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 _____ Антонов С.А.

Преподаватель _____ Жангиров Т.Р.

Санкт-Петербург

2020

Цель работы:

Ознакомиться с методами ассоциативного анализа из библиотеки MLxtend.

Ход работы:

Загрузка данных

1. На данном этапе был скачан и загружен датасет в датафрейм.

```
all_data = pd.read_csv('groceries - groceries.csv')
print(all_data)
```

	Item(s)	Item 1	Item 2	...	Item 30	Item 31	Item 32
0	4	citrus fruit	semi-finished bread	...	NaN	NaN	NaN
1	3	tropical fruit	yogurt	...	NaN	NaN	NaN
2	1	whole milk	NaN	...	NaN	NaN	NaN
3	4	pip fruit	yogurt	...	NaN	NaN	NaN
4	4	other vegetables	whole milk	...	NaN	NaN	NaN

Рисунок 1 Загруженный датасет

2. Данные были переформированы, а также были удалены все значения NaN.

```
np_data = all_data.to_numpy()
np_data = [[elem for elem in row[1:] if isinstance(elem,str)] for row in
np_data]
```

3. Получен список всех уникальных товаров.

```
unique_items = set()
for row in np_data:
    for elem in row:
        unique_items.add(elem)
```

4. Следующим шагом был сформирован датасет подходящий для частотного анализа.

```
dataset = [[elem for elem in all_data[all_data[1] == id][2] if elem in items]
for id in unique_items]
```

5. Получившийся список содержит 169 элементов.

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

```
{'cream', 'butter', 'cling film/bags', 'frozen chicken', 'liquor (appetizer)', 'soft cheese', 'preservation product
s', 'seasonal products', 'ham', 'whisky', 'potted plants', 'meat spreads', 'bottled water', 'prosecco', 'syrup', 'fro
zen fish', 'white wine', 'curd cheese', 'shopping bags', 'pastry', 'male cosmetics', 'salt', 'frozen fruits', 'rice',
'newspapers', 'hygiene articles', 'popcorn', 'pasta', 'baby food', 'cookware', 'flower (seeds)', 'yogurt', 'rubbing a
lcohol', 'soda', 'detergent', 'frozen potato products', 'ready soups', 'house keeping products', 'spices', 'cocoa dri
nks', 'pudding powder', 'roll products', 'semi-finished bread', 'nuts/prunes', 'berries', 'specialty chocolate', 'sou
nd storage medium', 'tea', 'cake bar', 'canned vegetables', 'sauces', 'hamburger meat', 'white bread', 'skin care',
'artif. sweetener', 'dessert', 'sweet spreads', 'candles', 'grapes', 'whole milk', 'herbs', 'salty snack', 'misc. bev
erages', 'citrus fruit', 'honey', 'bags', 'dental care', 'snack products', 'chewing gum', 'mayonnaise', 'beverages',
'specialty vegetables', 'turkey', 'condensed milk', 'zwieback', 'pet care', 'baby cosmetics', 'packaged fruit/vegetab
les', 'soups', 'long life bakery product', 'frankfurter', 'toilet cleaner', 'jam', 'hard cheese', 'tidbits', 'meat',
'frozen meals', 'candy', 'sparkling wine', 'salad dressing', 'tropical fruit', 'finished products', 'bottled beer',
'sliced cheese', 'organic sausage', 'instant coffee', 'chocolate marshmallow', 'coffee', 'margarine', 'photo/film',
'sugar', 'spread cheese', 'dishes', 'hair spray', 'Instant food products', 'frozen vegetables', 'baking powder', 'liq
ueur', 'domestic eggs', 'specialty fat', 'ketchup', 'butter milk', 'canned fish', 'light bulbs', 'beef', 'chicken',
'dog food', 'canned beer', 'cat food', 'fruit/vegetable juice', 'fish', 'other vegetables', 'softener', 'kitchen towe
ls', 'oil', 'onions', 'UHT-milk', 'flower soil/fertilizer', 'brandy', 'napkins', 'vinegar', 'kitchen utensil', 'make
up remover', 'red/blush wine', 'canned fruit', 'flour', 'brown bread', 'cereals', 'dish cleaner', 'abrasive cleaner',
'potato products', 'rum', 'specialty cheese', 'specialty bar', 'waffles', 'whipped/sour cream', 'ice cream', 'decalci
fier', 'processed cheese', 'chocolate', 'mustard', 'liver loaf', 'curd', 'rolls/buns', 'organic products', 'bathroom
cleaner', 'soap', 'liquor', 'cream cheese', 'root vegetables', 'female sanitary products', 'nut snack', 'cleaner', 'p
ip fruit', 'pork', 'sausage', 'cooking chocolate', 'frozen dessert', 'pickled vegetables'}
169
```

