license free.
    - The application should only have access to the required hardware and that too by user's permission after asking for it.
    - Personal information of the registered user shall only be accessed by themselves.

3.3 Solution Approach

With the widespread use of mobile-phones and tablets there has been an increase in the number of software applications that record and aim to improve people's food consumption behaviour. The

need for more suitable applications for tracking and understanding the nutritional intake of the consumer's food intake has sparked interest amongst behavioural, nutritional researchers and developers for these digital solutions. Smartphones and their implemented technologies such as image scanners, databases and chatbots have the potential to enhance the accuracy and efficiency of data collection. Various algorithms have made it easier for developers to collect data for research work on a real time basis.

Studies have found the market of packaged food items has been increasing since a long time with the rise of demand approximately exponentially.
We have projected a method where users can log the data based on image click of ingredients chart. This too can be automated if the dataset of ingredients of packaged food items with name is available.

We can generate the dataset based on Android Platform with Firebase where users can log their packaged food consumption data. Users will scan the ingredients section of the packaged food item which they are going to consume, if the item is present in our database then it is used from our database otherwise a new entry for that item is created. The ingredients scanned will be in JSON format and will be used for further analysis.

Data preprocessing is done on the ingredients scanned for example and redundant ingredients are dropped from the dataframe for example, carbon dioxide, food coloring, type of salt are redundant.
Once the data is gathered, we will find the most frequent ingredients consumed by the user. Data is again cleaned, necessary and the frequent items consumed in the dataset are analysed separately.

We will then test the data on multiple mining algorithms such as Apriori, Eclat, FP growth to find the most frequent sets of ingredients consumed by the user. Time complexities of these algorithms will be compared and analysed on various datasets. Only the best algorithm will be used to find the frequent combinations. The result is then displayed in the form of line plots and pie charts in the mobile application.

Using the disease dataset, the extracted frequent ingredients will be compared if they contain any kind of disease vulnerability. Once compared and analyzed the output will consist of suggestive ingredients which can be consumed and would not affect with any kind of exposure to prone disease.

4. Modeling and Implementation Details

4.1 Control Flow Diagrams

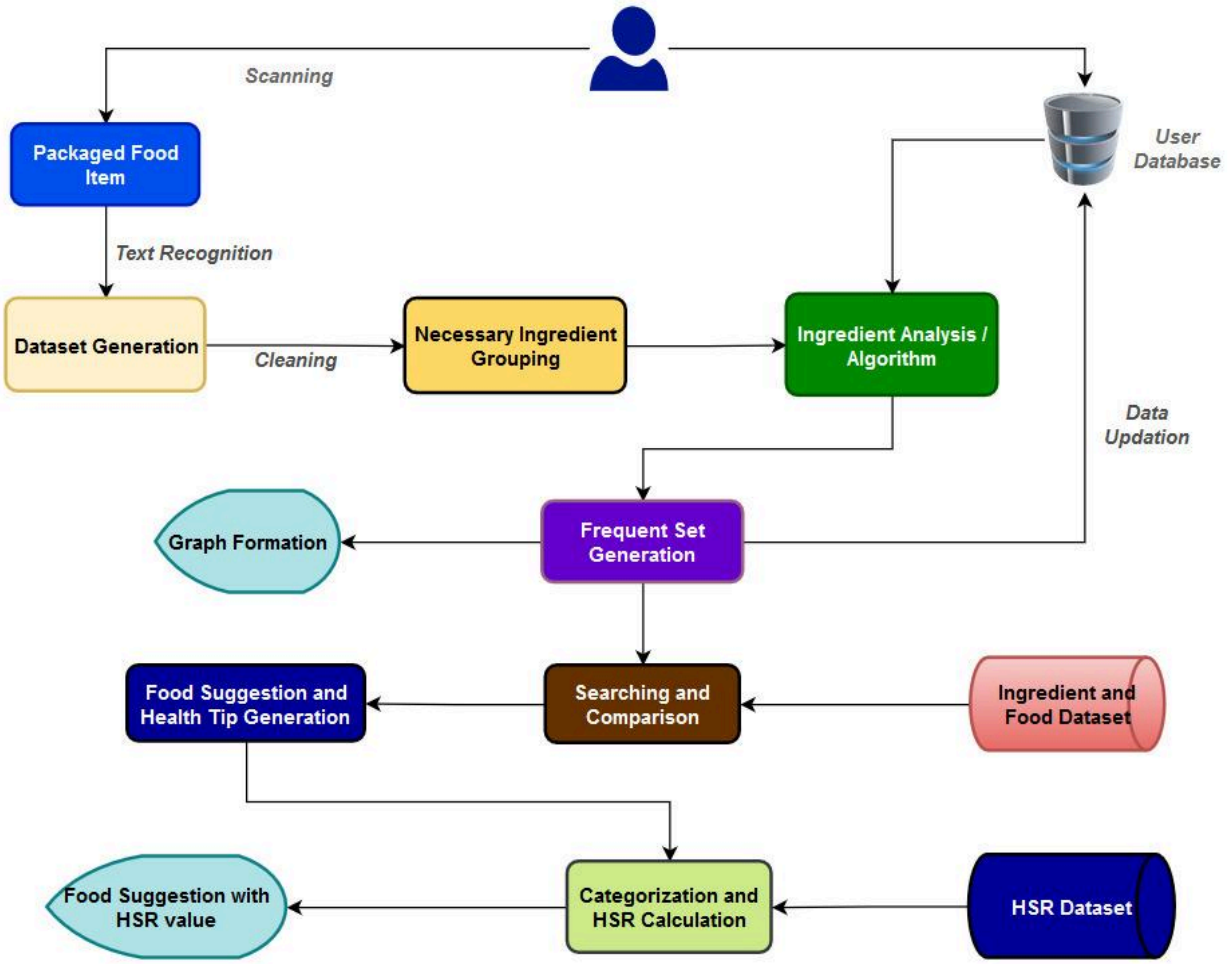


Fig.2 Control Flow Diagram

4.2 Implementation details and issues

We have used Android as a platform to generate the authentic database in a manner so that the ingredients being logged can be analyzed on our various algorithms. The android app is highly dependent upon Firebase for services like,

1. Secure Authentication

   To record data of food ingredient consumption can have many other perspectives and implementations. But to get the authentic data the user must feel safe about their privacy and type of data being stored. Firebase Authentication provided the platform where a secure onboarding of our target audience can be performed.

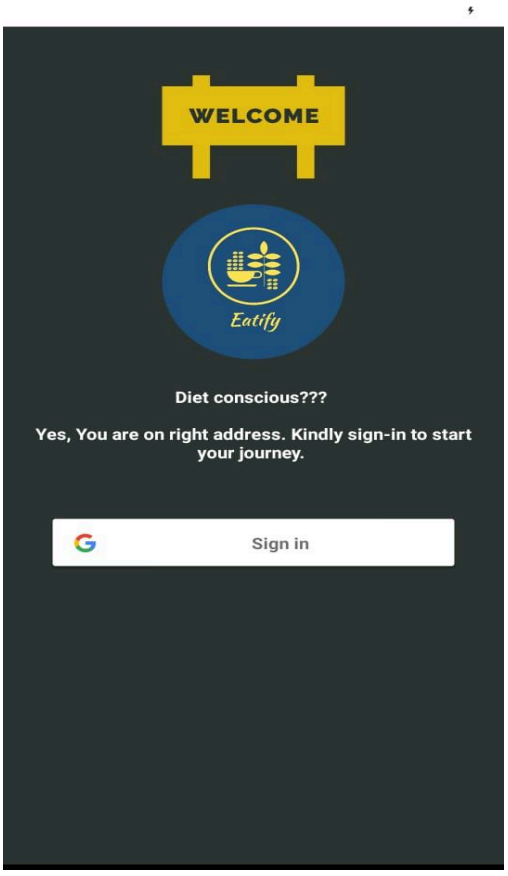


Fig.3 Firebase Authentication

2. Profile Details

   There are many such factors which can be responsible while analyzing and processing the data. Since the target user base was adolescents and college students, details like date of birth and gender were the key factors which should be recorded. That not only helped in

frequent ingredient pattern research, but also the kind, sub-age group and demography which is having keen interest in knowing about their food consumption pattern. At the time of registration, the user is asked to enter their health details, gender, weight, height, Body Mass Index, diseases (if any), deficiencies, etc. The data entered by the user will be updated later on according to the user's consumption of food.

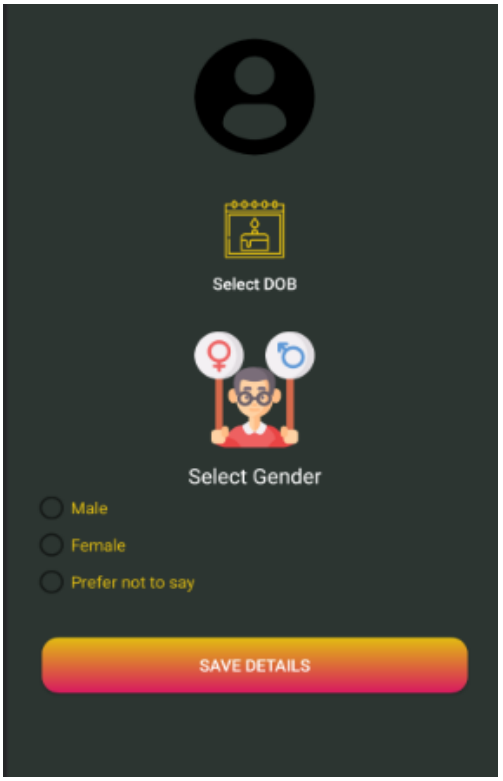


Fig.4 Profile Details

3. Data logging

The challenging part about collecting the dataset to have a proper well established user experience setup. For that we found out a simple click based approach to achieve the data logging easeness. Firebase MLKit played the major role in easing the task. MLKit text recognition service using OCR helped users to log the data in a much better way.

4. Data Storing

All the data is stored in the Firebase Database and is analyzed using the proper algorithm. Firebase provides services like a real-time database and backend. An API is provided to the

application developer which allows application data to be synchronized across clients and stored in Firebase's cloud. The ingredient data subset will be then analyzed properly with other profile data to give notifications to users if required.

