МИНОБРНАУКИ РОССИИ
САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ
ЭЛЕКТРОТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ
«ЛЭТИ» ИМ. В.И. УЛЬЯНОВА (ЛЕНИНА)
Кафедра МО ЭВМ

ОТЧЕТ

по лабораторной работе №4

по дисциплине «Машинное обучение»

Тема: Ассоциативный анализ

Студент гр. 6304 _____ Ваганов Н.

Преподаватель _____ Жангиров Т. Р.

Санкт-Петербург

2020

Цель работы

Ознакомиться с методами ассоциативного анализа из библиотеки MLxtend

Ход работы

Загрузка данных

1. Был загружен датасет. В наборе данных содержится информация о купленных товарах
2. В исходном датасете содержалось много значений NaN - они были удалены. Также был получен список уникальных товаров и их количество (рис. 1)

```
print(unique_items)
print(len(unique_items))
```

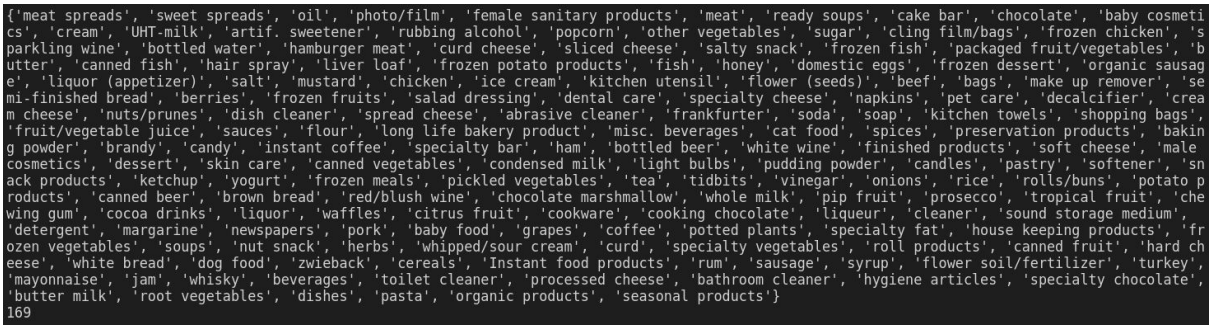


Рисунок 1 - Список уникальных товаров и их количество.

FPGrowth и FPMax

1. Данные были преобразованы к виду, удобному для анализа

```
from mlxtend.preprocessing import TransactionEncoder
te = TransactionEncoder()
te_ary = te.fit(np_data).transform(np_data)
data = pd.DataFrame(te_ary, columns=te.columns_)
```

2. Был проведен ассоциативный анализ с помощью FPGrowth при уровне поддержки 0.03 (рис. 2). Также был определен минимальный и максимальный уровень поддержки

```
from mlxtend.frequent_patterns import fpgrowth
result = fpgrowth(data, min_support=0.03, use_colnames = True)
result['length'] = result['itemsets'].apply(lambda x: len(x))
print(result)
```

	support	itemsets	length
0	0.082766	(citrus fruit)	1
1	0.058566	(margarine)	1
2	0.139502	(yogurt)	1
3	0.104931	(tropical fruit)	1
4	0.058058	(coffee)	1
..	...	...	...
58	0.033249	(whole milk, pastry)	2
59	0.047382	(other vegetables, root vegetables)	2
60	0.048907	(whole milk, root vegetables)	2
61	0.030605	(sausage, rolls/buns)	2
62	0.032232	(whole milk, whipped/sour cream)	2

[63 rows x 3 columns]

Рисунок 2 - Результат работы FPGrowth с минимальной поддержкой 0.03.

Таблица 1 - min/max поддержка для FPGrowth

	1	2
min support	0.03	0.03
max support	0.26	0.07

3. Аналогично с п.2, был произведен ассоциативный анализ для FPMax (рис. 3)

```
from mlxtend.frequent_patterns import fpmax
result_fp = fpmax(data, min_support=0.03, use_colnames = True)
result_fp['length'] = result_fp['itemsets'].apply(lambda x:
len(x))
result_fp
```

	support	itemsets	length
0	0.030402	(specialty chocolate)	1
1	0.031012	(onions)	1
2	0.032944	(hygiene articles)	1
3	0.033249	(berries)	1
4	0.033249	(hamburger meat)	1

Рисунок 3 - Частичный результат работы FPMax с минимальной поддержкой 0.03.

