УДК 336.74

# АНАЛИЗ МЕДИЦИНСКИХ ДАННЫХ С ПОМОЩЬЮ АЛГОРИТМА APRIORI НА ЯЗЫКЕ R

***В.С. Лянцевич***
*Студент кафедры математического и информационного обеспечения экономических систем*
*УО ГрГУ им. Я.Купалы,*
valeria1lyantsevich@gmail.com

***Н.В. Марковская***
*Доцент кафедры математического и информационного обеспечения экономических систем*
*УО ГрГУ им. Я.Купалы,*
n.markovskaya@grsu.by

***В.С. Лянцевич***
*Студентка 4 курса Гродненского государственного университета имени Янки Купалы кафедры математического и информационного обеспечения экономических систем.*

***Н.В. Марковская***
*Доцент кафедры математического и информационного обеспечения экономических систем УО ГрГУ им.Я.Купалы, кандидат физико-математических наук, доцент.*

**Аннотация.** Статья рассматривает использование алгоритма Apriori для выявления ассоциативных правил в медицинском наборе данных опухоли молочной железы. Анализируются особенности структур данных, проводится их изменение для применения алгоритма, оценивается реализация Apriori. Исследование проведено на основе преобразованного набора данных, выявлены ключевые закономерности, а также проанализировано влияние параметров алгоритма на качество извлекаемых правил. Сделаны выводы о применимости алгоритма Apriori в задачах анализа медицинских данных, предложены рекомендации по его настройке для различных типов данных.

**Ключевые слова:** ассоциативные правила, алгоритм Apriori, медицинские данные, опухоль молочной железы.

**Введение.** Обнаружение ассоциативных правил – это метод, направленный на выявление логических закономерностей между элементами. Он гарантирует нахождение сильных правил в базе данных с помощью указанных показателей.

Правила ассоциаций представляют выражениями: «если "некоторое условие", то "некоторое следствие или результат"». На практике их записывают определённой математической записью: {некоторое условие} => {некоторое следствие или результат}.

Главная проблема поиска ассоциативных правил – большое количество правил, которые появляются при анализе. Для уменьшения количества были придуманы ограничения: они сохраняют только те правила, значения мер которых превышают заданные пороги. К ним относятся поддержка, доверие, лифт и уверенность.

Один из способов поиска правил ассоциации – алгоритм Apriori. Это классический метод анализа данных, который находит скрытые закономерности и зависимости между различными элементами.

Apriori традиционно применяют для анализа покупательской корзины, поиска связей между приобретаемыми товарами. Однако эта статья посвящена использованию алгоритма

для медицинской сферы: найти характеристики, которые будут сопровождать различные опухоли.

**Основная часть.** Исследование проводится с использованием набора данных Breast Cancer Wisconsin (Diagnostic) Data Set. Наша цель – определить признаки, которые чаще всего встречаются вместе с конкретным видом образования.

Breast Cancer Wisconsin (Diagnostic) Data Set содержит идентификатор, поставленный диагноз и десять действительных характеристик [1]:

1 Радиус – среднее значение расстояний от центра до точек по периметру.

2 Периметр – общее расстояние между точками.

3 Площадь измеряется простым подсчетом количества пикселей внутри змеи и добавлением половины пикселей по периметру.

4 Компактность получается из объединения периметра и площади с использованием формулы 1 периметр$^2$/площадь. Это безразмерное число увеличивается с нерегулярностью границы.

5 Гладкость количественно определяется путём измерения разницы между длиной радиальной линии и средней длиной линий, её окружающих.

6 Вогнутостью является выраженность вогнутых частей контура.

7 Вогнутые точки – это характеристика, похожая на вогнутость, но измеряет только количество, а не величину вогнутостей контура.

8 Симметрия измеряется путём нахождения большой оси или самой длинной хорды, проходящей через центр.