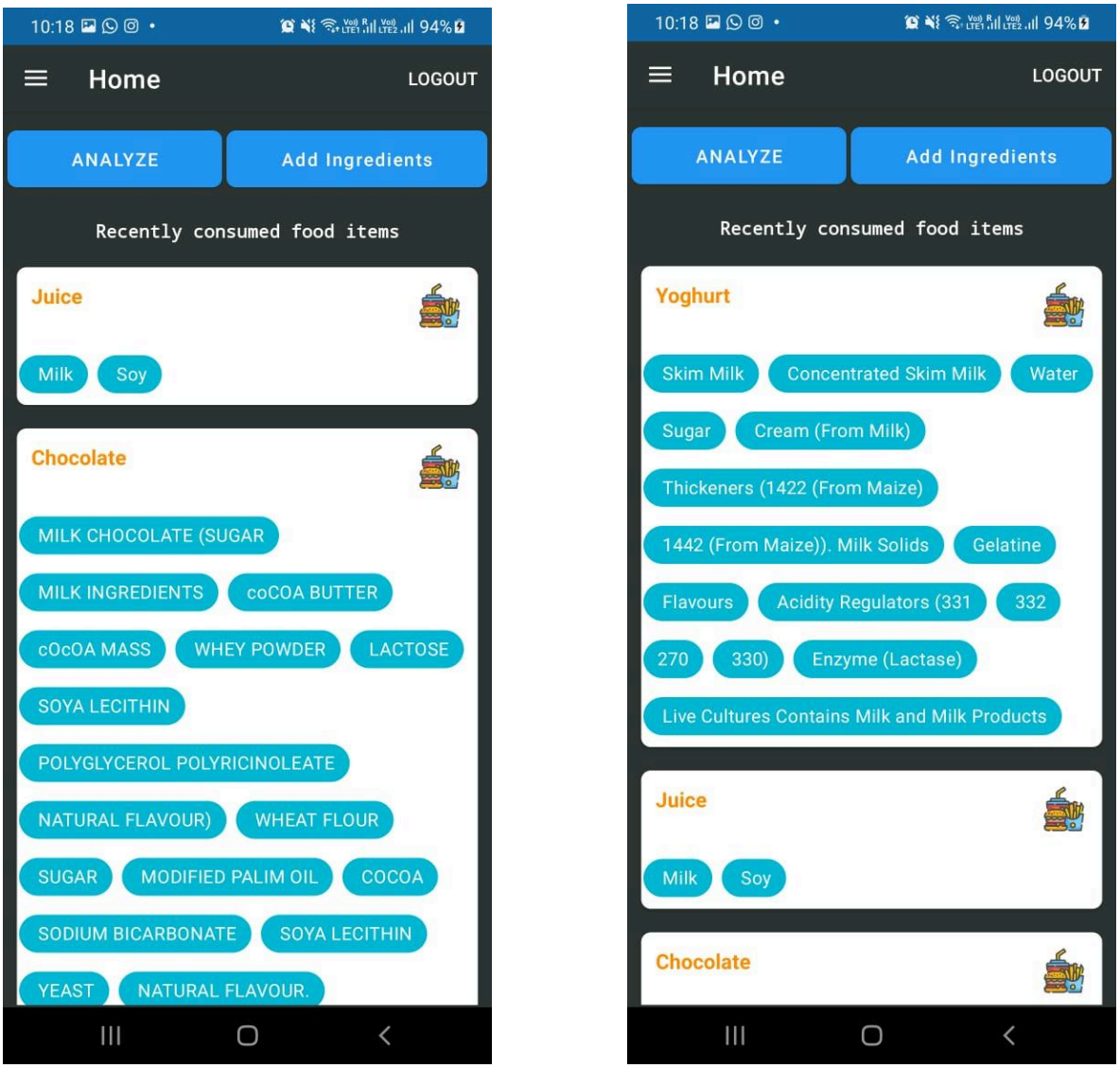


Fig.5 Scanned Ingredients

5. Text Recognition

Text Recognition will be done with the help of Text Recognition API, part of ML Kit for Firebase[26].

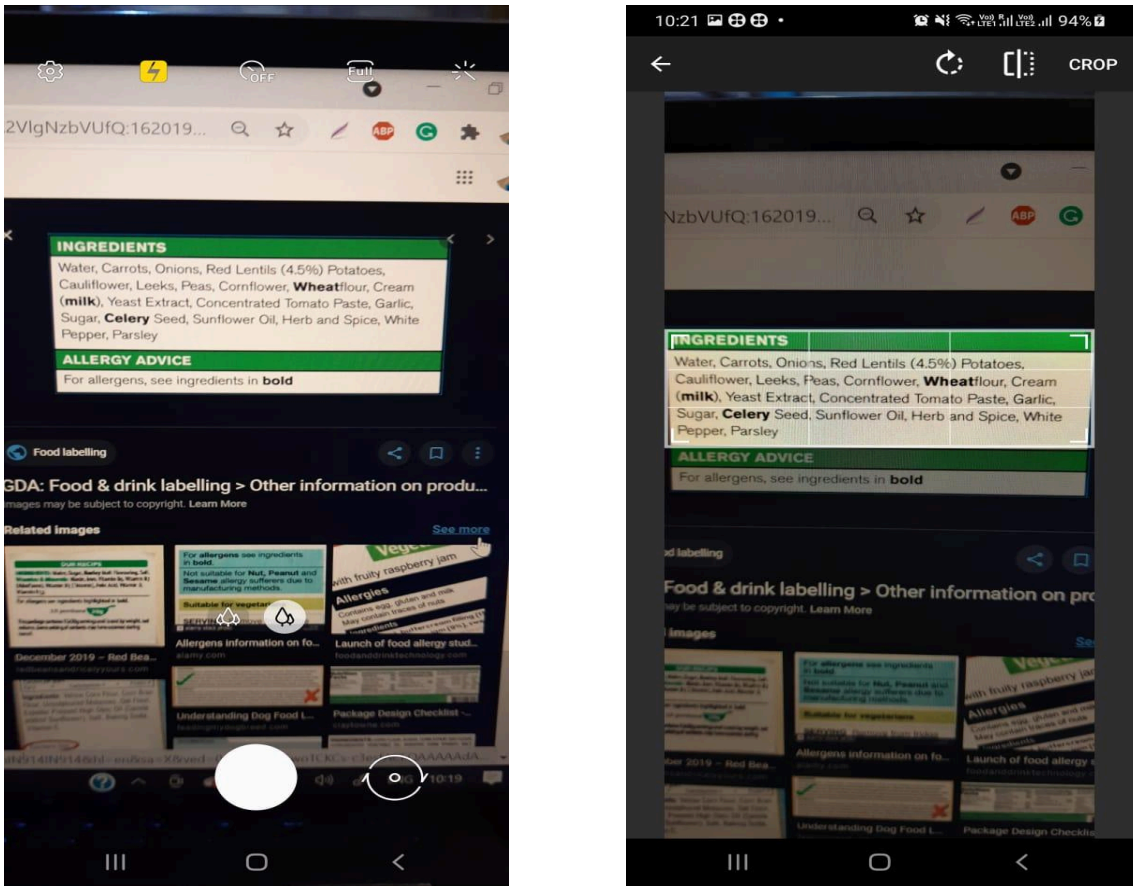


Fig.6 Scanning

Sample Image after text recognition is as follows. This data is then sent for further analysis.

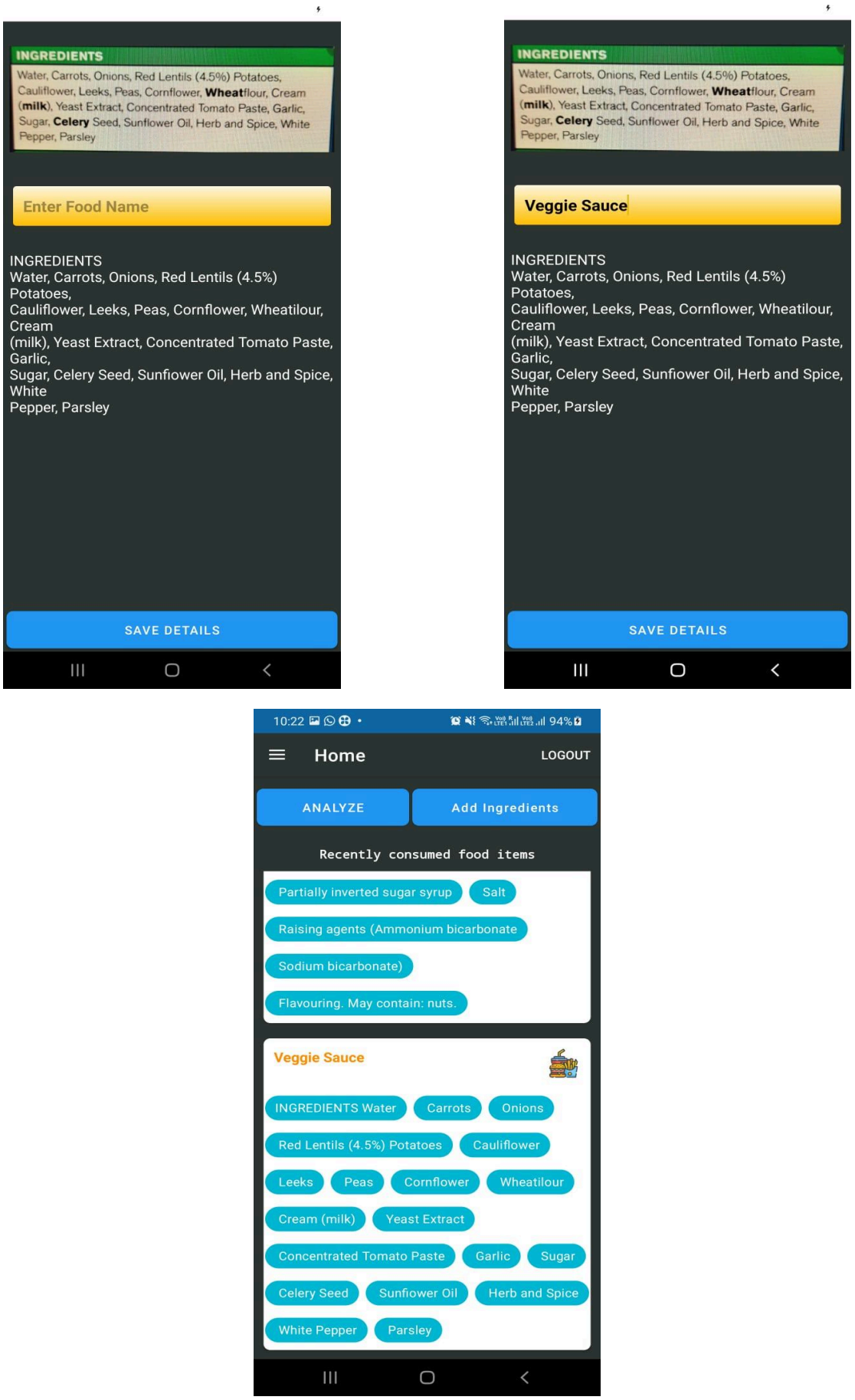


Fig.7 In App Text Extraction

Limitations:

The disadvantage with MLKit is that you might need to upgrade your Firebase plan to a paid one depending on your use.

- If you want a Machine Learning model of your own then a separate custom model should be trained and built with the help of TensorFlow Lite and then should be deployed with the help of MLKit.

There were several factors found to have affected the accuracy of the OCR Engine such as:

- Size of the text
- Font used
- Colour of text with respect to the background
- Any tilt in the text

Since a while there have been a lot of obstacles that keep on arising when it comes to automating the study of packaged food items. The obstacles arise due to different factors like,

- Lack of proper dataset of daily consumption of different age groups.
- Lack of resources to log the daily consumption data
- Lack of technical awareness among the masses about how to use devices in a manner such that data can be logged.
- Consumption of packaged food items randomly, etc

Still, the scope of research in this domain has never declined. Multiple institutions have found their own way to record and analyze consumption data.

6. Data Preprocessing

   Once the data has been logged, a proper record has been established, the data available on Firebase Realtime Database gets to the data preprocessing and algorithm part.

Ingredient names as they appear in the corpora are very noisy, examples of this can be seen in the below example.

Example:

```
100% Corn Flour,Water. This Product Is Produced In A Gluten-Free Facility.
100 Natural Scottish-Style Porridge Oats.
Peanuts And Salt. Contains Peanut Ingredients. Allergy Information: This Product
Is Made On Equipment That Also Makes Products Containing Tree Nuts.
```

Dataset preprocessing was done as follows:

a. Removal of unnecessary sub ingredients. For example: rock-salt, pink salt, black salt all were clustered under "salt", turmeric powder, red chilli powder and other seasonings were clustered under "spices".
b. Removal of redundant ingredients. For example: artificial flavour, food coloring, water, spices and co2 were dropped.
c. Information such as allergy information as shown in the example, deleted.

1. Adjectives and other words
   a. Deleted adjectives such as “exceptionally high quality”
   b. Keep adjectives that indicate food processing, e.g. “bleached white flour” and “roasted almonds”, “dried berries”.
   c. Keep “whole wheat” and “blue corn” because they are types that could have different ingredient implications compared to “wheat” and “corn”. Deleted “whole” in “whole strawberries” because this type doesn’t appear to change the ingredient list.
   d. Kept “organic”.
   e. Deleted vitamin ingredients and products
2. Deleted percentages in the ingredient list, e.g. 80% juice
3. Any “may contain” or “added” ingredients are treated as regular ingredients and parsed the same as the others.
4. Kept ambiguous ingredient references, e.g. “other flavors”, “various spices”, “spice mix”.
5. Deleted “Reverse Osmosis water”
6. Deleted non-food items, e.g. The One-time-use Pack! Wipe 70gx5 Bags
7. Kept clarifying terms, e.g. Hibiscus Flowers (Hibiscus Sabdariffa)
8. Ensured product name references the food. E.g. “Hershey Scare And Share” is a bag of assorted candy. Add clarifying terms to the product name if necessary.