Таблица 2 - min/max поддержка для FPMax

	1	2
min support	0.03	0.03
max support	0.09	0.07

4. Были сравнены минимумы и максимумы значений поддержки для FPMax и FPGrowth (таблицы 1-2). Их значения различаются только для максимальных значений поддержки наборов длины 1. Это связано с тем, что в FPGrowth набор длины 1 может быть использован как элемент для наборов большей длины, в отличие от FPMax, где такое исключено.
5. Была построена гистограмма для 10 наиболее встречающихся товаров. Результат на рис. 4

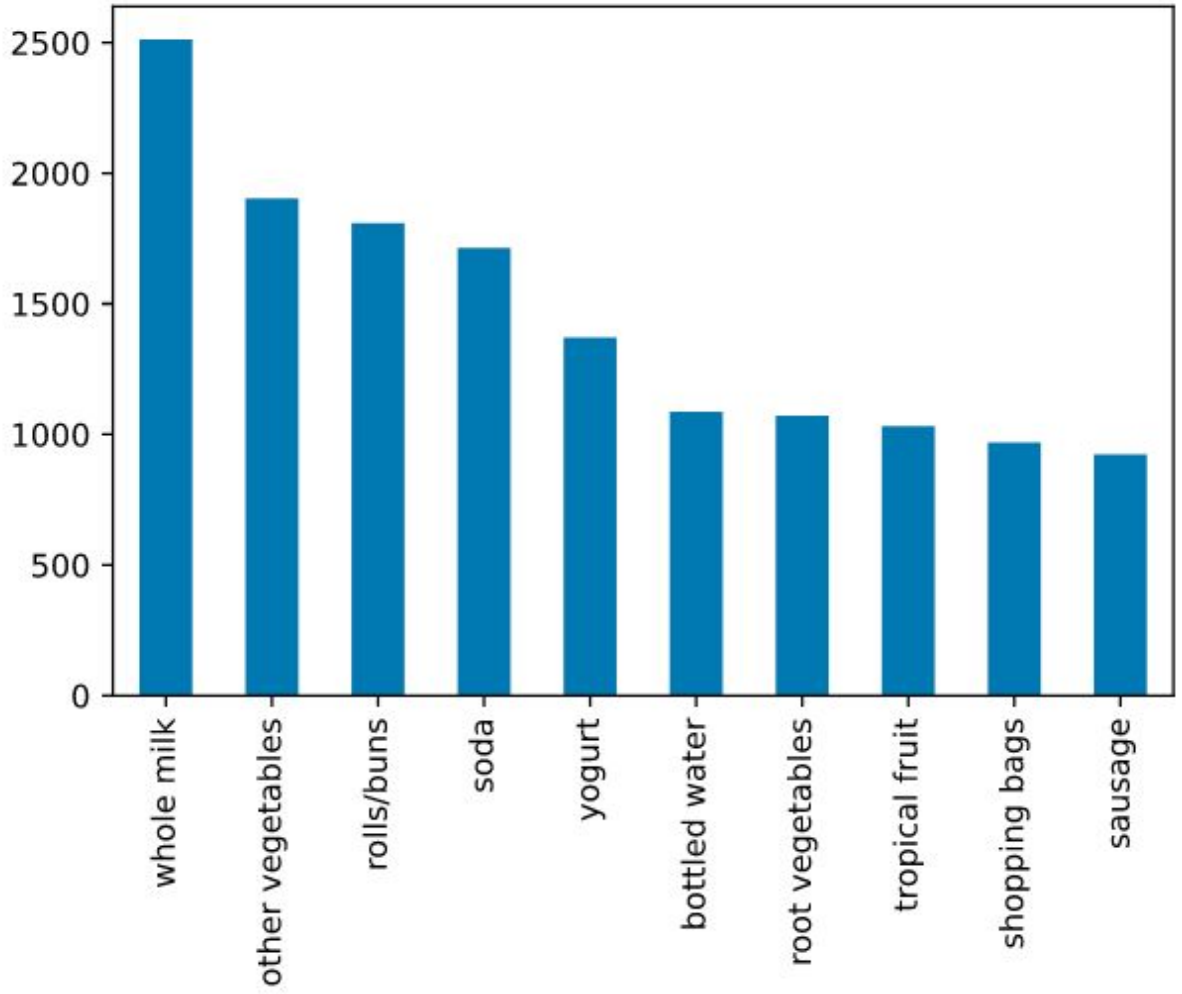


Рисунок 4 - Гистограмма наиболее встречающихся товаров.

