|  |
| --- |
| МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ |
| ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ |
| **«МОСКОВСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»** |
| **(МОСКОВСКИЙ ПОЛИТЕХ)** |
| Факультет информационных технологий |

Кафедра «Прикладная информатика»

Форма обучения: очная

**ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА**

по направлению 01.03.02 «Прикладная математика и информатика»

на тему «Прогнозирование материнского поведения в регионе по данным опросов на основе технологий больших данных»

Студентка Олеся Михайловна Репина

Руководитель работы

доцент, к.э.н. Александр Евгеньевич Рабинович

ДОПУСКАЕТСЯ К ЗАЩИТЕ

Заведующий кафедрой

профессор, к.э.н. Станислав Вадимович Суворов

МОСКВА 2023

|  |  |
| --- | --- |
|  | **УТВЕРЖДАЮ**  **Зав. кафедрой «Прикладная информатика»**  **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_С. В. Суворов** |

**ЗАДАНИЕ**

**На выпускную квалификационную работу (ВКР)**

**Студента** Репиной Олеси Михайловны  **группы** 191–341

1. **Тема:** «Прогнозирование материнского поведения в регионе по данным опросов на основе технологий больших данных»
2. **Утверждена** приказом **Московского политехнического университета №1759-УД** от **04.05.2023**
3. **Исходные данные к работе:** Данные анкетирования матерей в регионах США за 2011 год из системы наблюдения за поведением, отношением и переживаниями матерей до, во время и вскоре после беременности
4. **Содержание ВКР**

|  |  |  |
| --- | --- | --- |
| № п/п | Наименование раздела | Содержание раздела |
| 1 | Аналитическая часть | Анализ исследований материнского поведения |
| Источники информации для прогнозирования материнского поведения |
| Примеры применения методов машинного обучения для прогнозирования по опросам |
| Обзор набора данных |
| Предварительная постановка цели и задач |
| 2 | Теоретическая часть | Выбор технологии для прогнозирования по опросам |
| Обзор и выбор методов |
| Инструменты для реализации методов |
| Математическая постановка задачи |
| 3 | Проектная часть | Анализ набора данных и его подготовка для работы |
| Выявление причинно-следственных связей материнского поведения в штате США |
| Применение прогнозирования материнского поведения для сравнения матерей двух штатов |
| Возможности использования прогнозирования материнского поведения в различных сферах |

1. **Календарный график выполнения ВКР**

|  |  |  |
| --- | --- | --- |
| № п/п | Наименование раздела | Дата проведения консультаций |
| 1 | Аналитическая часть | 13.05.2023 |
| 2 | Теоретическая часть | 20.05.2023 |
| 3 | Проектная часть | 27.05.2023 |
| 4 | Оформление ВКР | 12.06.2023 |
| 5 | Подписание ВКР | 15.06.2023 |

1. **Срок сдачи студентом законченной работы 15.06.2023**

|  |  |
| --- | --- |
| Задание выдал 04.05.2023 | Задание получил 04.05.2023 |
| Руководитель \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ | Студент \_\_\_\_\_\_\_\_\_\_\_\_\_\_ |

**АННОТАЦИЯ**

Тема выпускной квалификационной работы: «Прогнозирование материнского поведения в регионе по данным опросов на основе технологий больших данных».

Работа содержит 94 страницы, 16 формул, 18 рисунков, 13 таблиц, 5 приложений и 28 источников.

Цель ВКР: выявление причин, приводящих к ухудшению состояния матери, и факторов, положительно влияющих на материнское здоровье, с помощью модели машинного обучения для прогнозирования материнского поведения.

Данная работа состоит из трёх частей:

В аналитической части был произведён анализ исследований материнского поведения и рассмотрены методы, позволяющие осуществлять прогнозирование по источнику данных для датасета — опросам.

В теоретической части были выбраны методы для реализации поставленных задач и рассмотрен набор инструментов для реализации: библиотеки pandas, NumPy, scikit-learn, matplotlib, graphviz и среда разработки Jupyter. Также была произведена математическая постановка задачи.

В проектной части была произведена реализация задач кластеризации, классификации и поиска ассоциативных правил на анализируемом наборе данных.

Ключевые слова: ПРОГНОЗИРОВАНИЕ ПО ОПРОСАМ, МАТЕРИНСКОЕ ПОВЕДЕНИЕ, КЛАСТЕРИЗАЦИЯ, КЛАССИФИКАЦИЯ, АССОЦИАТИВНЫЕ ПРАВИЛА, PYTHON, SCIKIT-LEARN.