9. Normalized terms, e.g. “popping corn” and “popcorn”, “oat” and “oats”
10. Unreadable characters were deleted, e.g. Waterï¿½ï¿½ï¿½Spring Water.Sub Bunï¿½ï¿½ï¿½
11. Fixed typos
12. All the ingredients were split from one cell to multiple cells. (one ingredient per cell)

Undecided Terms:

Example: *Gluten Free and Gluten*

a. If we keep these terms, should these terms have their own column?
b. Some products have gluten but don’t say so, and some are gluten free but don’t say so. This could skew results if they are in the same visualization with all other products. Perhaps we could special handle them or drop them?

7. Data for Analysis

   The most frequent ingredients consumed by the user were tracked and the top twenty most frequently consumed ingredients were displayed.
   a. The ingredients such as salt, sugar and oil were necessary as well as important were analysed separately, as they will give redundant set values.
   b. For mining algorithms, all the sets with less than two ingredients were dropped
   c. Length: 500

   The subset or frequencies of combination of food ingredients will be a key to analyze the data using suitable algorithms. There are several mining algorithms of association rules. Mining of frequent itemsets is an important phase in association mining which discovers frequent itemsets in transactions databases.In our project we have analysed and compared the time complexity of various mining algorithms.

1. Apriori Algorithm

One of the most popular mining algorithms is Apriori Algorithm.The algorithm [2] makes many searches in the database to find frequent itemsets where k itemsets are used to

generate k+1-itemsets. Each k-itemset must be greater than or equal to the minimum support threshold to be frequency. Otherwise, it is called candidate itemsets. In the first, the algorithm scans the database to find frequency of 1-itemsets that contain only one item by counting each item in the database.

Limitations:

Apriori algorithm has many disadvantages in spite of being simple and easy to implement. The main limitation is costly wasting of time to hold a vast number of candidate sets with much frequent itemsets, low minimum support or large itemsets.

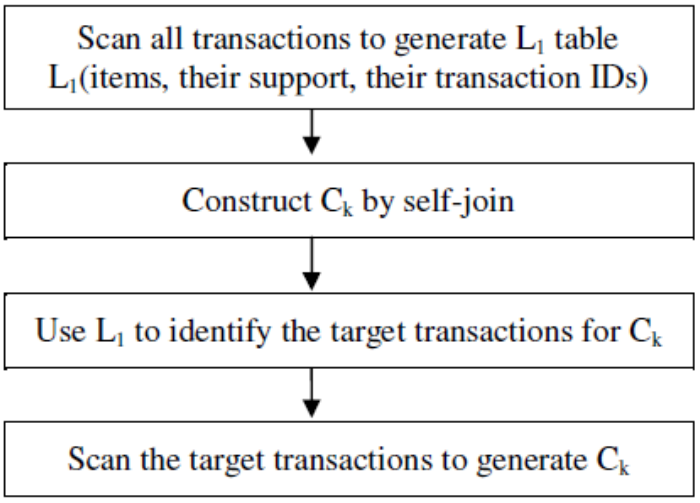
2. Improved Apriori Algorithm

We tried improving Apriori Algorithm and the improved algorithm is as follows:

Definition 1: Suppose T={T1, T2, … , Tm},(m1) is a set of transactions, Ti= {I1, I2, … , In},(n1) is the set of items, and k-itemset = {i1, i2, … , ik},(k1) is also the set of k items, and k-itemset ⊆ I.

Definition 2: Suppose (itemset), is the support count of itemset or the frequency of occurrence of an itemset in transactions.

Definition 3: Suppose Ck is the candidate itemset of size k, and Lk is the frequent itemset of size k.



- The improvement of algorithm can be described as follows:

//Generate items, items support, their transaction ID

(1) L1 = find_frequent_1_itemsets (T);

(2) For (k = 2; Lk-1 ; k++) {

//Generate the Ck from the LK-1

(3) Ck = candidates generated from Lk-1;

//get the item Iw with minimum support in Ck using L1, (1wk).

(4) x = Get _item_min_sup(Ck, L1);

// get the target transaction IDs that contain item x.

(5) Tgt = get_Transaction_ID(x);

(6) For each transaction t in Tgt Do

(7) Increment the count of all items in Ck that are found in Tgt;

(8) Lk= items in Ck min_support;

(9) End;

(10) }

In the prune step, we delete all itemsets c E ck such that some (k-1)-subset of c is not in Lk-1:
An example of the improved Apriori:

Limitations:

Apriori Algorithms can be slow. The main limitation is the time required to hold a vast number of candidate sets with many frequent itemsets, low minimum support, or large itemsets i.e. it is not an efficient approach for a large number of datasets. In our project we've also compared two popular mining algorithms for a large number of dataset.

3. ECLAT Algorithm

Eclat is a vertical database layout algorithm used for mining frequent itemsets. It is based on a depth first search algorithm. In the first step the data is represented in a bit matrix form. For example:

Transaction Id	Bread	Butter	Milk	Coke	Jam
T1	1	1	0	0	1
T2	0	1	0	1	0
T3	0	1	1	0	0
T4	1	1	0	1	0
T5	1	0	1	0	0
T6	0	1	1	0	0
T7	1	0	1	0	0
T8	1	1	1	0	1
T9	1	1	1	0	0

Table.1 Transaction Items for ECLAT algorithm

For the above itemset we conclude the following sets:

ITEMS BOUGHT	RECOMMENDED PRODUCTS
Bread	Butter
Bread	Milk
Bread	Jam
Butter	Milk
Butter	Coke
Butter	Jam
Bread and Butter	Milk
Bread and Butter	Jam

Table:2 Rules for ECLAT algorithm

4. FP-growth Algorithm

Frequent pattern growth also labeled as FP growth is a tree based algorithm to mine frequent patterns in the database the idea was given by (han et. al. 2000) . It is applicable to projected type databases. It uses the divide and conquer method. For example:

Transaction ID	Items
T1	{E,K,M,N,O,Y}
T2	{D,E,K,N,O,Y}
T3	{A,E,K,M}
T4	{C,K,M,U,Y}
T5	{C,E,I,K,O,O}

Table. 3 Transaction sets Fp Growth algorithm

For the above itemsets following is the conditional pattern base.

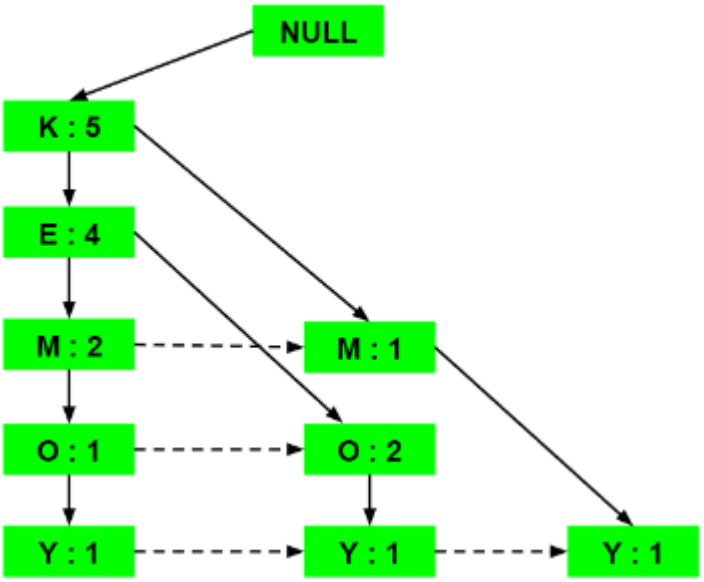


Fig. 8 Rules for FP GROWTH algorithm

Limitations:

FP-Growth algorithm in mining needs two times to scan the database, which reduces the efficiency of the algorithm.

9. Output

The time complexities of these algorithms are compared and the most efficient algorithm is used for displaying the most frequent set consumed to the user in the form of pie charts and line plots.

The chart can help the user to choose the food item according to his/her requirement. After the data analysis, all the results are shown to the user in the form of bar graphs and pie charts. Since the data analysis part also uses previous data, the graphs help the user to assume the after effects of consuming a food product. Consuming food items having the same type of ingredients or nutrient composition will increase the level of that particular nutrient in the graph. A popup notification will also alert the user regarding the matter. This will let the users, who are focused on their health condition, to choose the next food item wisely.

If the user consumes the chosen food product, the respective user data will be updated according to the latest analysis. The user data will include the level of respective ingredients or nutrients. This new data will be used for the next analysis to measure the resultant level after the next consumption process and make suggestions according to it.

And if the user chooses not to consume the food product the session will terminate and the next session will start from the beginning by clicking images or scanning the food item.

10. Ingredient Dataset

Since we are working on packaged food items, we had a dataset that consisted of names of 10 thousand packaged food items along with their respective brands, categories, ingredients contained and their default weight.

From this dataset, we made a program and obtained a dataset of ingredients and the food items that contained them.

```
From this: (10k food items)

        Food1 |  Ingredient1, I2, I3, I4,.................

        Food2 |  Ingredient1, I2, I3, I4,.................
        .
        .
        .
        .
```

To this: (target)

Ingredient1 | Food1, F2, F3,.....

Ingredient2 | Food1, F2, F3,.....

.

.

.

.

For this first from the 10k food items dataset the ingredient column was tokenized using NLTK library and cleaned. A list of useless words was formed that does not contribute to any food item or ingredient in particular. These words along with recurring words were filtered out from the tokenized data.

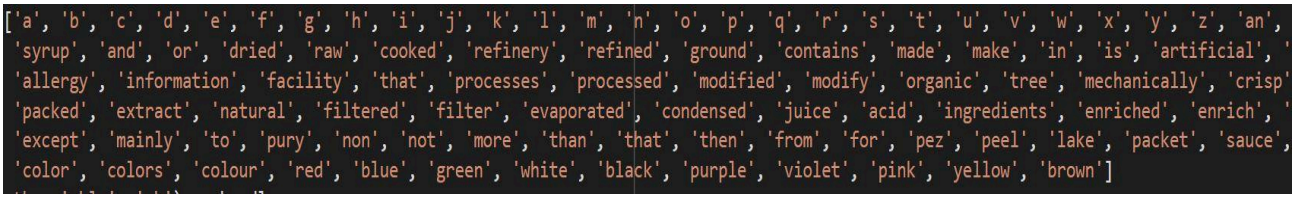


Fig. 9 Useless words

Fig. 10 Tokenized ingredients

Then a list of distinct ingredients was formed that contained every non repetitive and distinct ingredient throughout the 10k food item dataset.

Then from this list of distinct words, every word was looped from the ingredient column of the 10k food items dataset and all the food items that contained that particular ingredient was saved alongside it in an excel sheet and thus our new dataset was formed.