Рисунок 2 Список товаров.

FPGrowth и FPMax.

1. Данные были преобразованы к удобному для анализа виду:

```
te = TransactionEncoder()
te_ary = te.fit(np_data).transform(np_data)
data = pd.DataFrame(te_ary, columns=te.columns_)
print(data)
```

2. Был проведен ассоциативный анализ с использованием алгоритмов FPGrowth и FPMax при уровне поддержки 0.03.

```
result_fpgrowth = fpgrowth(data, min_support=0.03, use_colnames = True)
result_fpgrowth['length'] = np.fromiter(map(len,
result_fpgrowth['itemsets']),dtype=int)
result_fpmax = fpmax(data, min_support=0.03, use_colnames = True)
result_fpmax['length'] = np.fromiter(map(len,
result_fpmax['itemsets']),dtype=int)
```

3. Были проанализированы полученные результаты:

Количество элементов	Min/Max значение уровня	FPGrowth	FPMax
1	Min	0.0304	0.0304
	Max	0.2555	0.0985
2	Min	0.0300	0.0300
	Max	0.0748	0.0748

4. FPMax – это вариант FPGrowth, фокусирующийся на получении максимальных наборов предметов. Наборов элементов X максимальный, если X является частым и не встречается такого частого супер-шаблона, содержащего X. Т.е. X не может быть под-шаблоном более частого шаблона.
5. Частота встречаемости товара пропорционально значению уровня поддержки для конкретного товара:

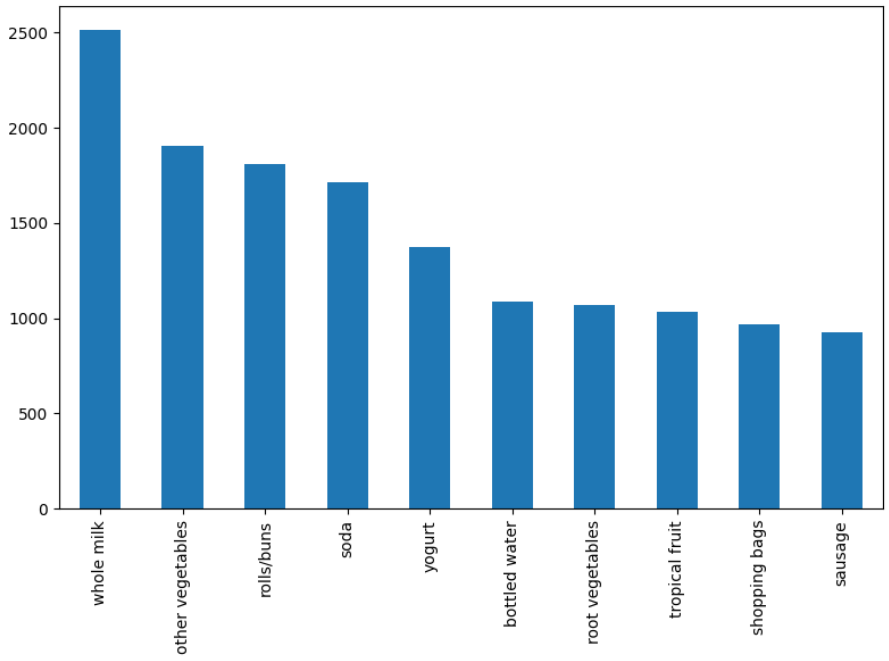

Рисунок 3 10 самых часто встречающихся товаров

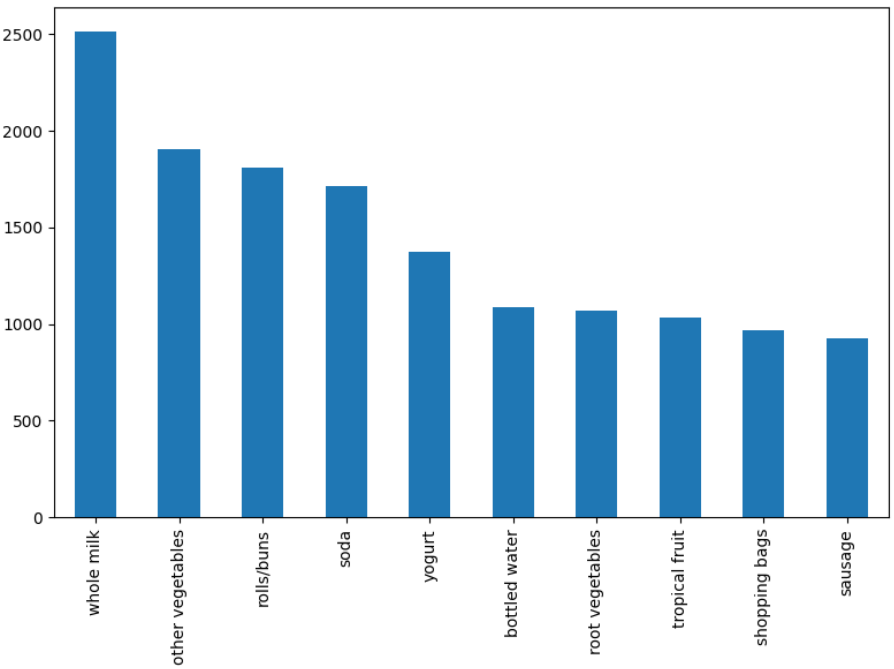

Рисунок 4 10 наборов с максимальным уровнем поддержки

6. Преобразуем набор так, чтобы он содержал ограниченный набор товаров:

```
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_f = all_data.to_numpy()
np_data_f = [[elem for elem in row[1:] if isinstance(elem,str) and elem in
items] for row in np_data_f]
```

7. Проведен анализ FPGrowth и FPMax для нового набора данных. Максимальные значения уровня поддержки не изменились, в то время как минимальные изменились. Причиной такого изменения является изменение самих товаров. Товар, уровень значения которого был минимален ранее, теперь удален, поэтому значение стало другим. Значения уровня поддержки товаров, которые остались – не изменились.

Количество элементов	Min/Max значение уровня	FPGrowth	FPMax
1	Min	0.0576	0.0576
	Max	0.2555	0.0985
2	Min	0.0305	0.0305
	Max	0.0748	0.0748

8. Было исследовано изменение количества получаемых правил от уровня поддержки.