**СОДЕРЖАНИЕ**

[ВВЕДЕНИЕ 7](#_Toc21821)

[1 АНАЛИТИЧЕСКАЯ ЧАСТЬ 9](#_Toc26254)

[1.1 Анализ исследований материнского поведения 9](#_Toc24538)

[1.2 Источники информации для прогнозирования материнского поведения 13](#_Toc31019)

[1.3 Примеры применения методов машинного обучения для прогнозирования по опросам 16](#_Toc14026)

[1.4 Обзор набора данных 22](#_Toc21926)

[1.5 Предварительная постановка цели и задач 25](#_Toc21655)

[2 ТЕОРЕТИЧЕСКАЯ ЧАСТЬ 26](#_Toc23846)

[2.1 Выбор технологии для прогнозирования по опросам 26](#_Toc11144)

[2.2 Обзор и выбор методов для прогнозирования материнского поведения 28](#_Toc24485)

[2.3 Инструменты для реализации методов 42](#_Toc24001)

[2.4 Математическая постановка задачи 47](#_Toc9583)

[3 ПРОЕКТНАЯ ЧАСТЬ 49](#_Toc6665)

[3.1 Анализ набора данных и его подготовка для работы 49](#_Toc32551)

[3.2 Выявление причинно-следственных связей материнского поведения в штате Колорадо 51](#_Toc15418)

[3.3 Применение прогнозирования материнского поведения для сравнения матерей двух штатов 62](#_Toc12513)

[3.4 Возможности использования прогнозирования материнского поведения в различных сферах 73](#_Toc4207)

[ЗАКЛЮЧЕНИЕ 75](#_Toc14161)

[СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ 77](#_Toc20729)

[ПРИЛОЖЕНИЕ А 81](#_Toc31812)

[ПРИЛОЖЕНИЕ Б 88](#_Toc26636)

[ПРИЛОЖЕНИЕ В 90](#_Toc6254)

[ПРИЛОЖЕНИЕ Г 91](#_Toc13834)

[ПРИЛОЖЕНИЕ Д 92](#_Toc29470)

# ВВЕДЕНИЕ

Ежегодно на свет появляется более 100 миллионов человек, у каждого из них есть мать. Мать имеет непосредственное влияние на развитие ребёнка, который в дальнейшем становится полноправным членом общества. Отношение матери к ребёнку во многом зависит от её здоровья и привычек.

В период беременности и после родов у матери ослаблен иммунитет, а также проявляется повышенная чувствительность нервной системы к различного рода возбудителям, как хороших, так и плохих. Поэтому необходимо более основательно подходить к вопросам диагностики отклонений здоровья женщин в этот период.

Проведение профилактических мероприятий и мониторинг здоровья матери могут быть положительно сказаны на состоянии как самой матери, так и её детей. Проблема состоит в определении негативных факторов, профилактика которых должна производиться, и причин возникновения этих факторов, а также определении факторов, положительно влияющих на состояние матери.

Таким образом, использование технологий интеллектуального анализа данных для прогнозирования материнского поведения имеет ценность как для общества, так и для самой матери.

Цель работы — выявление причин, приводящих к ухудшению состояния матери, и факторов, положительно влияющих на материнское здоровье, с помощью модели машинного обучения для прогнозирования материнского поведения.

Объектом исследования является материнское поведение.

Предметом исследования являются данные анкетирования матерей в регионах США в 2011 году, представляющие собой выборку, состоящую из 200 столбцов и более 500 тысяч строк.