	A	B	C	D	E	F	G	H	I	J	K	L
1	ingredient	food_items										
2	prunes	['Simon Fischer Fruit Bttr Prune Lekvar', 'Flora Dried Fruit and Nut Gift Tray', 'Golden State Fruit Pacific Coast Classic Dried										
3	water	['Simon Fischer Fruit Bttr Prune Lekvar', 'McCORMICK GRILL MATES MOLASSES BACON SEASONING 1 x 77g JAR AMERICAN										
4	corn	['Simon Fischer Fruit Bttr Prune Lekvar', 'McCORMICK GRILL MATES MOLASSES BACON SEASONING 1 x 77g JAR AMERICAN										
5	sugar	['Simon Fischer Fruit Bttr Prune Lekvar', 'McCORMICK GRILL MATES MOLASSES BACON SEASONING 1 x 77g JAR AMERICAN										
6	pectin	['Simon Fischer Fruit Bttr Prune Lekvar', 'Hero Fruit Sprd Blk Currant-12 Oz -pack of 8', 'Bionaturae Organic Wild Berry Fruit										
7	salt	['McCORMICK GRILL MATES MOLASSES BACON SEASONING 1 x 77g JAR AMERICAN IMPORT', 'Jolly Time Popcorn', 'Simply										
8	molasses	['McCORMICK GRILL MATES MOLASSES BACON SEASONING 1 x 77g JAR AMERICAN IMPORT', "Erin Baker's Homestyle Gra										
9	caramel	['McCORMICK GRILL MATES MOLASSES BACON SEASONING 1 x 77g JAR AMERICAN IMPORT', 'Simply Asia Noodle Bowl M										
10	spices	['McCORMICK GRILL MATES MOLASSES BACON SEASONING 1 x 77g JAR AMERICAN IMPORT', 'Italian Bread Crumbs', 'Corn										
11	pepper	['McCORMICK GRILL MATES MOLASSES BACON SEASONING 1 x 77g JAR AMERICAN IMPORT', 'Simply Asia Noodle Bowl M										
12	garlic	['McCORMICK GRILL MATES MOLASSES BACON SEASONING 1 x 77g JAR AMERICAN IMPORT', 'Italian Bread Crumbs', "Stel										
13	onion	['McCORMICK GRILL MATES MOLASSES BACON SEASONING 1 x 77g JAR AMERICAN IMPORT', 'Simply Asia Noodle Bowl M										
14	tapioca	['McCORMICK GRILL MATES MOLASSES BACON SEASONING 1 x 77g JAR AMERICAN IMPORT', 'Simply Asia Noodle Bowl M										
15	maltodextrin	['McCORMICK GRILL MATES MOLASSES BACON SEASONING 1 x 77g JAR AMERICAN IMPORT', 'Trolli Sour Brite Eggs, 4.0 O										

Fig. 11 Ingredients to food Dataset

11. Suggestion of packaged food items

Based on the results of the Fp growth algorithm, most frequent items consumed by the users are then used for suggesting other packaged food items of their choice.

Algorithm:

1. Get a list of *highest frequency items*.
2. Initialize a list variable as *mylist.*
3. Loop through each value in the *highest freq items* list.
4. For every ingredient in the *Ingredient food dataset*.
5. If the highest frequency item name is equal to ingredient name.
    a. then append mylist.
6. else
    a. do nothing
7. end if
8. make a set of *mylist* items and print the set items.

12. To generate the dataset consisting of disease prone flags, we have created 2 columns (Ingredient Name and Disease related to the ingredients). Based on the data available in research papers [34, 35] the ingredients were classified into 0 and 1.

Flag “1” was stating that the ingredient is prone to the disease, Flag “0” was stating that the ingredient is safe to consume keeping the same disease in mind.

13. Health tip is generated with the help of the ingredients dataset available with values 0 and 1 for specific disease prone. This tip generation plays a vital role in the development of the scenario as the further implementations are based on this health tip generation.

For example, diseases like Asthma, hypertension, iron deficiency, decrease in liver transaminases are found to be prone with MSG which is available in Corn starch, Corn syrup, Modified food starch, Lipolyzed butter fat, Dextrose, Rice syrup, Brown rice syrup, Milk powder, etc.

Algorithm:

1. Get a list of *highest frequency items*.
2. Initialize a list variable as *mylist.*
3. Loop through each value in the *highest freq items* list.
4. For every ingredient *i* in *the disease dataset.*
5. If the highest frequency item name is equal to ingredient name.
    a. Then print "*Ingredient i* is harmful for your health."
6. else
    a. do nothing
7. end if

The average time complexity of this algorithm is O(n*m). Where n is the number of ingredients in the disease dataset and m is the number of highest frequency items and m < n.

As the number of ingredients are already sorted during preprocessing of the dataset. We tried to make this algorithm more efficient using a binary search algorithm.

14. Binary Search Algorithm

This search algorithm works on the principle of divide and conquer. For this algorithm to work properly, the data collection should be in the sorted form.Which is why we've used this for our algorithm.

Algorithm:

```
Procedure binary_search
   A ← sorted array
   n ← size of array
   x ← value to be searched

   Set lowerBound = 1
   Set upperBound = n

   while x not found
      if upperBound < lowerBound
         EXIT: x does not exists.

      set midPoint = lowerBound + ( upperBound - lowerBound ) / 2

      if A[midPoint] < x
         set lowerBound = midPoint + 1

      if A[midPoint] > x
         set upperBound = midPoint - 1

      if A[midPoint] = x
         EXIT: x found at location midPoint
   end while

end procedure
```

Improved Algorithm for health tip generation:

1. Get a list of *highest frequency items*.
2. Initialize a list variable as *mylist.*
3. Loop through each value in the *highest freq items* list.
4. Apply *binary search algorithms* for every ingredient in the *highest frequency list.*
5. initialize a variable *val.*
6. If the ingredient *i* is found in the *Disease dataset.*
    a. Then the index of the ingredient is returned.
    b. set val -> index of ingredient returned
7. Else
    a. -1 is returned
    b. set val as -1
8. end if
9. if val is equal to -1
    a. Then print "*Ingredient i* is harmful for your health."
10. else
    a. do nothing

The average time complexity of binary search algorithms is O(logn). Due to which the efficiency of our algorithm is reduced to O(m*logn). Where n is the number of ingredients in the disease dataset and m is the number of highest frequency items and m < n.

15. Categorization

Regulatory frameworks all around the globe have primarily started to realize the target of making certain food safety and protection of shopper interests. Each of these objectives need that regulators analyse the knowledge on numerous food safety and regulatory aspects throughout the organic phenomenon, as well as estimating dietary exposure for closing scientific risk assessments. Keeping in mind the large diversity within the food merchandise being consumed, regions during which they consumed, the population teams concerned etc., it becomes nearly not possible to trace the knowledge on a personal product basis. except the sheer volume of knowledge, the matter gets more combined by the utilization of multiple languages, dialects and regional variations.

The preparation of reliable information on food needs precise words and careful description of foods. Even information of excellent quality is often a supply of error if they're derived from foods that aren't clearly outlined. Moreover, it's troublesome to exchange information on foods, or to grasp and compare numerous parameters like biological process standing, consumption patterns, risk analysis profiles etc. for various regions, states or people, while not a coherent description of foods in databases.

To objectively analyse the relevant information, regulators have used the categorization approach, wherever teams of comparable merchandise area units clubbed along joined class. This class is employed because the basic unit for capturing data and driving higher cognitive processes in regulatory frameworks.

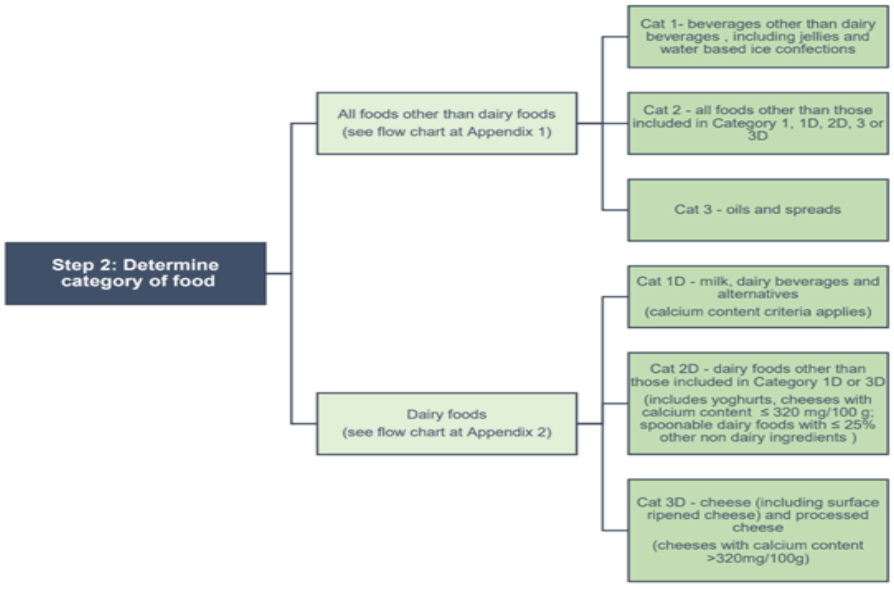


Fig.12 Categories of products in the HSR calculator

16. Health Star Rating

Each of the categories has been assigned a pre-calculated Mean Health Star value which we used in our program for calculating Health Star Rating of the food item.

Major Food Category	No. Products	Nutrient Profiling Summary Score		Level of Processing
		Mean HSR (SD)	Proportion 'Healthy' HSR ≥ 3.5 (%)	Proportion Ultra-Processed (%)
Eggs	487	3.9 ± 0.5	94.3	0.0
Fruit, vegetables, nuts and legumes	38,032	3.7 ± 1.1	70.3	18.3
Seafood and seafood products	4325	3.7 ± 0.8	81.7	20.3
Cereal and grain products	18,024	3.2 ± 1.1	52.8	61.8
Convenience foods	17,980	3.2 ± 0.8	51.9	90.9
Edible oils	3246	3.0 ± 1.3	55.9	8.4
Foods for specific dietary use	2369	2.9 ± 1.1	41.4	100.0
Dairy	27,839	2.9 ± 1.4	40.7	61.5
Snack foods	12,231	2.6 ± 1.2	29.6	100.0
Non-alcoholic beverages	19,954	2.5 ± 1.7	35.7	83.1
Sauces, dressings, spreads, and dips	21,772	2.5 ± 1.2	32.9	92.3
Meat and meat alternatives	12,249	2.2 ± 1.3	35.7	75.3
Bread and bakery products	30,194	2.1 ± 1.1	20.0	98.6
Sugars, honey, and related products	4625	1.5 ± 0.9	6.7	73.0
Confectionery	16,829	1.1 ± 0.7	1.4	100.0
All	230,156	2.7 ± 1.4	40.2	70.9

HSR, Health Star Rating. SD, standard deviation

Table 4 Healthfulness of the US food and beverage supply

Based on the above resources, we in our project categorized the food ingredients into 12 categories:

- Eggs
- Seafood
- Cereals
- Oil products
- Dairy items
- Non-alcoholic Beverages
- Sauces and Dressings
- Meat products
- Bread and Bakery products
- Sugar and Honey items
- Confectionaries

A dataset of ingredients categorized according to the above categories were made using INDIAN FOOD CODE [38]:

```
Eggs = ['eggs', 'egg', 'yolk']

#(fruits, vegetables, nuts, legumes)
FVNL = ['fruits', 'fruit', 'vegetables', 'vegetable', 'vegan', 'nuts', 'nut', 'legumes', 'legume', 'pulse', 'pulses', 'apple', 'apricots', 'avocado', 'b
        'carambola', 'cherimoya', 'cherry','cherries', 'clementine', 'coconut', 'cranberries', 'date', 'durian', 'elderberries', 'feijoa', 'figs', 'goos
        'plum', 'jujube', 'kiwifruit', 'kumquat', 'lemon', 'lime', 'longan', 'loquat', 'lychee', 'mandarin', 'mango', 'mangosteen', 'mulberries', 'necta
        'pear', 'persimmon', 'pitaya', 'pineapple', 'pitanga', 'plantain', 'plums', 'pomegranate', 'prunes', 'pummelo', 'quince', 'raspberries', 'rhubar
        'tamarind', 'tangerine', 'watermelon', 'tomato', 'onion', 'garlic', 'ginger','cerely', 'parsley', 'corriander', 'seed', 'seeds', 'aniseed', 'pis

Seafood = ['seafood', 'seafoods', 'fish', 'cod', 'crab', 'prawns']

Cereal = ['wheat', 'barley', 'rice', 'oats', 'oat', 'bran', 'corn']

Oil = ['oils', 'oil', 'mustard']

Dairy = ['dairy', 'milk', 'cheese', 'curd', 'yogurt','cream']

NABeverage = ['juice', 'juices', 'tea', 'coffee', 'water', 'lemonade', 'cola']

Sauce = ['sauce', 'sauces', 'dressing', 'spreads', 'dips', 'syrup', 'ketchup', 'chutney', 'mayonnaise', 'salt', 'vinegar', 'spice', 'spices', 'herbs']

Meat = ['meat', 'chicken', 'lamb', 'goat', 'beef', 'pork', 'turkey']

Bread = ['bread', 'breads', 'cake', 'pie', 'biscuits', 'cookies', 'pastry', 'muffin']

Sugar = ['sugar', 'cane' , 'honey', 'maltose', 'glucose', 'fructose', 'maple']

Confectionery = ['sweets', 'chocolate', 'chocolates', 'pastilles', 'vanilla', 'candy', 'fudge']
```

Fig. 13 Categories of Ingredients

Once the dataset was made, the food items suggested previously in the program, according to the ingredients, were searched throughout the 10k food item dataset and the corresponding list of ingredients were fetched.