9 Фрактальная размерность клетки аппроксимируется с помощью «приближения береговой линии», описанного Мандельбротом. Периметр ядра измеряется с помощью все более крупных «линеек». По мере увеличения размера линейки, уменьшая точность измерения, наблюдаемый периметр уменьшается.

10 Текстура ядра клетки измеряется путем нахождения дисперсии интенсивностей серой шкалы в составляющих пикселях.

В наборе содержится информация о 357 доброкачественных образованиях и 212 – о злокачественных. Для каждой записи определены среднее значение, экстремальное (наихудшее) значение и стандартное отклонение каждой характеристики [2].

Характеристики представлены в виде численных значений. Однако алгоритм Apriori предназначен для поиска ассоциативных правил в категориальных данных. Следовательно, для последующего применения нужно преобразовать числовые данные.

Для реализации алгоритма будут использоваться средние значения признаков. По этой причине из файла данных удалим ненужные столбцы и изменим названия других.

```
breastc <- subset(breastc, select = -c(id, X, perimeter_se, area_se, smoothness_se, compactness_se, concavity_se, concave.points_se, symmetry_se, fractal_dimension_se, radius_worst, texture_worst, perimeter_worst, area_worst, smoothness_worst, compactness_worst, concavity_worst, concave.points_worst, symmetry_worst, texture_se, radius_se, fractal_dimension_worst)).
```

```
breastc <- breastc %>% rename(radius = radius_mean, texture = texture_mean, perimeter = perimeter_mean, area = area_mean, smoothness = smoothness_mean, compactness = compactness_mean, concavity = concavity_mean, concave_points = concave.points_mean, symmetry = symmetry_mean, fractal_dimension = fractal_dimension_mean)
```

Далее данные разделим на три группы, выделяя три категории: с низкими, средними и высокими значениями. В языке R это можно осуществить с помощью функции cut(). Она автоматически определяет минимальное и максимальное значение переменной, а затем делит диапазон на равные части.

В результате вместо чисел будут использоваться соответствующие слова. После этого приводим данные к виду транзакций.

```
breastc_discr <- breastc %>%
+       mutate(across(where(is.numeric), ~ cut(., breaks = 3, labels = c("Low", "Medium", "High"))))
```

```
breastc_trans <- as(breastc_discr, "transactions").
```

Количество доброкачественных и злокачественных наблюдений в наборе данных отличается. Поэтому поиск правил проводится раздельно.

Для злокачественного случая пороговое значение поддержки будет равно 0,1, а доверия – 0,8. Следствием будет выступать злокачественная опухоль, условием – признаки. Показатель доверия отражает частоту выполнения правила, поэтому отсортируем результаты по этому критерию.

```
rules_m <- apriori(breastc_trans,
      parameter = list(support = 0.1, confidence = 0.8, minlen = 2),
      appearance = list(rhs = "diagnosis=M", default = "lhs"))
inspect(head(sort(rules_m, by = "confidence"), 5))
plot(rules_m)
```

```
    lhs                          rhs             support confidence  coverage
[1] {area=Medium,
     concavity=Medium}       => {diagnosis=M} 0.1019332          1 0.1019332
[2] {area=Medium,
     concave_points=Medium}  => {diagnosis=M} 0.1318102          1 0.1318102
[3] {texture=Medium,
     area=Medium}            => {diagnosis=M} 0.1265378          1 0.1265378
[4] {area=Medium,
     symmetry=Medium}        => {diagnosis=M} 0.1212654          1 0.1212654
[5] {area=Medium,
     smoothness=Medium}      => {diagnosis=M} 0.1458699          1 0.1458699
```