Данные гистограммы полностью совпадают с результатами FPGrowth. Например - частота встречаемости для whole milk равна 2500, а уровень поддержки - 0.25. Аналогичная зависимость актуальна и для других товаров из списка.

6. Исходный датасет был преобразован. Результат представлен на рис. 5

```
items = ['whole milk', 'yogurt', 'soda', 'tropical fruit',
'shopping bags','sausage','whipped/sour cream', 'rolls/buns',
'other vegetables', 'root vegetables',
'pork', 'bottled water', 'pastry', 'citrus fruit', 'canned beer',
'bottled beer']
np_data = all_data.to_numpy()
np_data = [[elem for elem in row[1:] if isinstance(elem,str) and
elem in
items] for row in np_data]
```

	bottled beer	bottled water	canned beer	citrus fruit	other vegetables	pastry	pork	rolls/buns	root vegetables	sausage	shopping bags	soda	tropical fruit	whipped/sour cream	whole milk	yogurt
0	False	False	False	True	False	False	False	False	False	False	False	False	False	False	False	False
1	False	False	False	False	False	False	False	False	False	False	False	False	True	False	False	True
2	False	False	False	False	False	False	False	False	False	False	False	False	False	False	True	False
3	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	True
4	False	False	False	False	True	False	False	False	False	False	False	False	False	False	True	False
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
9830	False	False	False	True	False	False	False	False	True	True	False	False	False	True	True	False
9831	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False	False
9832	False	False	False	True	True	False	False	True	False	False	False	False	False	False	False	True
9833	True	True	False	False	False	False	False	False	False	False	False	True	False	False	False	False
9834	False	False	False	False	True	False	False	False	False	False	True	False	True	False	False	False

9835 rows × 16 columns

Рисунок 5 - Вид полученного датасета (16 товаров).

7. Был проведен анализ FPGrowth и FPMax для этого набора данных. Результаты представлены на рис. 6-7. В сравнении с предыдущими результатами, было найдено меньше наборов, однако для найденных наборов уровень поддержки остался прежним.

```
te = TransactionEncoder()
te_ary = te.fit(np_data).transform(np_data)
data_limited = pd.DataFrame(te_ary, columns=te.columns_)
result = fpgrowth(data_limited, min_support=0.03, use_colnames =
True)
result['length'] = result['itemsets'].apply(lambda x: len(x))
result
result_fp = fpmax(data_limited, min_support=0.03, use_colnames =
True)
result_fp['length'] = result_fp['itemsets'].apply(lambda x:
len(x))
result_fp
```

	support	itemsets	length
0	0.082766	(citrus fruit)	1
1	0.139502	(yogurt)	1
2	0.104931	(tropical fruit)	1
3	0.255516	(whole milk)	1
4	0.193493	(other vegetables)	1
5	0.183935	(rolls/buns)	1
6	0.080529	(bottled beer)	1
7	0.110524	(bottled water)	1
8	0.174377	(soda)	1
9	0.088968	(pastry)	1
10	0.108998	(root vegetables)	1
11	0.077682	(canned beer)	1
12	0.093950	(sausage)	1
13	0.098526	(shopping bags)	1
14	0.071683	(whipped/sour cream)	1
15	0.057651	(pork)	1
16	0.030503	(whole milk, citrus fruit)	2
17	0.056024	(yogurt, whole milk)	2
18	0.034367	(rolls/buns, yogurt)	2
19	0.043416	(yogurt, other vegetables)	2
20	0.035892	(other vegetables, tropical fruit)	2
21	0.042298	(whole milk, tropical fruit)	2
22	0.074835	(whole milk, other vegetables)	2
23	0.042603	(rolls/buns, other vegetables)	2
24	0.056634	(rolls/buns, whole milk)	2
25	0.034367	(bottled water, whole milk)	2
26	0.038332	(soda, rolls/buns)	2
27	0.040061	(soda, whole milk)	2
28	0.032740	(soda, other vegetables)	2
29	0.033249	(pastry, whole milk)	2
30	0.047382	(other vegetables, root vegetables)	2
31	0.048907	(whole milk, root vegetables)	2
32	0.030605	(sausage, rolls/buns)	2
33	0.032232	(whipped/sour cream, whole milk)	2

Рисунок 6 - Результат работы FPGrowth.