Each of these ingredients were checked for the categories they lie into. If they were found to be lying in any category the respective MHS value of that category were added to the rating and an average rating was calculated.

This average rating calculated was the desired Health Star Rating of the respective food item of whom the ingredients were. The food item was further grouped in “Low” (0-2 HSR), “Medium” (2-3 HSR) and “High” (3<) according to the HSR value and was suggested to the user to consume accordingly.

```
['BACK TO NATURE CEREAL ORANGE CRUNCH GRAN', '2.73', 'Medium']
["Erin Baker's Homestyle Granola", '2.50', 'Medium']
['Russell Stover Assorted Fine Chocolates, 7.0 Oz', '2.70', 'Medium']
['Honey Mustard', '2.73', 'Medium']
['Iced Tea - Unsweetened Brewed', '2.00', 'Medium']
["Herr's Real Chocolate Covered Pretzel Sticks, 5 Oz", '2.57', 'Medium']
['Russell Stover Assorted Fine Chocolates, 34.0 Oz', '2.70', 'Medium']
["Nature's Path Organic Heritage Crunch Cereal, 14.0 Oz", '2.50', 'Medium']
['Quaker Real Medleys Fruit & Nut Multigrain Bars, Cherry Pistachio, 1.34 Oz Bar, 10/box -qkr31799', '2.73', 'Medium']
['Bakery On Main Gluten Free Non-gmo Nut Crunch Snack', '2.54', 'Medium']
```

Fig. 14 Output after HSR rating assignment

5. Testing

For testing we used Sample Packaged food item dataset which contains 32 processed food items and their ingredients.

First, the frequency of all the food items was calculated and top 20 most frequent items were displayed as follows:

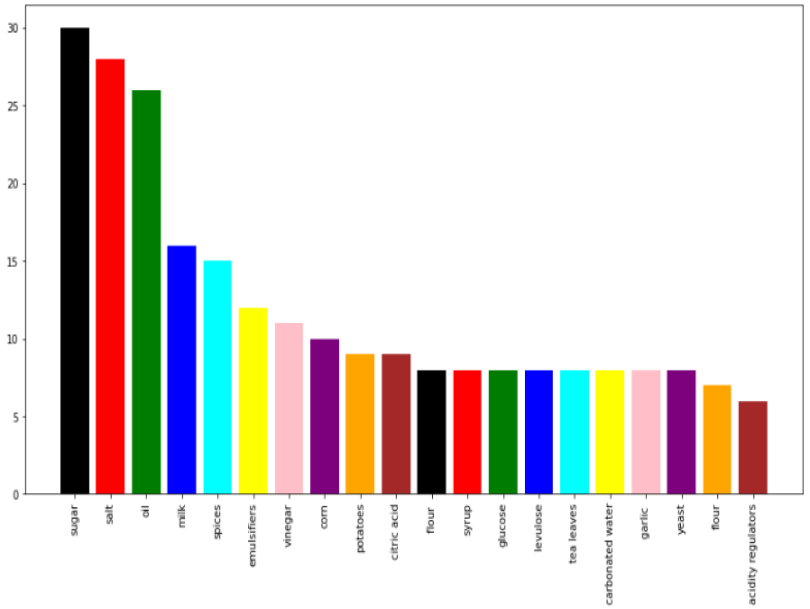


Fig. 15 Frequent ingredients consumed

As we can see the most frequent and necessary ingredients are salt, sugar and oil. These affect the health of adolescents the most. So we will be tracking these separately.

All the redundant items such as spices were also dropped.

The time taken to calculate frequent items was **0.010021 secs.**

To analyse the most frequent sets of ingredients we used the different mining Algorithms on our sample dataset. The result is as follows:

1. **Apriori Algorithm:**

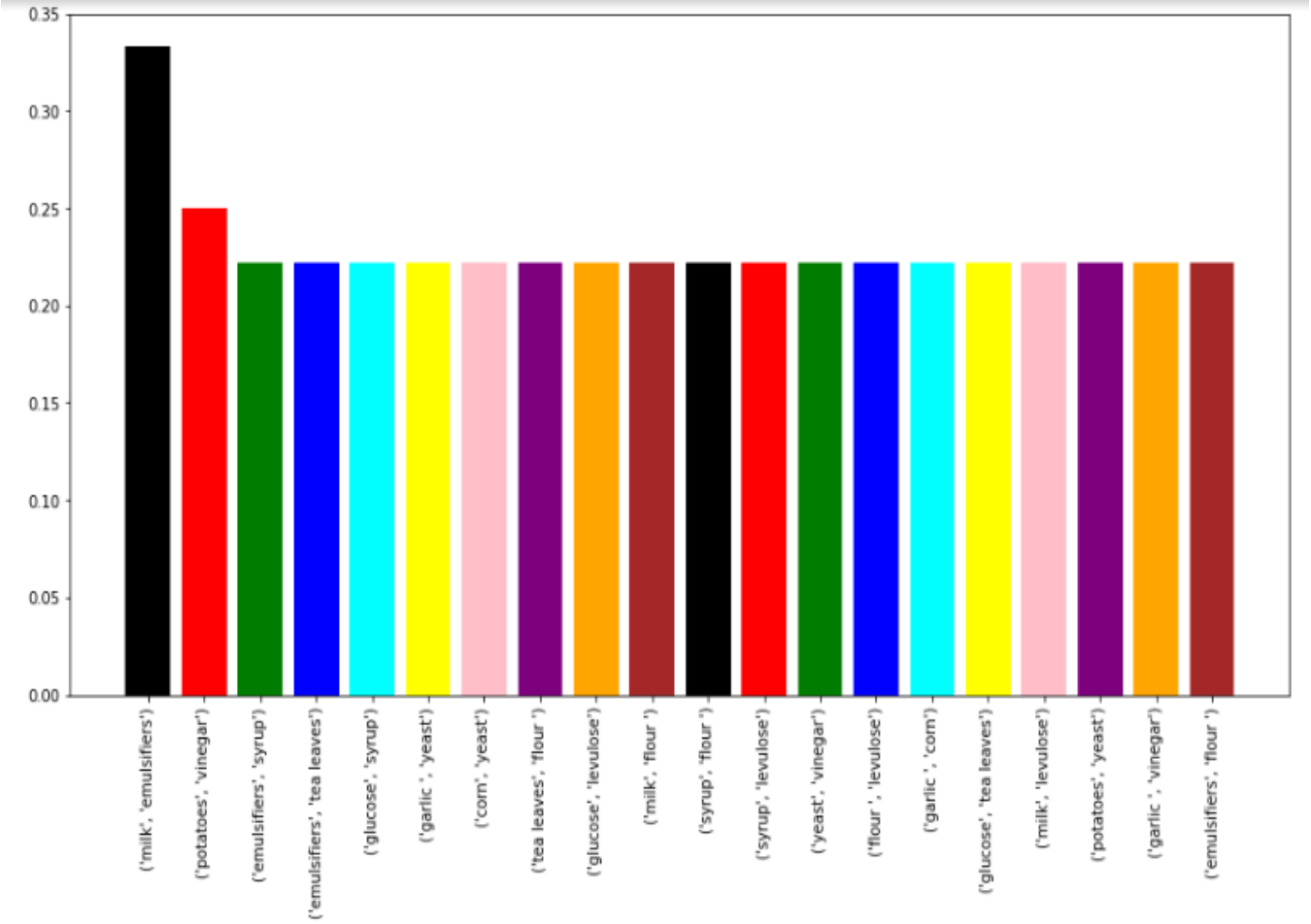


Fig. 16 Frequent sets using Apriori algorithm

Minimum Support: 20%

Confidence: 80%

Minimum items in set: 2

Total Items found: 222

Time Taken: 0.02108 secs

2. Improved Apriori Algorithm

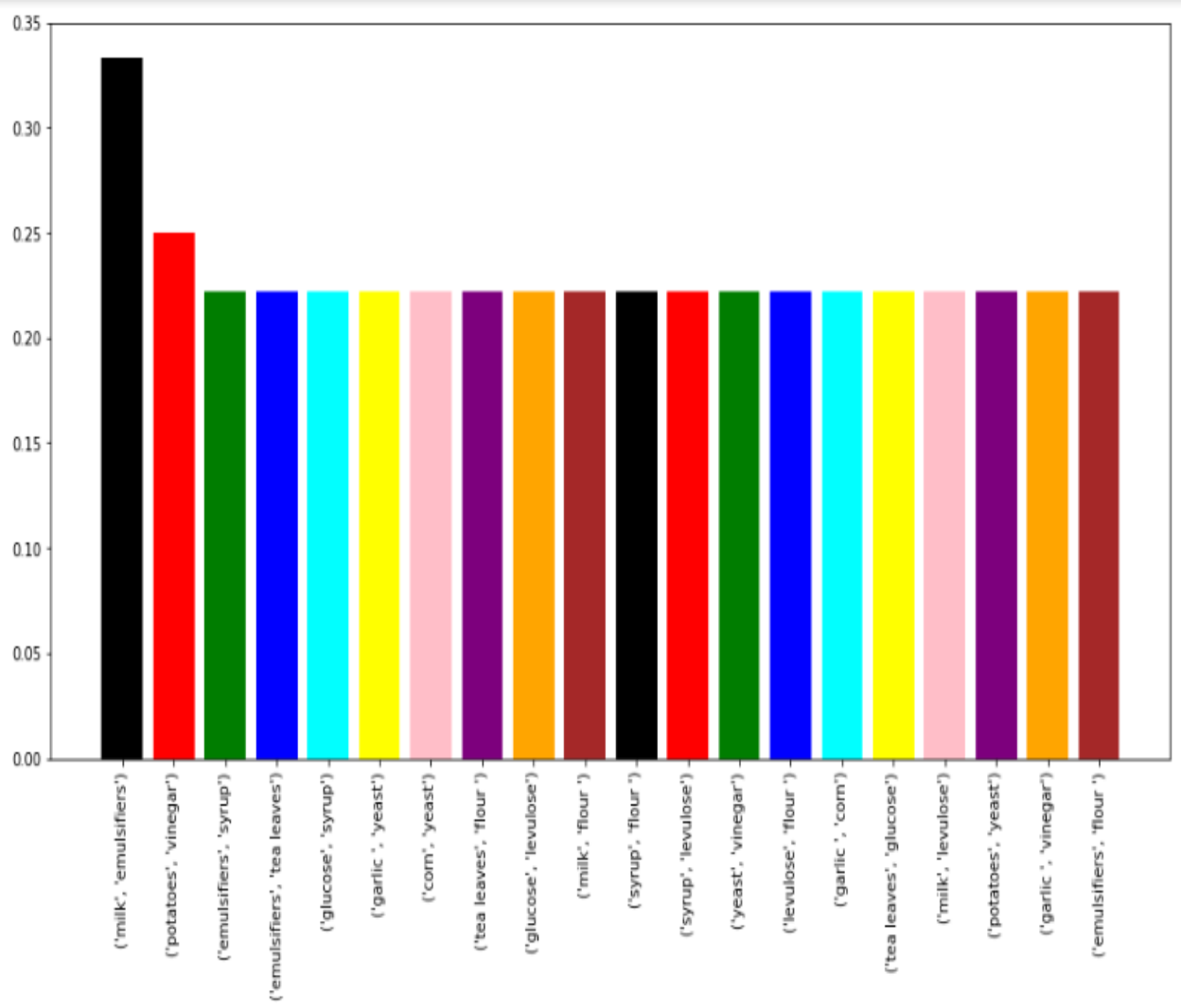


Fig. 17 Frequent sets using Improved Apriori algorithm

Minimum Support: 20%

Confidence: 80%

Minimum items in set: 2

Total Items found: 162

Time Taken: 0.00672 secs

From the above result we can see that the time complexity was reduced after pruning. Hence, we can infer that the Apriori Algorithm can be improved.