*Рисунок 1.* Ассоциативные правила со злокачественным диагнозом в правой части

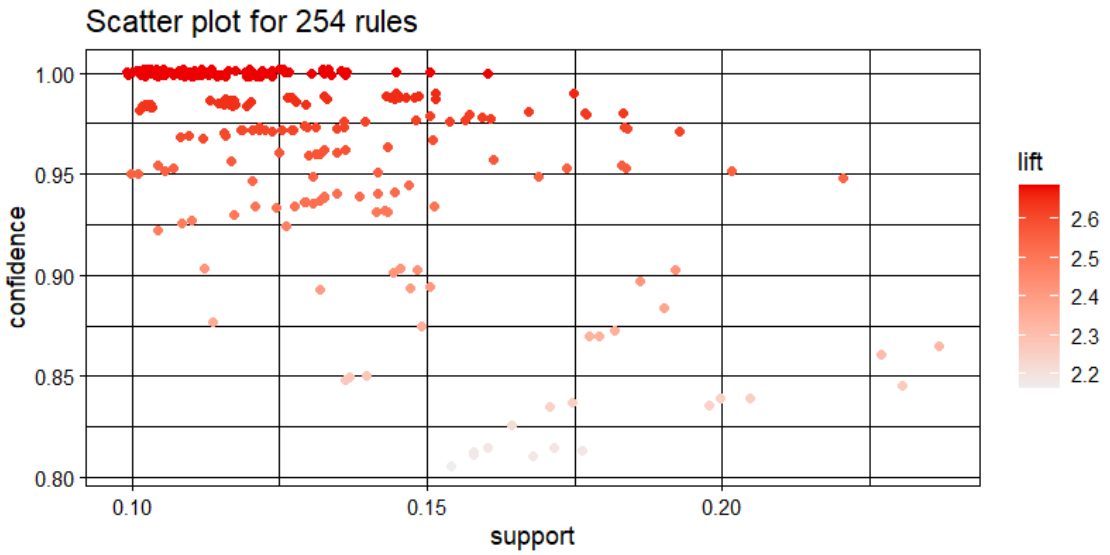


Рисунок 2. Диаграмма рассеяния ассоциативных правил со злокачественным диагнозом в правой части

Итого получилось пять важных правил.

1 Правило {среднее значение площади, среднее значение вогнутости} => {злокачественная опухоль} подразумевает, что сочетание площади и выраженности вогнутых частей контура указывает на заболевание.

2 Правило {среднее значение площади, среднее значение вогнутых точек} => {злокачественная опухоль} значит, что величина площади и количество вогнутых частей контура характерны для злокачественных образований.

3 Правило {среднее значение площади, среднее значение текстуры} => {злокачественная опухоль} выражает влияние площади и стандартного отклонения значений по серой шкале на конечный диагноз.

4 Правило {среднее значение площади, среднее значение симметрии} => {злокачественная опухоль} указывает на взаимосвязь между признаками условия и признаком следствия.

5 Правило {среднее значение площади, среднее значение гладкости} => {злокачественная опухоль} означает, что средние величины площади и локального изменения длин радиусов негативно сказываются на характере образования.

У полученных правил ассоциации значения доверия равны единице, то есть правило полностью достоверно. Хотя значения поддержки невелики, можно сделать вывод, что такого вида правило будет встречаться в каждой десятой записи.

Рисунок 1 показывает, что достаточно средних значений характеристик для выявления злокачественности опухоли. На Рисунке 2 видно, что, хотя правила встречаются редко, они достоверны и признаки будут встречаться вместе часто.

Высокие значения признаков будут только подтверждать неблагоприятный диагноз. Тем не менее, найдём такие правила, где слева будет хотя бы один показатель с высоким значением. Для этого изменим порог поддержки на 0,01. Здесь рассмотрим десять правил.

```
rules_m <- apriori(breastc_trans,
        parameter = list(support = 0.01, confidence = 0.8, minlen = 2),
        appearance = list(rhs = "diagnosis=M", default = "lhs"))
rules_m_high <- subset(rules_m, subset = lhs %pin% "High")
inspect(head(sort(rules_m_high, by = "confidence"), 10))
```

```
     lhs                                rhs           support    confidence coverage
[1]  {area=High}                     => {diagnosis=M} 0.01405975 1          0.01405975
[2]  {compactness=High}              => {diagnosis=M} 0.02284710 1          0.02284710
[3]  {radius=High}                   => {diagnosis=M} 0.03866432 1          0.03866432
[4]  {concave_points=High}           => {diagnosis=M} 0.03866432 1          0.03866432
[5]  {perimeter=High}                => {diagnosis=M} 0.03866432 1          0.03866432
[6]  {area=High, concavity=High}     => {diagnosis=M} 0.01054482 1          0.01054482
[7]  {radius=High, area=High}        => {diagnosis=M} 0.01405975 1          0.01405975
[8]  {area=High, concave_points=High} => {diagnosis=M} 0.01405975 1          0.01405975
[9]  {perimeter=High, area=High}     => {diagnosis=M} 0.01405975 1          0.01405975
[10] {area=High, compactness=Medium} => {diagnosis=M} 0.01054482 1          0.01054482
     count
```