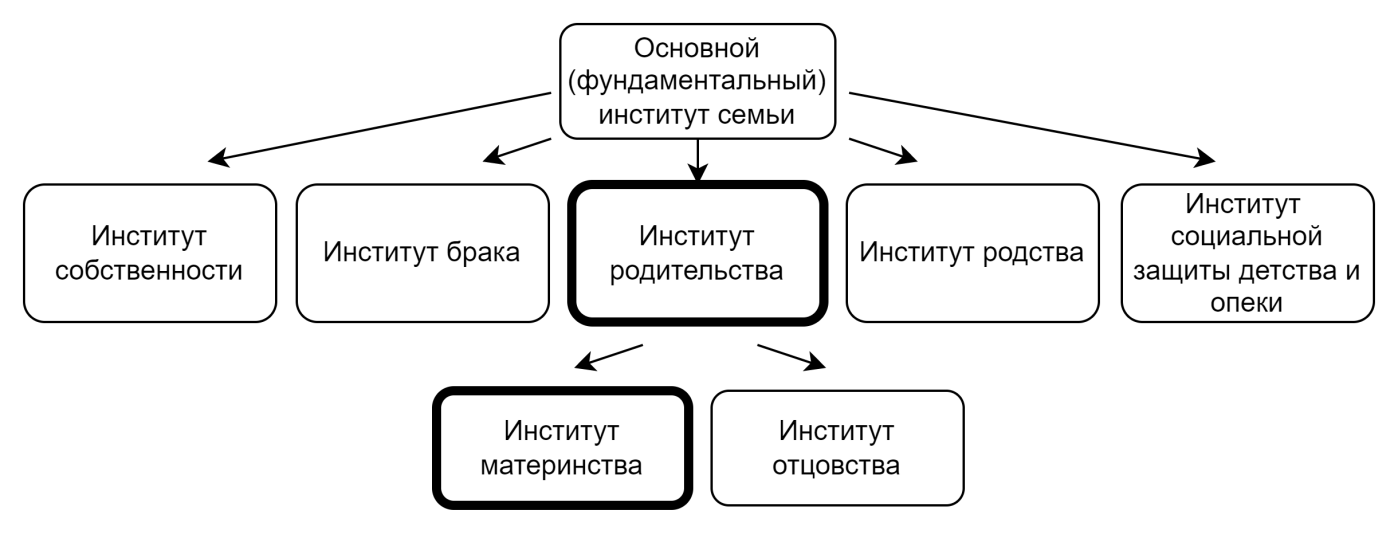
В данной работе планируется использование таких методов машинного обучения, как кластеризация построение деревьев решений и поиск ассоциативных правил.

Источник данных — система мониторинга оценки риска беременности (PRAMS) США.

# АНАЛИТИЧЕСКАЯ ЧАСТЬ

## Анализ исследований материнского поведения

Институт материнства — часть частного института родительства, который является частью фундаментального института семьи. Наглядно положение института материнства в контексте института семьи представлено на схеме на рисунке 1.1.



1. Положение института материнства в контексте института семьи

Формирование института семьи началось в V—IV веке до н.э. мыслителями Платоном и Геродотом. Тогда внимание уделялось отношениям мужчины и женщины.

Первые исследования родительства были историческими. Проводили их в первой половине XX века в основном женщины американского гражданства: Маргарет Мид (антрополог), Хелен Дойч (психолог), Карен Хорни (психолог), Нэнси Чодороу (социолог).

Также большой вклад в исследование материнства и детства внесли французские исследователи: Филипп Арьес (историк) и Элизабет Бадинтер (историк и философ). Многие труды были написаны либо феминистками, либо отражали их точку зрения о навязывании женщине образа матери и о том, что материнство — главная её социальная функция [1].

Основными источниками информации для многих работ первых исследованиях были эго-документы. Эго-документ представляет собой текст, созданный автором для него самого или близких ему людей, и не подразумевает публичного показа.

Изучение материнства в психологии и социологии началось во второй половине XX века, когда началось разделение толкования термина «материнство» на два направления: материнство как социальный институт (motherhood) и материнство как субъективное восприятие (mothering).

В 90-е годы прошлого столетия проблема материнства и детства привлекла внимание властей в США, Великобритании и Канаде. В связи с этим интерес исследователей начал смещаться с изучения «motherhood» на раскрытие «mothering». К 2010 году трактовка термина «материнство» разделилась окончательно.

Под материнством как социальным институтом (motherhood) подразумевается изучение следующих характеристик:

• официальный статус матери;

• права и обязанности матери;

• отношение государства и общества к материнской роли;

• система социальной защиты матерей и т.д.

Субъективное восприятие (mothering) включает в себя:

• повседневные материнские практики:

- грудное вскармливание;

- гигиена ребёнка и матери во время беременности;

- воспитание;

- времяпрепровождение ребёнка в разлуке с матерью (когда женщине нужно на работу или по бытовым делам) и т.д.

• эмоциональные переживания матерей;

• отношение женщин к роли матери;

• реализацию женщины посредством рождения ребёнка [1,2].