3. ECLAT Algorithm

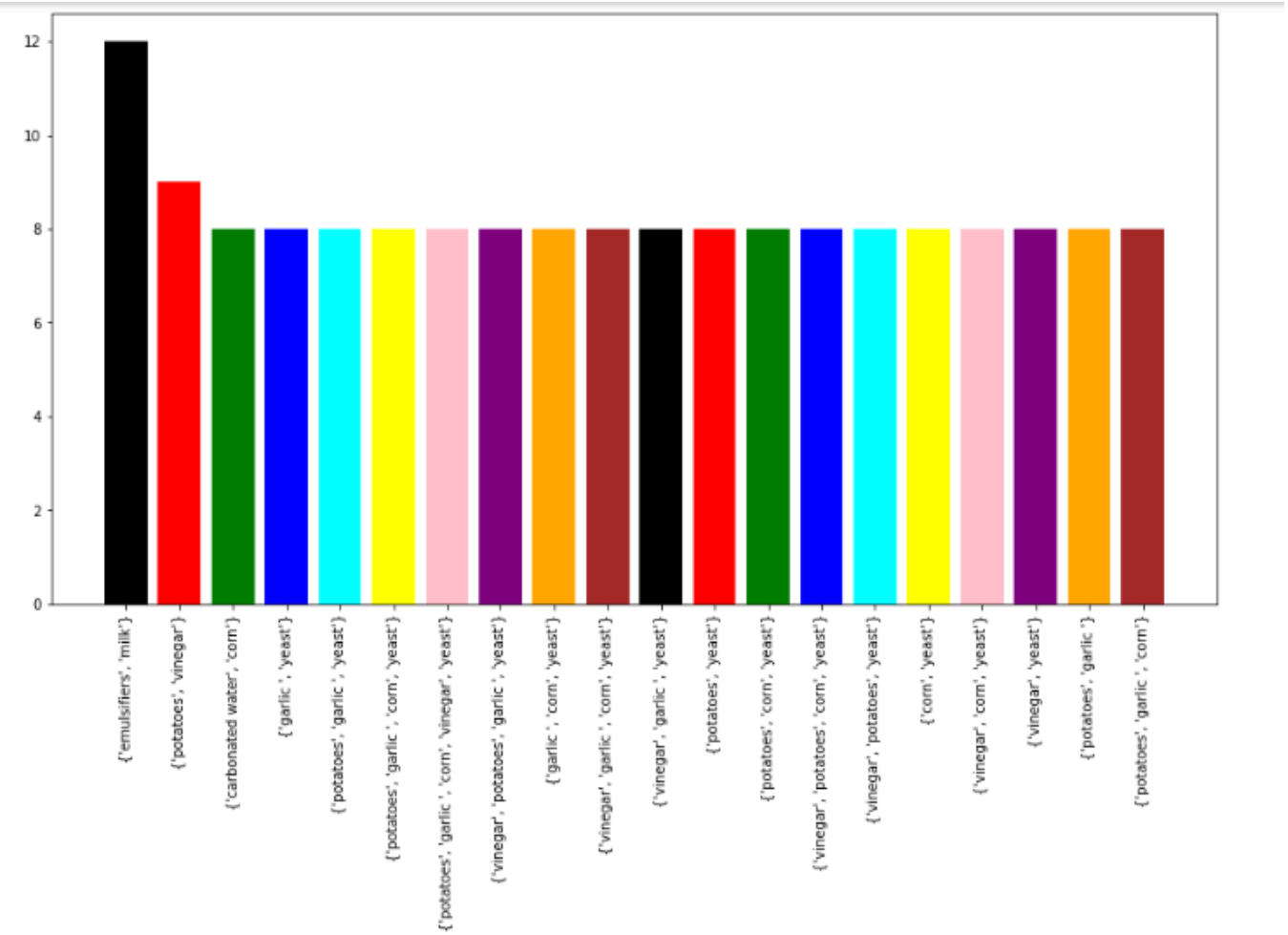


Fig. 18 Frequent sets using Eclat algorithm

Minimum Support: 20%

Confidence: 80%

Minimum items in set: 2

Total Items found: 192

Time Taken: 0.00069 seconds

From the above result we can see that the time complexity was reduced. Hence, we can infer that the ECLAT is better than Apriori.

4. FP - Growth Algorithm

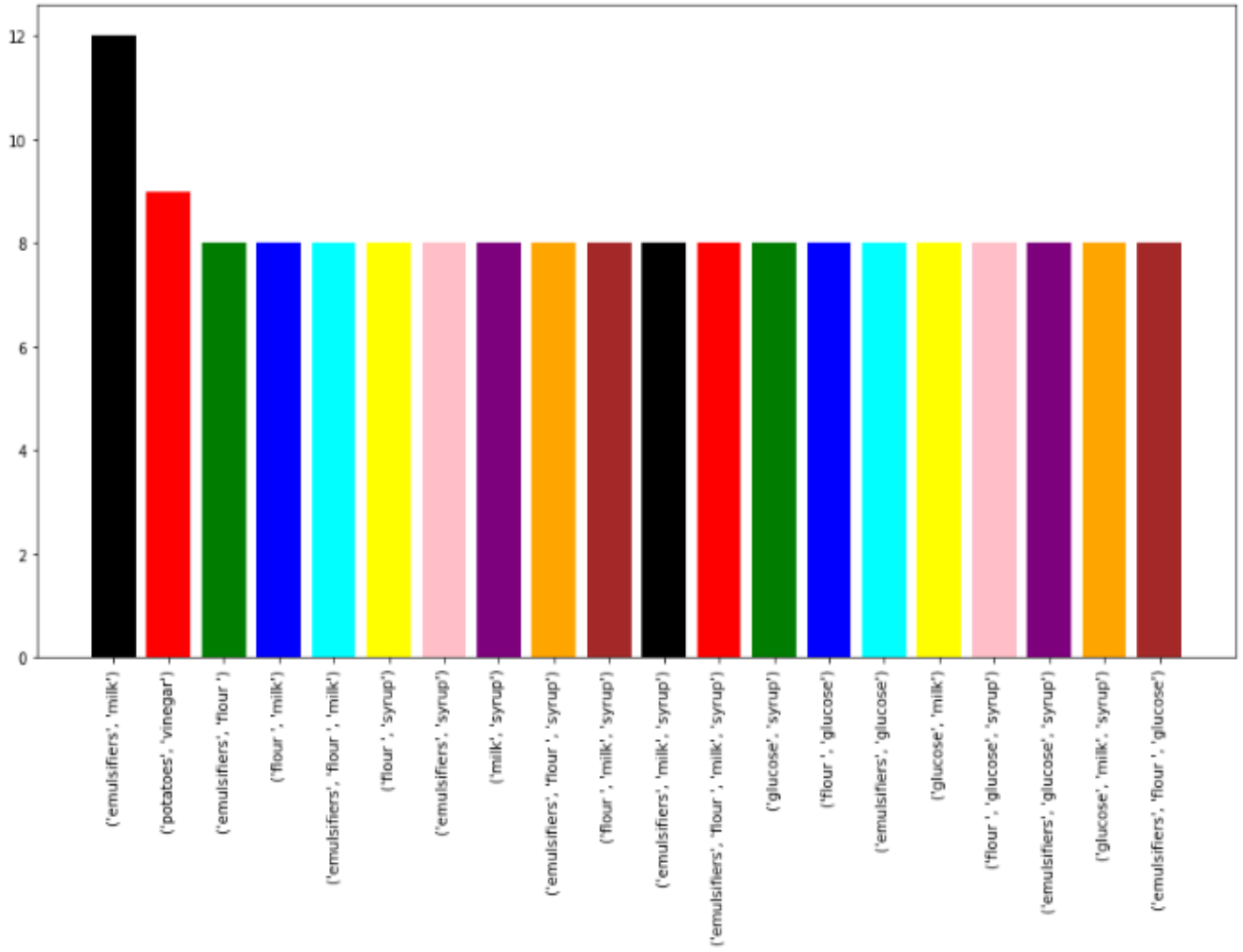


Fig. 19 Frequent sets using FP-GROWTH algorithms

Minimum Support: 20%

Confidence: 80%

Minimum items in set: 2

Total Items found: 168

Time Taken: 0.00174 seconds

Now, to find out the most efficient algorithm for large datasets we tested all these algorithms on various support values and the results were as follows:

confidence	support	Time Apriori	Time Improved Apriori	Time eclat	Time FP Growth	Sets Apriori	Sets Improved Apriori	Sets eclat	Sets FP Growth
0.8	0.3	0.01496	0.00137	0.00013	0.00056	76	4	5	5
0.8	0.25	0.01648	0.00283	0.00062	0.00151	77	8	161	161
0.8	0.2	0.02322	0.01157	0.00121	0.00147	222	162	192	168
0.8	0.15	0.03343	0.02323	0.00205	0.00374	315	257	659	649
0.8	0.14	0.09994	0.02157	0.00241	0.00685	315	257	659	649
0.8	0.13	0.19816	0.17187	0.00242	0.00868	712	661	659	649
0.8	0.12	0.19813	0.21580	0.12624	0.09789	712	661	33421	33414
0.8	0.115	0.24003	0.23559	0.12701	0.08782	712	661	33421	33414
0.8	0.113	0.24946	0.22691	0.13708	0.08814	712	661	33421	33414
0.8	0.112	>700	644	0.28931	0.09253	-	661	33421	33414
0.8	0.111	>700	680	0.25214	0.22767	-	66189	33421	33414
0.8	0.108	>700	688	0.22429	0.22121	-	66189	33421	33414
0.8	0.105	>700	689	0.21135	0.16961	-	66189	33421	33414
0.8	0.1	>700	690	0.22862	0.23438	-	66189	33421	33414

Table. 5 Performance Analysis of Algorithms with changing support values

From the above table, we can infer that the Eclat algorithm and frequency growth pattern algorithm is better than both Apriori and the improved version of Apriori Algorithm.
Also, Fp growth works better on large datasets and Eclat works better on small datasets.