*Рисунок 3.* Ассоциативные правила со злокачественным диагнозом в правой части и хотя бы одним высоким показателем в левой части

Рисунок 3 показывает, что высокие значения характеристик являются отличительной чертой злокачественного течения болезни. Видно, что первые пять правил состоят из одного параметра в левой части: площади, компактности, радиуса, вогнутых точек и периметра. Этого достаточно, чтобы установить злокачественный диагноз.

1 Правило {высокое значение площади, высокое значение вогнутости} => {злокачественная опухоль} показывает, как признаки из левой части воздействуют на результат правой части.

2 Правило {высокое значение площади, высокое значение радиуса} => {злокачественная опухоль} подтверждает, что такие величины площади и радиуса свидетельствуют о злокачественности.

3 Правило {высокое значение площади, высокое значение вогнутых точек} => {злокачественная опухоль} работает аналогичным образом, как и предыдущие правила.

4 Правило {высокое значение площади, высокое значение периметра} => {злокачественная опухоль} содержит в левой части одинакового рода характеристики, которые негативно сказываются на течении болезни.

5 Правило {высокое значение площади, среднее значение компактности} => {злокачественная опухоль} интересует больше остальных. Видно, что показатели признаков отличаются, но величина компактности никак не влияет на правдивость правила.

Значения доверия этих ассоциативных правил всё ещё равны единице в отличие от поддержки. Дело в меньшем количестве высоких значений. Проблема легко решается добавлением записей с высокими показателями.

Теперь перейдём к ситуации, когда опухоль доброкачественная. Таких случаев больше, поэтому значение поддержки будет равно 0,3, а значение доверия останется прежним.

```
rules_b <- apriori(breastc_trans,
      parameter = list(support = 0.3, confidence = 0.8, minlen = 2),
      appearance = list(rhs = "diagnosis=B", default = "lhs"))
inspect(head(sort(rules_b, by = "confidence"), 5))
plot(rules_b)
```

```
    lhs                      rhs             support confidence  coverage
[1] {texture=Low,
     perimeter=Low,
     compactness=Low,
     concave_points=Low} => {diagnosis=B} 0.3725835  0.9860465 0.3778559
[2] {texture=Low,
     perimeter=Low,
     compactness=Low,
     concavity=Low,
     concave_points=Low} => {diagnosis=B} 0.3725835  0.9860465 0.3778559
[3] {texture=Low,
     perimeter=Low,
     area=Low,
     compactness=Low,
     concave_points=Low} => {diagnosis=B} 0.3725835  0.9860465 0.3778559
[4] {texture=Low,
     perimeter=Low,
     area=Low,
     compactness=Low,
     concavity=Low,
     concave_points=Low} => {diagnosis=B} 0.3725835  0.9860465 0.3778559
[5] {radius=Low,
     texture=Low,
     compactness=Low,
     concave_points=Low} => {diagnosis=B} 0.3637961  0.9857143 0.3690685
```