	support	itemsets	length
0	0.057651	(pork)	1
1	0.032232	(whipped/sour cream, whole milk)	2
2	0.077682	(canned beer)	1
3	0.080529	(bottled beer)	1
4	0.030503	(whole milk, citrus fruit)	2
5	0.033249	(pastry, whole milk)	2
6	0.030605	(sausage, rolls/buns)	2
7	0.098526	(shopping bags)	1
8	0.035892	(other vegetables, tropical fruit)	2
9	0.042298	(whole milk, tropical fruit)	2
10	0.047382	(other vegetables, root vegetables)	2
11	0.048907	(whole milk, root vegetables)	2
12	0.034367	(bottled water, whole milk)	2
13	0.034367	(rolls/buns, yogurt)	2
14	0.043416	(yogurt, other vegetables)	2
15	0.056024	(yogurt, whole milk)	2
16	0.032740	(soda, other vegetables)	2
17	0.038332	(soda, rolls/buns)	2
18	0.040061	(soda, whole milk)	2
19	0.042603	(rolls/buns, other vegetables)	2
20	0.056634	(rolls/buns, whole milk)	2
21	0.074835	(whole milk, other vegetables)	2

Рисунок 7 - Результат работы FPMax.

8. Был построен график изменения количества получаемых правил от уровня поддержки (рис. 8)

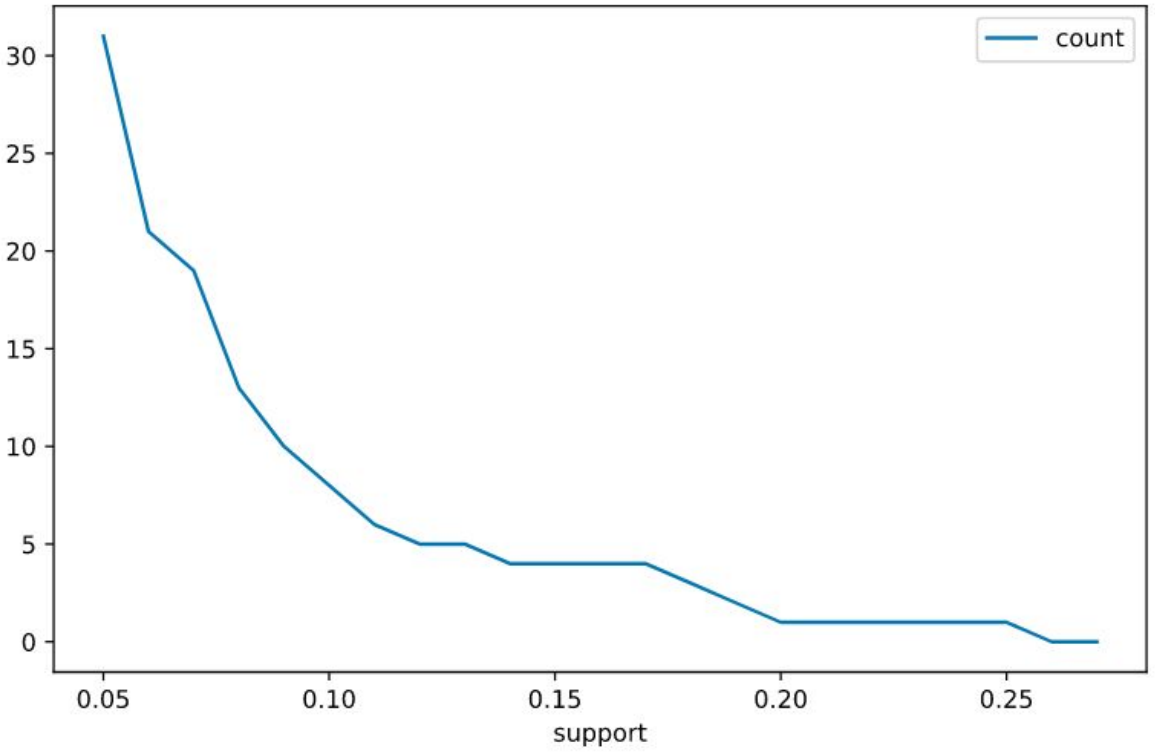


Рисунок 8 - График изменения количества получаемых правил от уровня поддержки.