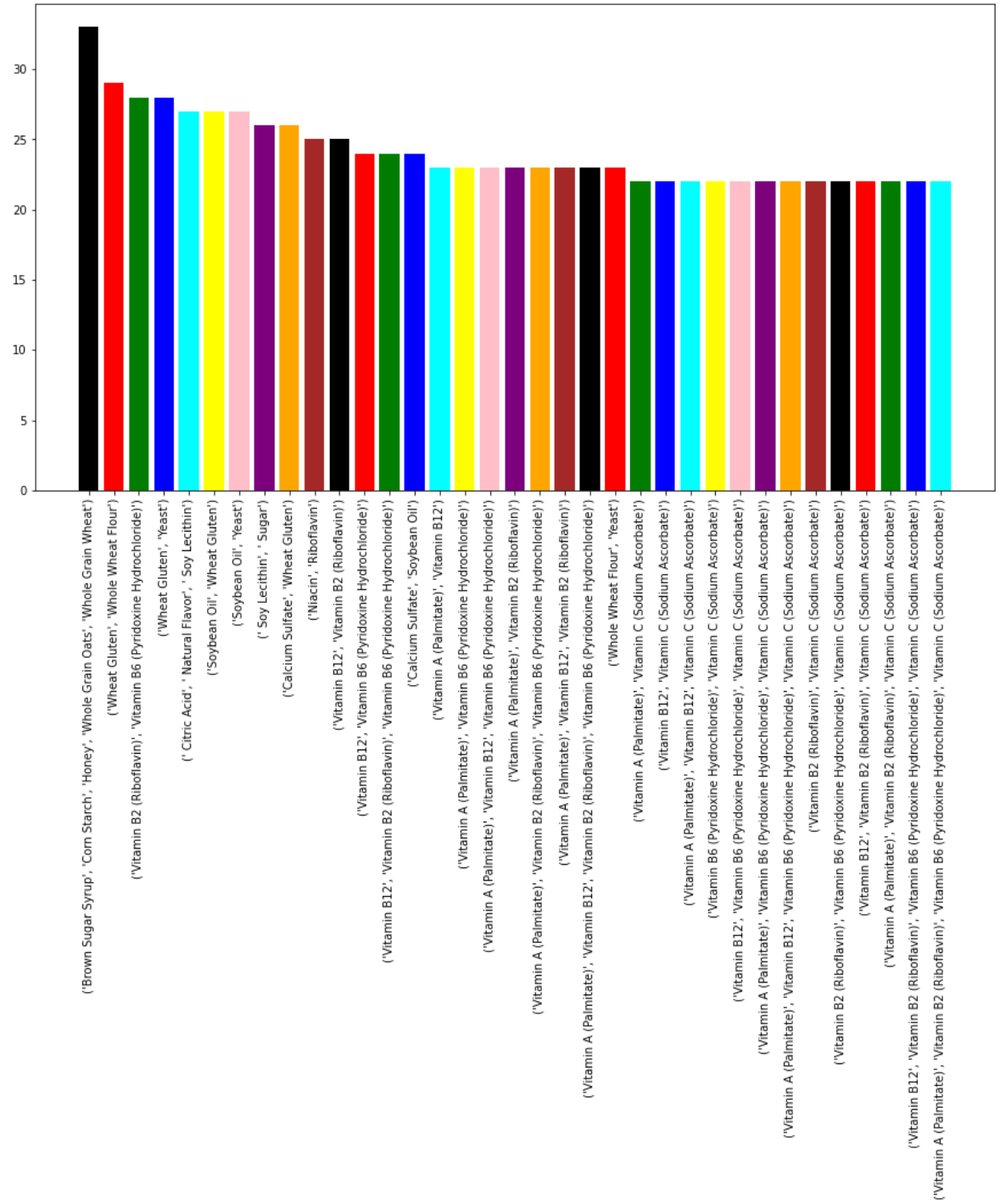


Fig. 20 Applying Fp growth on preprocessed dataset

Minimum Support: 20%

Confidence: 80%

Minimum items in set: 2

Total Items found: 8490

Time Taken: 0.12604 seconds

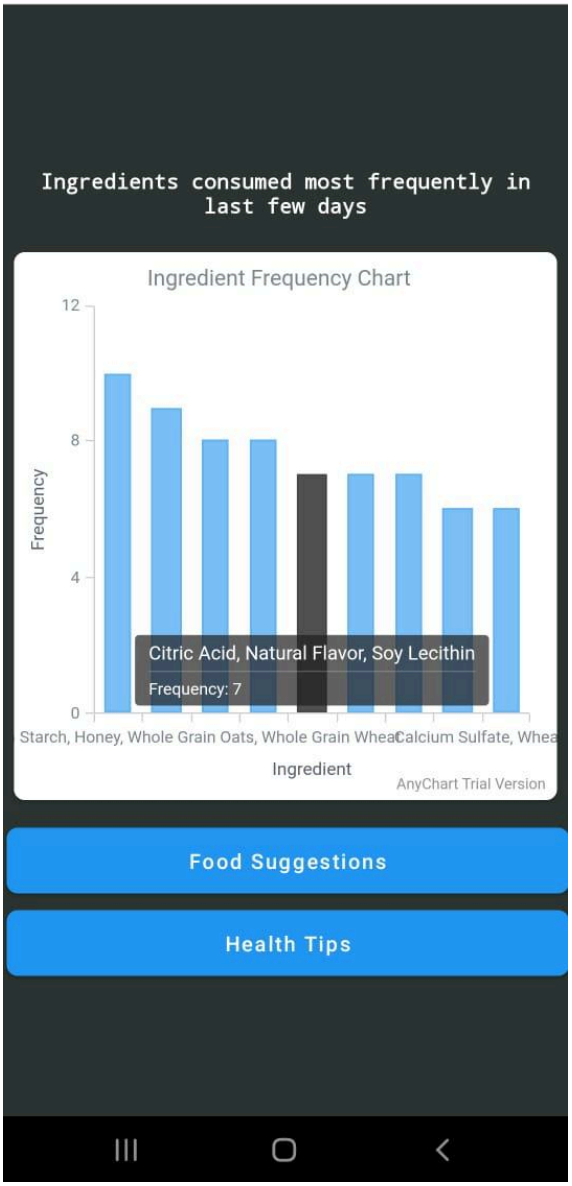


Fig. 21 Output of Fp Growth

```
['BACK TO NATURE CEREAL ORANGE CRUNCH GRAN', '2.73', 'Medium']
["Erin Baker's Homestyle Granola", '2.50', 'Medium']
['Russell Stover Assorted Fine Chocolates, 7.0 Oz', '2.70', 'Medium']
['Honey Mustard', '2.73', 'Medium']
['Iced Tea - Unsweetened Brewed', '2.00', 'Medium']
["Herr's Real Chocolate Covered Pretzel Sticks, 5 Oz", '2.57', 'Medium']
['Russell Stover Assorted Fine Chocolates, 34.0 Oz', '2.70', 'Medium']
["Nature's Path Organic Heritage Crunch Cereal, 14.0 Oz", '2.50', 'Medium']
['Quaker Real Medleys Fruit & Nut Multigrain Bars, Cherry Pistachio, 1.34 Oz Bar, 10/box -qkr31799', '2.73', 'Medium']
['Bakery On Main Gluten Free Non-gmo Nut Crunch Snack', '2.54', 'Medium']


[['BACK TO NATURE CEREAL ORANGE CRUNCH GRAN', '2.73', 'Medium'], ["Erin Baker's Homestyle Granola", '2.50', 'Medium'],
```

Fig. 22 HSR rating assignment

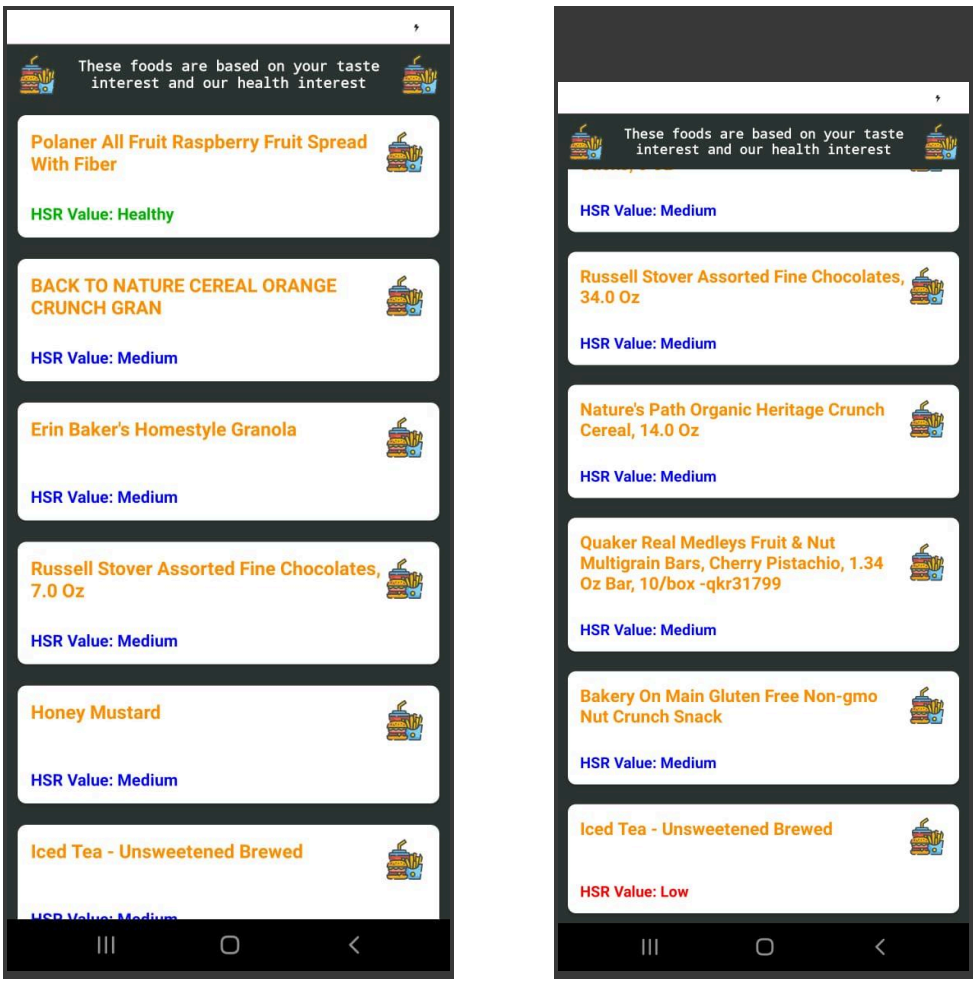


Fig. 23 Output screenshots - Food suggestion

Result after health tip generation:

```
for i in highest_freq:
  val=binarySearch(list(diseases['ingredients']),i)
  if val!=-1:
    print("{} is harmful for your health, it may cause D1 ".format(i))
end=time.time()
print("TIME COMPLEXITY IMPROVED: {}".format(end-start))
#TIME COMPLEXITY IS M*logn
```

```
Corn Starch is harmful for your health, it may cause D1
TIME COMPLEXITY NORMAL: 0.002455472946166992
Corn Starch is harmful for your health, it may cause D1
TIME COMPLEXITY IMPROVED: 0.0005736351013183594
```

Fig. 24 Health Tip Generation

As we can see from the above figure, using a binary search algorithm, the time complexity was initially 0.00245 sec, after improvising the algorithm we got the time complexity of 0.00057 sec.

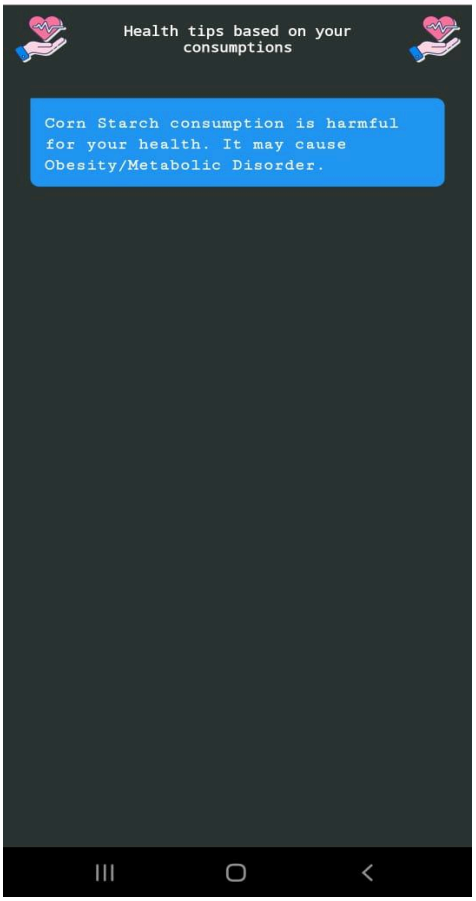


Fig. 25 Output screenshot - Health Tip

6. Conclusion, and Future Work

6.1 Conclusion

After the current study and implementation, we came to a conclusion that there can be different sets of combinations which can be analyzed. There can be multiple patterns which can help us track patterns of the consumption.