Исследования материнского поведения зачастую имеют немало противоречий. Например, в 2012 году в журнале “Child Abuse & Neglect” было опубликовано исследование о равнодушии матерей, переживших опыт сексуального насилия до 21 года, к собственным детям в возрасте до 18 месяцев. Солидарная точка зрения была выражена в 2017 году в журнале “Infant Mental Health Journal”. Также в этих работах утверждается, что при наличии опыта физического насилия мать может проявлять враждебность по отношению к ребёнку возраста полутора лет.

В противовес упомянутым выше исследованиям в 2015 году исследование Анны Фукс показало, что до пяти месяцев эмоциональная отзывчивость матери, имевшей опыт насилия, по отношению к ребёнку не отличается от матерей, не испытывающих подобного. В данной работе рассматривались матери с опытом и физического, и сексуального насилия [3].

Разногласия возникают не только при исследованиях в направлении motherhood, но и в направлении mothering. Так в работах 2015 года исследование ценности семьи у матерей дошкольников и бездетных женщин показало противоречивые результаты. В научном журнале «Вестник Томского государственного педагогического университета» утверждается, что женщины, имеющие детей дошкольного возраста, больше ценят семью и нравственные качества. Респондентами являлись 2 группы женщин по 25 человек с детьми и бездетные с высшим или неполным высшим образованием.

Иные выводы были опубликованы в журнале «Педагогическое образование в России». Данные показали, что у 74% матерей дошкольников отдают больше предпочтения обучению, хобби и общественной деятельности нежели семье. Выборку составляло 80 матерей дошкольников (3-7 лет) и детей до 3 лет преимущественно с высшим образованием и средним возрастом равным 28 годам. В обоих исследованиях методом сбора данных был опрос [4].

Изучение материнского поведения активно производится не только в области психологии и социологии, но и в сфере биологии. В 2019 году в журнале “Psychoneuroendocrinology” опубликовано исследование о влиянии неблагоприятных детских переживаний матери на преждевременное старение плаценты. Было выявлено, что негативные события в детстве могут быть связаны с преждевременными родами и различными осложнениями беременности, многие из которых имеют независимую связь с физиологией и функцией плаценты. Информация о детской тревоге и стрессе была получена из данных опросов матерей. Для изучения было взято 268 образцов у 67 плацент [5].

В текущем (2023) году в марте в журнале “Scientific Reports” было обнародовано исследование выявившее, что у матерей по сравнению с не матерями более отзывчивая, гибкая и эффективная система регуляции эмоций. Учёные пришли к выводу, что материнский мозг более эффективен, гибок и отзывчив в состоянии покоя, чем мозг женщин, которые никогда не были беременны. Различия в мозговой активности распространяются на различные формы поведения в послеродовой период, включая те, которые связаны с материнским уходом, а также на не материнское поведение, то есть познание и социальное познание. В исследовании принимали участие 40 матерей, родивших впервые, у которых прошёл год после родов, и 39 женщин, которые никогда не были беременными. В обеих группах женщины были одного возраста и образования [6].

В области больших данных довольно мало научных работ, связанных с анализом поведения матерей. В 2022 году на 6-ой международной конференции по вычислительным методологиям и коммуникации представитель из Ордосского технологического института (Китай) выступил с докладом на тему разработки системы сбора и интеллектуального анализа данных о корреляции между послеродовой депрессией и социальной поддержкой у женщин в период внутриутробного развития. В докладе отмечено, что частота послеродовой депрессии у матерей, родивших второго ребёнка, относительно высока и социальная поддержка может способствовать улучшению физическому и психическому здоровью [7].

Из представленных выше научных работ можно выделить ряд проблем, которые возникают при исследовании материнского поведения:

• небольшой размер выборки. В социологических опросах и биологических экспериментах количество участников обычно не превышает 100 человек.