*Рисунок 4.* Ассоциативные правила с доброкачественным диагнозом в правой части

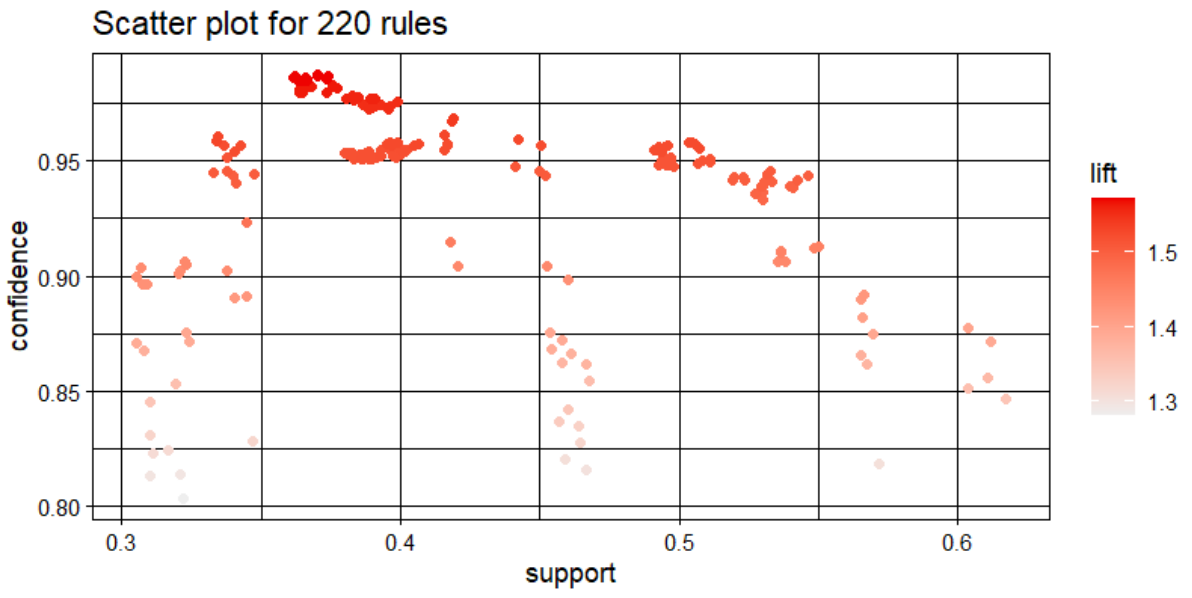


*Рисунок 5.* Диаграмма рассеяния ассоциативных правил с доброкачественным диагнозом в правой части

Получены пять важных правил.

1 Правило {низкое значение текстуры, низкое значение периметра, низкое значение компактности, низкое значение вогнутых точек} => {доброкачественная опухоль} занимает первое место и указывает на связь низких показателей с качеством заболевания.

2 Правило {низкое значение текстуры, низкое значение периметра, низкое значение компактности, низкое значение вогнутости, низкое значение вогнутых точек} => {доброкачественная опухоль} позволяет понять, что на конечный диагноз повлияли представленные характеристики.

3 Правило {низкое значение текстуры, низкое значение периметра, низкое значение площади, низкое значение компактности, низкое значение вогнутых точек} => {доброкачественная опухоль} означает, что низкие величины периметра, площади, компактности, стандартного отклонения значений по серой шкале и количество вогнутых точек негативно сказываются на характере образования.

4 Правило {низкое значение текстуры, низкое значение периметра, низкое значение площади, низкое значение компактности, низкое значение вогнутости, низкое значение вогнутых точек} => {доброкачественная опухоль} выражает влияние признаков на доброкачественность опухоли.

5 Правило {низкое значение радиуса, низкое значение текстуры, низкое значение компактности, низкое значение вогнутых точек} => {доброкачественная опухоль} значит, что сочетание радиуса, текстуры, компактности и количества вогнутых частей контура с их значениями чаще всего гарантирует отсутствие злокачественности.

Значения доверия ассоциативных правил близко к единице, когда как значения поддержки выше в сравнении со злокачественными случаями. На это влияет большее количество доброкачественных диагнозов. Такого вида правила будут встречаться в тридцати процентах случаев.