Algorithms used to study the scenario played a major role and was the key part of our research.
We compared the time consumption of all the three Algorithms.
Apriori, ECLAT and FP Growth Algorithms gave significantly better understanding about the dataset. The time complexity plays a major role when it comes to analyzing big chunks of data and therefore the results we got along with all the Algorithms were satisfying upto the expectations.
We implemented an improved version of apriori Algorithm as well which helped us to reduce the time consumption and complexity with a good margin.

In this project we surveyed the pattern mining algorithms namely apriori, Eclat and FP Growth on ingredients. It is found that apriori uses join and prune method, Eclat works on vertical datasets and FP Growth constructs the conditional frequent pattern tree which satisfies the minimum support.
The major weakness of Apriori algorithm is producing a large number of candidate itemsets and large number of database scans which is equal to maximum length of frequent itemset . It is very much expensive to scan large databases. A true reason of apriori failure is it lacks an efficient processing method on the database. FP Growth is the best among the three algorithms and is thus most scalable. Eclat performs poorer than FP Growth and Apriori performs the worst.

Our research helped us to authenticate that study on packaged food items is possible if we utilize the resources effectively. Algorithms can help us to tackle the time complexity when it comes to big amounts of data as well as user base.

6.2 Future Work

In this application, the program only considers ingredients as the inputs and totally ignores their share or weightage in the packed food item. This is a significant issue that can change the values and graphs generated as output. Which a user uses to plan his/her future diet.

We have decided to keep this topic as a future reference as much research work is required to include the weightage as inputs. Also planning and designing a program to calculate the output is also to be researched on.

Also, we can extend this application to calculate other nutrition facts such as calories, fat, carbohydrates, proteins, vitamins etc. For this, we might need to explore more on the data collection part. And, we can use better Natural Language Processing techniques in the extraction phase for better results but might take more time and a big dataset for training.

In a major leap, this app can be extended to generate diet plans specific to user requirements and necessities to balance the excessive and inadequate nutrients. Or suggest the best diet plan according to the region specific to the user and for those users who are going through some health issues that require specific quantities of certain types of nutrients. A chatbot can also be included to help the user regarding some problems for searching or planning and a connectivity with other users can also be formed to create a social media environment for further application related communication.

Also, in the future improvements, different algorithms can be used to enhance the efficiency and accuracy. Some words that only contribute to enhancing the quality or share of some ingredients can be used to add some meaning and increase the accuracy and specifications for a better food suggestion to the user.

As of now, the health tip is generated based on comparison and searching the most frequent item dataset. We can develop a machine learning model which will help in generating the tips with the help of binary classification problems.

References

[1] Connie M Weaver, Johanna Dwyer, Victor L Fulgoni III, Janet C King, Gilbert A Leveille, Ruth S MacDonald, Jose Ordovas, and David Schnakenberg. “PROCESSED FOODS: CONTRIBUTIONS TO NUTRITION”. Am J Clin Nutr 2014; 99:1525–42.

[2] Swarna Sadasivam Vepa, Programme Director and Ford Foundation Chair for Women and Sustainable Food Security. “IMPACT OF GLOBALIZATION ON THE FOOD CONSUMPTION OF URBAN INDIA”. M.S. Swaminathan Research Foundation, Chennai.

[3] Elizabeth K Dunford, Jacqueline L Webster, Rama K Guggilla, Pallab K Maulik, Bruce C Neal. “THE ADHERENCE OF PACKAGED FOOD PRODUCTS IN HYDERABAD, INDIA WITH NUTRITIONAL LABELLING GUIDELINES”. Asia Pac J Clin Nutr 2015;24(3):540-545.

[4] Nida I Shaikh, Shailaja S Patil, Shiva Halli, Usha Ramakrishnan and Solveig A Cunningham. “GOING GLOBAL: INDIAN ADOLESCENTS’ EATING PATTERNS”. Public Health Nutrition: 19(15), 2799–2807.

[5] Subhalakshmi K., Dhanasekar M.. “A STUDY ON FAST FOOD CULTURE AMONG TEENAGERS IN URBAN INDIA”. International Journal of Pure and Applied Mathematics Volume 120 No. 5 2018, 215-238.

[6] Rosemary Green, James Milner, Edward J. M. Joy, Sutapa Agrawa and Alan D. Dangour. “DIETARY PATTERNS IN INDIA: A SYSTEMATIC REVIEW”. British Journal of Nutrition (2016), 116, 142–148.

[7] Barbara J Rolls, Erin L Morris, and Liane S Roe. “PORTION SIZE OF FOOD AFFECTS ENERGY INTAKE IN NORMAL-WEIGHT AND OVERWEIGHT MEN AND WOMEN”. Am J Clin Nutr 2002; 76:1207–13.

[8] Mirre Viskaal-van Dongen, Frans J. Kok, Cees de Graaf. “EATING RATE OF COMMONLY CONSUMED FOODS PROMOTES FOOD AND ENERGY INTAKE”. Appetite 56 (2011) 25–31.

[9] Sharadendu Bali, Maneshwar Singh Utaal. “INCOMPATIBLE FOODS: AN IMPORTANT CONCEPT TO UNDERSTAND TO PREVENT AUTO-IMMUNE DISORDERS”. Bali S et al. Int J Sci Rep. 2019 Sep; 5:266-270.

[10] T. Longvah, R. Ananthan, K. Bhaskarachary, K. Venkaiah. “INDIAN FOOD COMPOSITION TABLES 2017”. National Institute of Nutrition, India.

[11] “INTERNATIONAL COLLABORATIVE PROJECT TO COMPARE AND MONITOR THE NUTRITIONAL COMPOSITION OF PROCESSED FOODS”. European Journal of Preventive Cardiology 19(6) 1326–1332.

[12] Neha Rathi, Lynn Riddell and Anthony Worsley. “FOOD CONSUMPTION PATTERNS OF ADOLESCENTS AGED 14–16 YEARS IN KOLKATA, INDIA”. Rathi et al. Nutrition Journal (2017).

[13] “NUTRITIONAL INTAKE IN INDIA 2004-2005”. National Sample Survey Organisation Ministry of Statistics & Programme Implementation, Government of India 2007.

[14] C. Bren d’Amoura, B. Pandey, M. Reba, S. Ahmadd, F. Creutzig, K.C. Seto. “URBANIZATION, PROCESSED FOODS, AND EATING OUT IN INDIA”. Global Food Security.

[15] Marcus Maringer, Pieter van’t Veer, Naomi Klepacz, Muriel C. D. Verain, Anne Normann, Suzanne Ekman, Lada Timotijevic, Monique M. Raats and Anouk Geelen. “USER-DOCUMENTED FOOD CONSUMPTION DATA FROM PUBLICLY AVAILABLE APPS: AN ANALYSIS OF OPPORTUNITIES AND CHALLENGES FOR NUTRITION RESEARCH”. Maringer et al. Nutrition Journal (2018).

[16] Mohammed Al-Maolegi, Bassam Arkok. “AN IMPROVED APRIORI ALGORITHM FOR ASSOCIATION RULES”. International Journal on Natural Language Computing (IJNLC) Vol. 3, No.1, 2014.

[17] Kanwal Garg, Deepak Kumar. “COMPARING THE PERFORMANCE OF FREQUENT PATTERN MINING ALGORITHMS”. International Journal of Computer Applications (0975 – 8887) Volume 69– No.25, 2013.

[18] Rakesh Agrawal, Ramakrishnan Sant. “FAST ALGORITHMS FOR MINING ASSOCIATION RULES”.

[19] Melissa Harrell, Jose Medina, Blanche Greene-Cramer, Shreela V. Sharma, Monika Arora, and Gaurang Nazar. “UNDERSTANDING EATING BEHAVIORS OF NEW DELHI'S YOUTH”. Journal of Applied Research on Children: Informing Policy for Children at Risk, Volume 6 Article 8, 2015.

[20] Anand Ramanathan, Avinash Chandani. “INDUSTRY 4.0 IN FOOD INDUSTRY”. India Food Report 2018.

[21] Ashish Raina, Dhiraj Pathak, Varinder Singh Rana, Gaurav Bathla. “CONSUMPTION PATTERNS FOR READY TO EAT FOODS ITEMS IN PHAGWARA DISTRICT OF PUNJAB”. International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8, Issue- 9S2, July 2019.

[22] Ingmar Weber, Palakorn Achananuparp. “INSIGHTS FROM MACHINE-LEARNED DIET SUCCESS PREDICTION” Pacific Symposium on Biocomputing 2016.

[23] Wesley Tansey, Edward W. Lowe, Jr., James G. Scott. “DIET2VEC: MULTI-SCALE ANALYSIS OF MASSIVE DIETARY DATA”. 2016

[24] Nick Comly, Marissa Lee, William Locke. “PREDICTING WEIGHT LOSS USING THE MYFITNESSPAL DATASET”. 2017

[25] Fanny Thomas, Sonia Capelli “THE EFFECT OF THE NUMBER OF INGREDIENT IMAGES ON PACKAGE EVALUATION AND PRODUCT CHOICE”

[26] Mr. Bhavin M. Mehta, Mr. Nishay Madhani, Mrs. Radhika Patwardhan. “FIREBASE: A PLATFORM FOR YOUR WEB AND MOBILE APPLICATIONS”. International Journal of Advance Research in Science and Engineering, Vol 6, 2017.

[27] Smallbusiness.chron.com, How are teenagers affected by advertisements for fast food, Jeffrey Care.

[28] Shetty P (2013). “NUTRITION TRANSITION AND ITS HEALTH OUTCOMES”. Indian J Pediatr 80, Suppl. 1, S21–S27.

[29] Srinivas T. “THE COSMOPOLITAN INDIAN FAMILY, AUTHENTIC FOOD AND THE CONSTRUCTION OF CULTURAL UTOPIA”. International Journal of Sociology of the Family. 2006;32:191-221.

[30] National Sample Survey Organisation. Nutritional Intake in India 2004-2005. 2007 [cited 2013/7/20]; Available from: http://mospi.nic.in/rept%20_%20pubn/513_final.pdf.

[31] www.researchgate.net, Aysha Karnataka Baig, Munazza Saeed. “REVIEW OF TRENDS IN FAST FOOD CONSUMPTION”. May 2012

[32] www.news.com.aw, “THE REASONS WE EAT JUNK FOOD”

[33] https://livehealthy.chron.com/effects-food-preservatives-human-body-6876.html. Jessica Bruso. “FACTS ON LACK OF NUTRIENTS FROM A FAST FOOD DIET”

[34] Food Chemicals Induces Toxic Effect on Health: Overview Mukta Sharma1 , Anupama Rajput1*, Chhaya Rathod2 , Shobharam Sahu2 *1Department of Chemistry, FET, MRIIRS, Faridabad-121001, India 2Rajiv Gandhi college of Pharmacy, Nautanwa-Maharajganj, UP-273164,India*

[35] EXTENSIVE USE OF MONOSODIUM GLUTAMATE: A THREAT TO PUBLIC HEALTH? Kamal Niaz, Elizabeta Zaplatic, Jonathan Spoor *Department of Pharmacology and Toxicology, Faculty of Bioscience and Agri-Food and Environmental Technology*

[36] Formalizing Food Ingredients for Data Analysis and Knowledge Organization Usashi Chatterjee, Vinit Kumar & Devika P. Madalli

[37] Consumer Understanding of Food Quality, Healthiness, and Environmental Impact: A Cross-National Perspective Dacinia Crina Petrescu *Iris Vermeir and Ruxandra Malina Petrescu-Mag*

[38] Indian Food Code: Food Categorization System

[39] Health Star Rating system: Calculator and Style Guide, September 2020 Version 1

[40] The Healthfulness of the US Packaged Food and Beverage Supply: A Cross-Sectional Study *Abigail S. Baldridge ,Mark D. Huffman, Fraser Taylor Dagan Xavier, Brooke Bright, Linda V. Van Horn, Bruce Neal and Elizabeth Dunford*