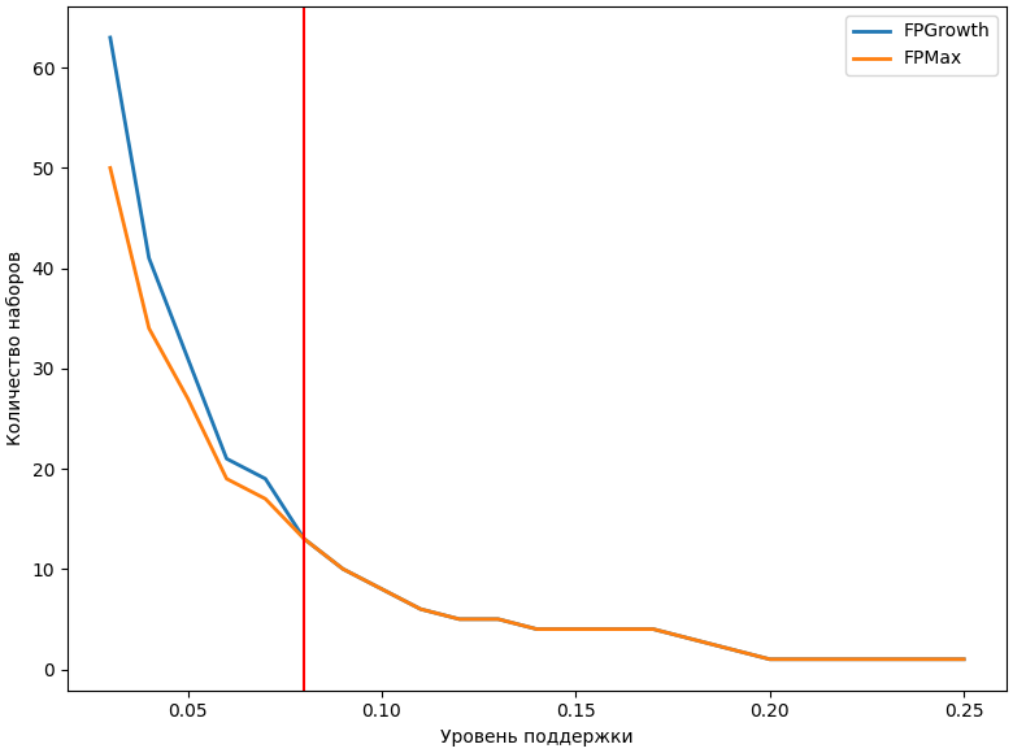

Рисунок 5 Зависимость количества наборов от уровня поддержки

Ассоциативные правила

1. Сформирован набор данных из определенных товаров, чтобы размер транзакции был 2 и более. После чего получены частоты наборов с использованием алгоритма FPGrowth.

```
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]
np_data = [row for row in np_data if len(row) > 1]
result = fpgrowth(data, min_support=0.05, use_colnames = True)
print(result.sort_values('support'))
```

2. Проведен ассоциативный анализ, по умолчанию расчет производится на основе метрики *Confidence*

```
rules = association_rules(result, min_threshold = 0.3)
```

```
              antecedents   consequents  ...  leverage  conviction
0              (yogurt)  (whole milk)  ...  0.020379    1.244132
1  (other vegetables)  (whole milk)  ...  0.025394    1.214013
2        (rolls/buns)  (whole milk)  ...  0.009636    1.075696
```

Рисунок 6 Результаты ассоциативного анализа.

Confidence (Уверенность) – вероятность увидеть консеквент в транзакции при условии, что оно также содержит антецедент. Метрика не является симметричной или направленной. Уверенность равна 1 – максимальная для правила $A \rightarrow B$, если консеквент и антецедент всегда встречаются вместе.

$$\text{confidence}(A \rightarrow B) = \frac{\text{support}(A \rightarrow B)}{\text{support}(A)}, \text{range: } [0,1]$$

Lift (Подъем) – насколько чаще предшествующее и последующее действие правила $A \rightarrow B$ встречается вместе, чем ожидалось, если бы они были статически независимыми. Если A и B независимы, оценка *Lift* будет равно 1.

$$\text{lift}(A \rightarrow B) = \frac{\text{confidence}(A \rightarrow B)}{\text{support}(B)}, \text{range: } [0, \infty]$$

Leverage (Рычаг) – разница между наблюдаемой частотой появления A и B вместе и частотой, которую можно было бы ожидать, если бы A и B были независимыми. Значение *Leverage = 0* указывает на независимость.