На Рисунке 4 видно, что на доброкачественность опухоли одновременно влияют несколько признаков. Из Рисунка 5 понятно, что у правил с заданными порогами высокая достоверность и положительная корреляция признаков.

Как и в случае со злокачественной ситуацией, попробуем поискать правила с хотя бы одним средним значением параметра. Значение порога поддержки уменьшится до 0,2, доверия – до 0,5. Также определим ограничение на максимальную длину, равную 4.

```
rules_b <- apriori(breastc_trans,
   parameter = list(support = 0.2, confidence = 0.5, minlen = 2, maxlen = 4),
   appearance = list(rhs = "diagnosis=B", default = "lhs"))
rules_b_med <- subset(rules_b, subset = lhs %pin% "Medium")
inspect(head(sort(rules_b_med, by = "confidence"), 5))
```

```
    lhs                                                    rhs           support   confidence coverage
[1] {texture=Low, perimeter=Low, smoothness=Medium}      => {diagnosis=B} 0.2302285 0.9424460  0.2442882
[2] {radius=Low, texture=Low, smoothness=Medium}         => {diagnosis=B} 0.2284710 0.9420290  0.2425308
[3] {texture=Low, smoothness=Medium, concave_points=Low} => {diagnosis=B} 0.2530756 0.9411765  0.2688928
[4] {radius=Low, compactness=Low, symmetry=Medium}       => {diagnosis=B} 0.2214411 0.9402985  0.2355009
[5] {perimeter=Low, compactness=Low, symmetry=Medium}    => {diagnosis=B} 0.2214411 0.9402985  0.2355009
```

*Рисунок 6.* Ассоциативные правила с доброкачественным диагнозом в правой части и хотя бы одним средним показателем в левой части

Получилось пять правил.

1 Правило {низкое значение текстуры, низкое значение периметра, среднее значение гладкости} => {доброкачественная опухоль} со средним показателем гладкости отражает влияние таких характеристик, как текстура, периметр, гладкость, и их значений на благоприятный диагноз.

2 Правило {низкое значение радиуса, низкое значение текстуры, среднее значение гладкости} => {доброкачественная опухоль} похоже на предыдущее и показывает воздействие признаков из левой части на правую.

3 Правило {низкое значение текстуры, среднее значение гладкости, низкое значение вогнутых точек} => {доброкачественная опухоль} значит, что полученные величины гладкости, стандартного отклонения значений по серой шкале и количества вогнутых точек влияют на качество образования.

4 Правило {низкое значение радиуса, низкое значение компактности, среднее значение симметрии} => {доброкачественная опухоль} содержит другую характеристику со средним значением – симметрию. При совпадении признаков из правила, высока вероятность, что опухоль окажется доброкачественной.

5 Правило {низкое значение периметра, низкое значение компактности, среднее значение симметрии} => {доброкачественная опухоль} подразумевает благоприятный диагноз в большинстве случаев при наблюдении указанных параметров.

Из полученных выше правил только характеристики симметрия и гладкость принимают средние значения, а остальные низкое. Вполне возможно, что при использовании других значений поддержки и доверия было бы выделено большее количество правил и результат получился бы более разнообразным. Видно, что у данных правил достаточно высокая вероятность появления, согласно которой они встречаются в каждой пятой записи. Их достоверность ожидаемо ниже, хотя всё ещё достаточно высока.

**Заключение.** Таким образом, в статье была рассмотрена реализация алгоритма Apriori на языке R для анализа медицинского набора данных Breast Cancer Wisconsin (Diagnostic) Data Set с численными значениями, которые были преобразованы в категориальные для дальнейшей работы.

Преобразование данных в подобных наборах необходимо для корректной работы алгоритма. Иначе Apriori будет искать правила с аналогичными числовыми значениями в записях. Пороги будут минимальными, потому что точное совпадение величин площади, периметра или любого другого признака маловероятно. С преобразованием, где числовые значения разбиты на три категории, легко выявить ассоциативные правила, полезные для анализа взаимосвязей характеристик в медицинских данных.