Ассоциативные правила

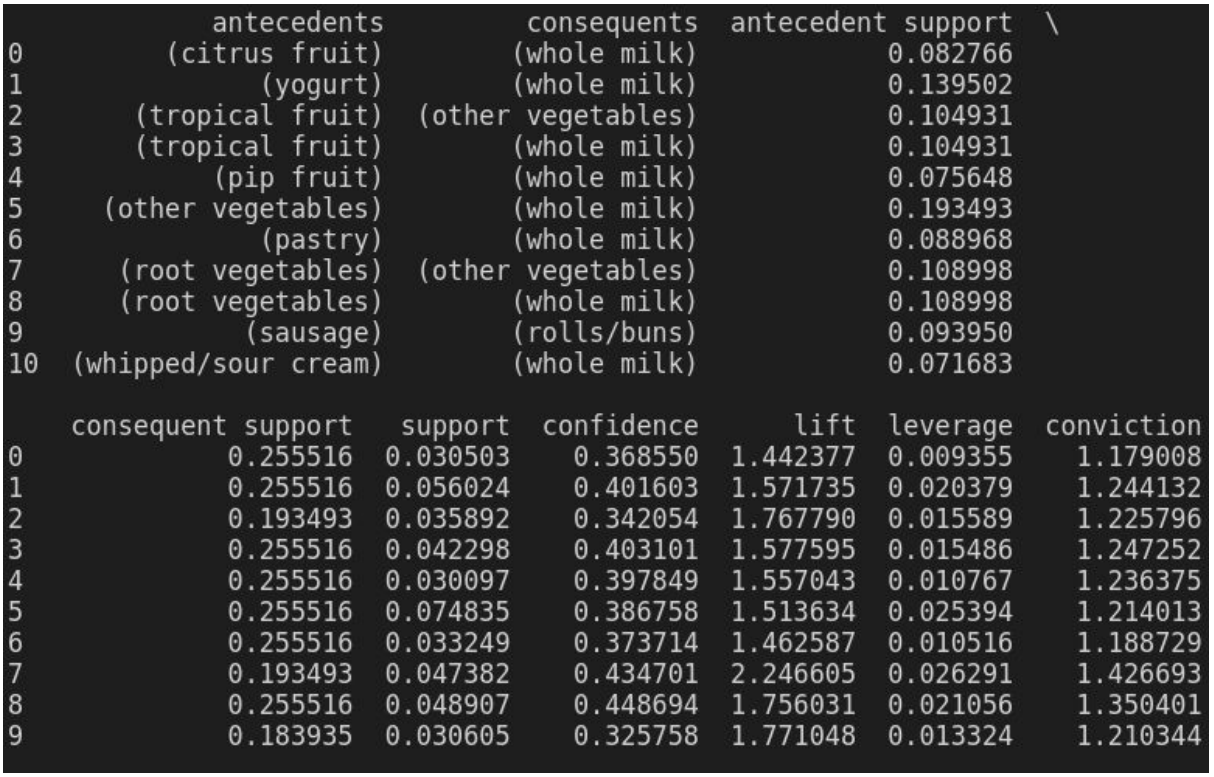
1. Набор данных был сформирован так, чтобы размер транзакции был 2 и более
2. Были получены частоты наборов, используя алгоритм FPGrowth с уровнем поддержки 0.03. Результат на рис. 9

	support	itemsets
0	0.082766	(citrus fruit)
1	0.058566	(margarine)
2	0.139502	(yogurt)
3	0.104931	(tropical fruit)
4	0.058058	(coffee)
...	...	...
58	0.033249	(pastry, whole milk)
59	0.047382	(other vegetables, root vegetables)
60	0.048907	(whole milk, root vegetables)
61	0.030605	(sausage, rolls/buns)
62	0.032232	(whipped/sour cream, whole milk)

63 rows × 2 columns

Рисунок 9 - Результат работы FPGrowth.