$$\text{levarage}(A \rightarrow B) = \text{support}(A \rightarrow B) - \text{support}(A) \times \text{support}(B),$$
$$\text{range: } [-1,1]$$

Conviction (Убеждение) – насколько консеквент сильно зависит от антецедента. Как и в случае с *Lift*, если предметы независимы, *Conviction = 1*.

$$\text{conviction}(A \rightarrow B) = 1 - \frac{\text{support}(B)}{1 - \text{confidence}(A \rightarrow B)},$$
$$\text{range: } [0, \infty]$$

3. Проведено построение ассоциативных правил для различных метрик. Значение *min_treshold* выбрано на основе того, чтобы выводилось не менее 10 правил.

```
association_rules_res = association_rules(result_fpgrowth, metric='confidence',
min_threshold = 0.34)
association_rules(result_fpgrowth, metric='lift', min_threshold = 1.75)
association_rules(result_fpgrowth, metric='leverage', min_threshold = 0.016)
association_rules(result_fpgrowth, metric='conviction', min_threshold = 1.18)
```

4. Рассчитаны описательные статистики для метрик.

```
association_rules_res.iloc[:,2:].describe()
```

	Antecedent support	Consequent support	Support	Confidence	Lift	Leverage	Conviction
count	10	10	10	10	10	10	10
mean	0.107992	0.243111	0.0431	0.4006	1.6655	0.01687	1.266513
std	0.036035	0.026151	0.0142	0.0353	0.2374	0.00612	0.081671
min	0.071683	0.193493	0.0300	0.3420	1.4423	0.00935	1.179008
25%	0.084316	0.255516	0.0324	0.3769	1.5244	0.01155	1.216959
50%	0.104931	0.255516	0.0390	0.3997	1.5746	0.01553	1.240253
75%	0.108998	0.255516	0.0485	0.4268	1.7588	0.02088	1.324614
max	0.193493	0.255516	0.0748	0.4496	2.2466	0.02629	1.426693

5. Построен граф для существующего анализа.

```
rules = association_rules(result, min_threshold = 0.4, metric='confidence')
```

Каждая вершина графа отображает набор товаров. Граф ориентирован от антецедента к консеквенту. Ширина ребра отображает уровень поддержки, а подпись на ребре отображает уверенность.