Применение алгоритма Apriori в медицинской сфере демонстрирует его потенциал в выявлении скрытых закономерностей, которые можно использовать для улучшения диагностики, прогнозирования и персонализации лечения. Важно учитывать, что точность и интерпретируемость полученных результатов во многом зависят от качества исходных данных и параметров алгоритма – поддержки и доверия.

Найденные ассоциативные правила могут помочь выявить, что, например, средние и высокие значения характеристик коррелируют с постановкой злокачественного диагноза, тогда как низкие значения чаще указывают на доброкачественную природу образования.

Если диагнозы одного вида превалируют над другими, полезно сделать нахождение ассоциативных правил для каждого случая отдельно. Для исследования конкретных случаев с определённой группой признаков можно выполнить сортировку по необходимым критериям.

Другие результаты исследований с помощью алгоритма Apriori можно найти в статьях [3-5].

## Список литературы

[1] Интернет ресурс «Kaggle» [Электронный ресурс] / Breast Cancer Wisconsin (Diagnostic) Data Set. – 2017. – Режим доступа: https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data. – Дата доступа: 17.03.2025.

[2] Street, W. N. Nuclear feature extraction for breast tumor diagnosis / W. N. Street, W. H. Wolberg, O. L. Mangasarian // Electronic imaging. - 1993. - С. 3- 5.

[3] Лянцевич, В. С. Использование алгоритма Apriori для анализа данных выбросов загрязняющих веществ в атмосферу / В. С. Лянцевич; науч. рук. Н. В. Марковская // Студенческий вестник. – 2024. – Т. Ч.7. – № 42(328). – С. 66-67.

[4] Лянцевич, В. С. История создания и разработка алгоритма Apriori / В. С. Лянцевич; науч. рук. Н. В. Марковская // Молодежь и наука XXI века: реалии и перспективы: Междунар. науч.-практ. конф., Астана, 12 апреля 2024 г. – Астана: ИПЦ; Esil University; 2024. – С. 771-772.

[5] Лянцевич, В. С. Библиотеки языка Python для анализа данных с помощью алгоритма Apriori / В. С. Лянцевич; науч. рук. Н. В. Марковская // От Альфа к Омеге...: сб. материалов XIV Респ. науч.-практ. конф.-конкурса науч.-исслед. работ учащихся средних, средних спец. учебных заведений и студентов вузов; От Альфа к Омеге...; / Гродн. гос. ун-т им. Янки Купалы; гл. ред. А. В. Кузьмич; редкол.: А. В. Кузьмич [и др.]. – Гродно: ГрГУ им. Янки Купалы, 2024. – С. 71-72.

**Авторский вклад**

**Лянцевич Валерия Сергеевна** – выполнение исследования, написание программного алгоритма, анализ полученных результатов.

**Марковская Наталья Вацлавовна** – постановка задачи исследования, анализ полученных результатов.

# MEDICAL DATA ANALYSIS USING THE APRIORI ALGORITHM IN R

***V.S. Lyantsevich***
*Student of the Department of Mathematical and Information Support of Economic Systems, YKSUG*

***N.V. Markovskaya***
*Associate professor of the Department of Mathematical and Information Support of Economic Systems, YKSUG, PhD of Mathematical Sciences, Associate Professor*

**Abstract.** The article considers the use of the Apriori algorithm for identifying association rules in a medical breast tumor dataset. The features of data structures are analyzed, they are modified for the application of the algorithm, and the implementation of Apriori is assessed. The study is based on the transformed dataset, key patterns are identified, and the influence of the algorithm parameters on the quality of the extracted rules is analyzed. Conclusions are made about the applicability of the Apriori algorithm in the tasks of analyzing medical data, and recommendations are offered for its configuration for various types of data.

**Key words:** association rules, Apriori algorithm, medical data, breast tumor.