3. Был проведен ассоциативный анализ. Результат представлен на рис. 10. Полученные столбцы имеют следующий смысл:
    a. **antecedent** - антецент
    b. **consequent** - консеквент
    c. **antecedent support** - поддержка антецента
    d. **consequent support** - поддержка консеквента
    e. **support** - поддержка набора антецент -> консеквент
    f. **confidence** - вероятность получить консеквент в транзакции, содержащей антецент
    g. **lift** - частота встречаемости антецента и консеквента в отличие от случая, когда эти величины независимы
    h. **leverage** - разница между частотой появления антецента и консеквента вместе и частотой, ожидаемой при их статистической независимости
    i. **conviction** - зависимость консеквента от антецента



	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(citrus fruit)	(whole milk)	0.082766	0.255516	0.030503	0.368550	1.442377	0.009355	1.179008
1	(yogurt)	(whole milk)	0.139502	0.255516	0.056024	0.401603	1.571735	0.020379	1.244132
2	(tropical fruit)	(other vegetables)	0.104931	0.193493	0.035892	0.342054	1.767790	0.015589	1.225796
3	(tropical fruit)	(whole milk)	0.104931	0.255516	0.042298	0.403101	1.577595	0.015486	1.247252
4	(pip fruit)	(whole milk)	0.075648	0.255516	0.030097	0.397849	1.557043	0.010767	1.236375
5	(other vegetables)	(whole milk)	0.193493	0.255516	0.074835	0.386758	1.513634	0.025394	1.214013
6	(pastry)	(whole milk)	0.088968	0.255516	0.033249	0.373714	1.462587	0.010516	1.188729
7	(root vegetables)	(other vegetables)	0.108998	0.193493	0.047382	0.434701	2.246605	0.026291	1.426693
8	(root vegetables)	(whole milk)	0.108998	0.255516	0.048907	0.448694	1.756031	0.021056	1.350401
9	(sausage)	(rolls/buns)	0.093950	0.183935	0.030605	0.325758	1.771048	0.013324	1.210344
10	(whipped/sour cream)	(whole milk)	0.071683						

Рисунок 10 - Результат проведенного ассоциативного анализа.

4. Расчет проводится на основе выбранной метрики, по умолчанию используется метрика **confidence**. Проведено построение ассоциативных правил для каждой метрики. Параметры полученных распределений представлены в таблице 3

Таблица 4 - Характеристики полученных результатов по каждой метрике

	threshold	mean	median	mse
support	0.04	0.05	0.05	0.02
confidence	0.21	0.32	0.31	0.04
lift	1.34	1.67	1.59	0.09
leverage	0.01	0.01	0.01	0.006
conviction	1.06	1.17	1.17	0.07

5. Был построен граф для анализа (рис. 11). С помощью графа легко определить зависимости между антецедентами и консеквентами. В этом случае можно сказать, что для продуктов yogurt, sour cream, root vegetables, tropical food в транзакции, вероятно, будет присутствовать еще и whole milk. Для root vegetables почти

равновероятно в транзакции будут присутствовать whole milk или other vegetables.