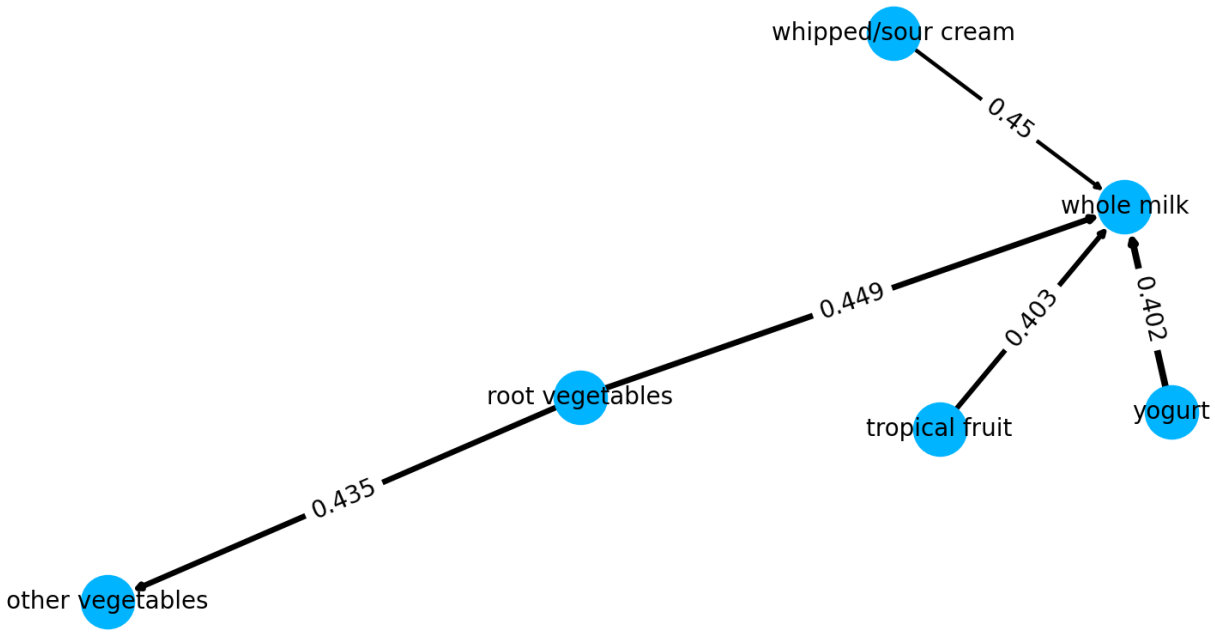

Рисунок 7 Граф набора товаров

6. Из графа можно сделать выводы, что если в транзакции есть предметы tropical fruit, yogurt и др., то с высокой вероятностью в транзакции будет присутствовать whole milk, а если root vegetables, то other vegetables.
7. Альтернативные способы отображения правил.

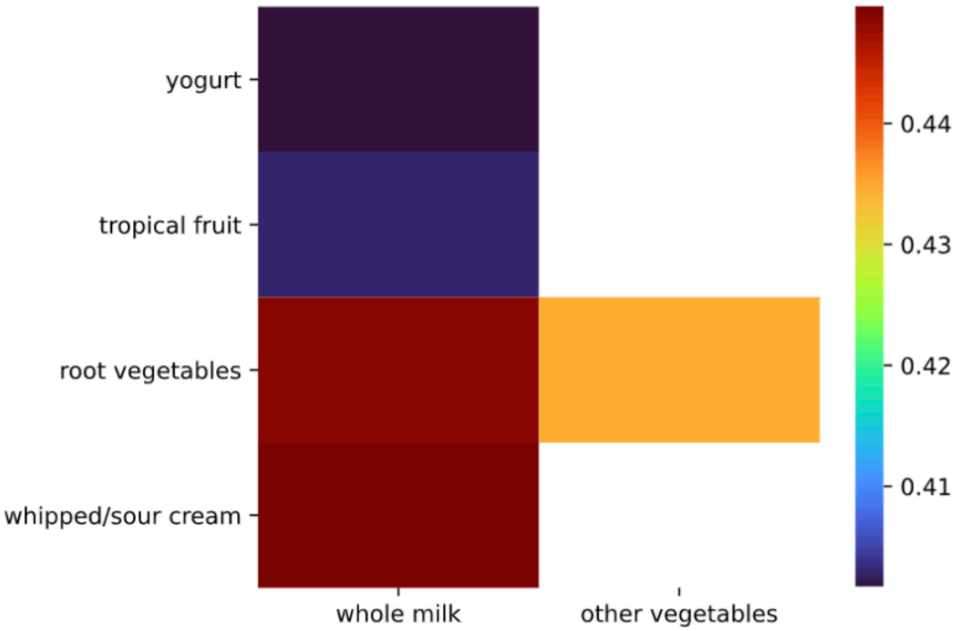

Рисунок 8 Heatmap значений Confidence

	whole milk	other vegetables
yogurt	0.401603	NaN
tropical fruit	0.403101	NaN
root vegetables	0.448694	0.434701
whipped/sour cream	0.449645	NaN

Рисунок 9 Текстовое представление датафрейма правил

Выводы:

В результате выполнения лабораторной работы были изучены методы ассоциативного анализа из библиотеки MLxtend. Были рассмотрены алгоритмы FPGrowth и FPMax, а также построение ассоциативных правил с помощью association_rules. Такие алгоритмы применяются для построения рекомендаций.

.