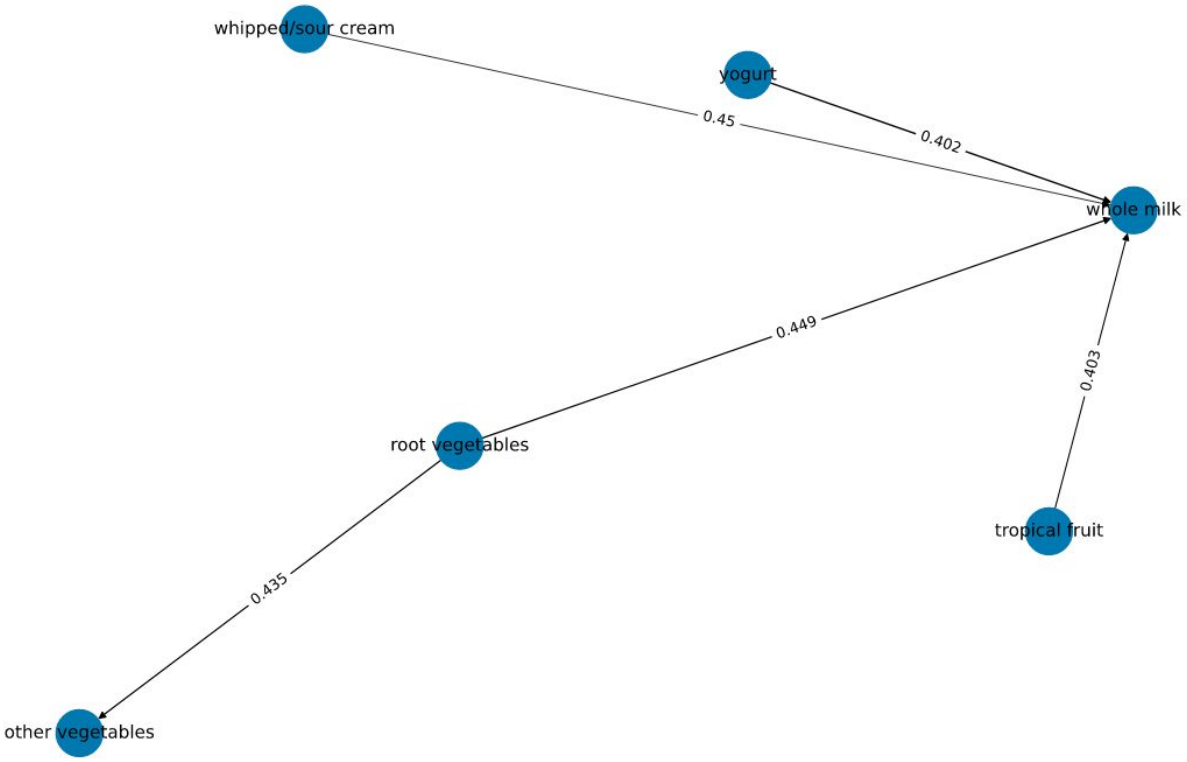


Рисунок 11 - Результат ассоциативного анализа, представленный в виде графа.

6. Полученный в ходе ассоциативного анализа результат также можно представить в виде, например, heatmap (рис.12)

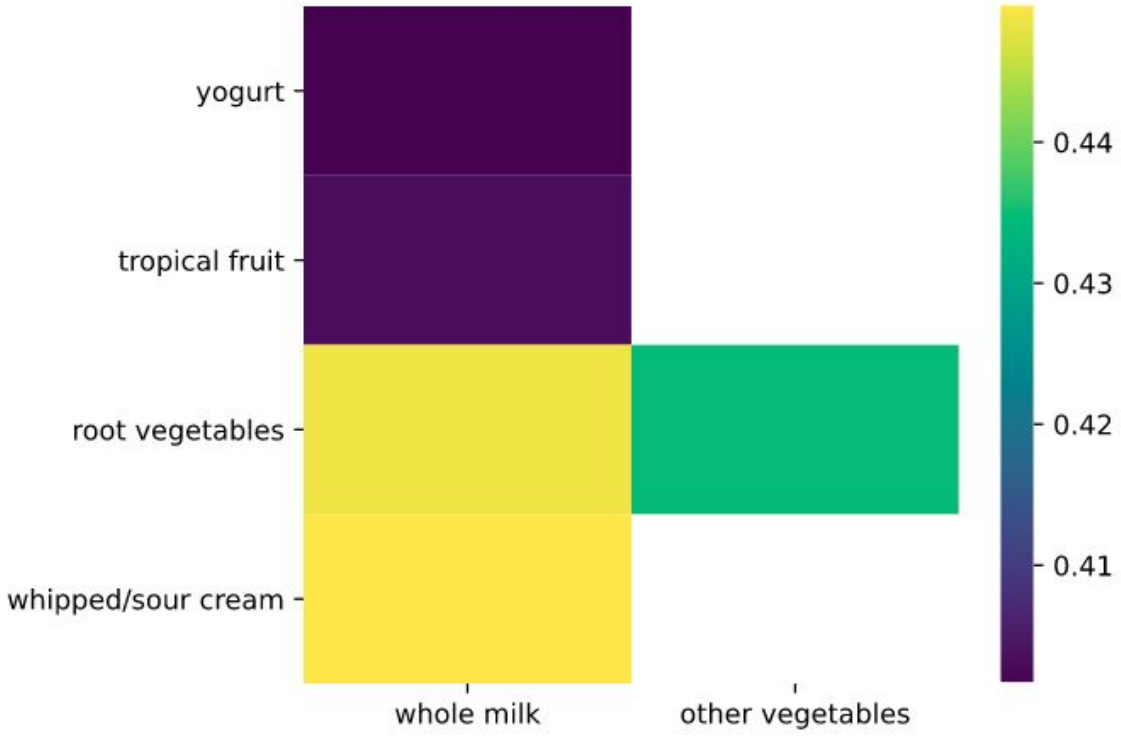


Рисунок 12 - Результат ассоциативного анализа, представленный в виде heatmap.

Вывод

Были изучены методы ассоциативного анализа из библиотеки MLxtend. В частности, были рассмотрены FPGrowth, FPMax, построение ассоциативных правил при помощи association_rules и дальнейшая их визуализация с помощью графа и heatmap.