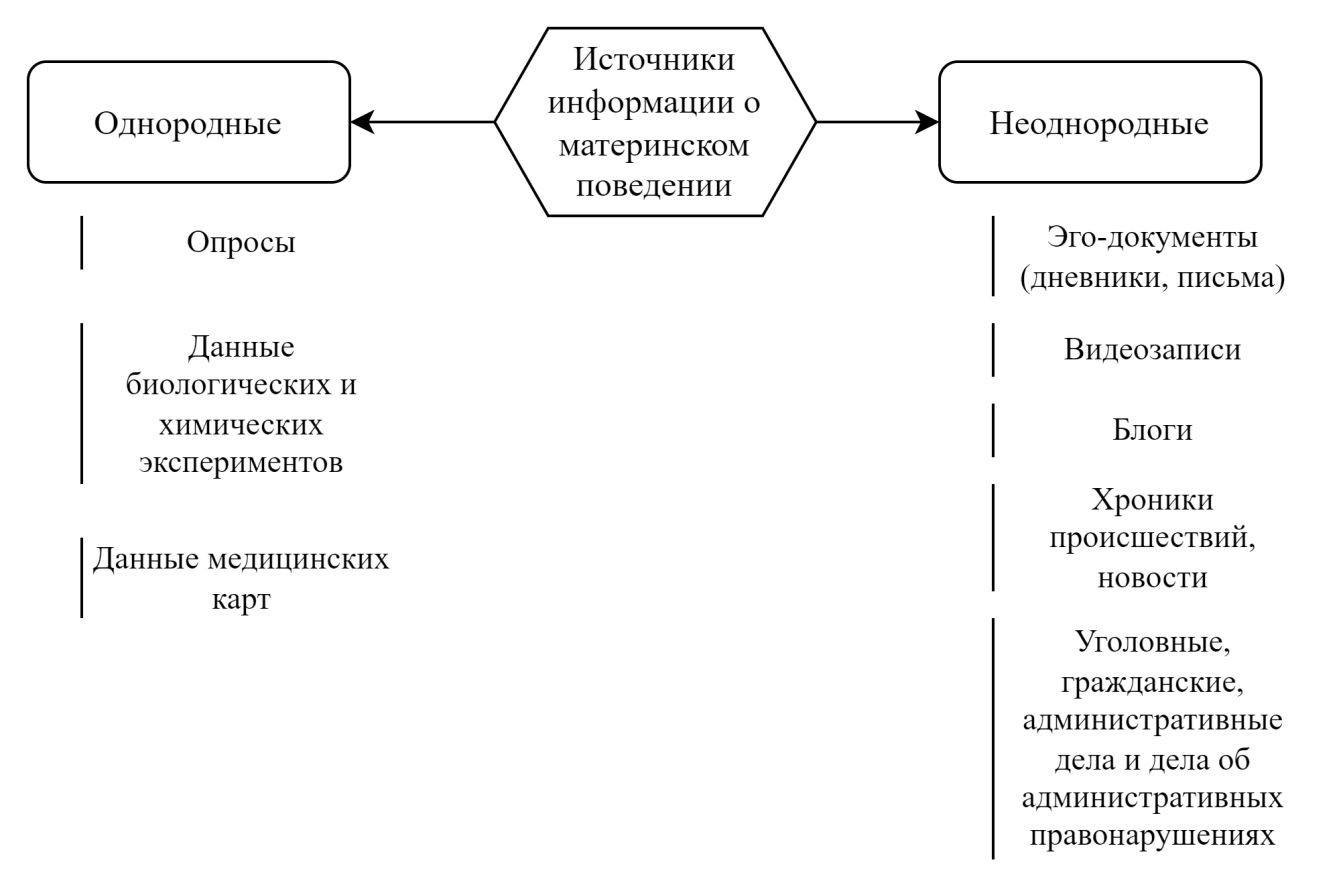
• однородность участников конкретного исследования. Как правило участниками исследований являются женщины одного возраста, образования и региона проживания. Сами исследователи также часто отмечают схожесть участников.

• неоднородность участников разных исследований. Для каждого опроса или эксперимента отбираются определённые участники (люди одного возраста, образования или с наличием определённых вредных привычек), из-за чего возникают противоречия в результатах исследований.

• субъективная оценка. Большинство исследований используют опросы в качестве метода сбора информации. Во многих биологических исследованиях точные показания состояния организма сопоставляют с данными опросов участников.

## Источники информации для прогнозирования материнского поведения

Материнское поведение исследуют во всех областях науки, поэтому источников информации можно выделить много. По структуре содержания источники о психическом и физическом состоянии матерей можно разделить на однородные, для обработки которых можно использовать алгоритмы обработки данных, и неоднородные, содержание которых анализируется в частном порядке. Однородные источники отличаются тем, что одни документы имеют одинаковую структуру с другими. Это позволяет получить данные в единой унифицированной форме. Примеры однородных и неоднородных источников наглядно представлены в виде схемы на рисунке 1.2.



1. Источники информации о материнском поведении

Как видно из схемы, неоднородные источники либо имеют небольшую выборку относительно общего исследования, либо отражают исключительные случаи. Не все матери ведут блог, дневник или совершают преступления. Однородные источники позволяют охватить больший круг женщин и получить более точные данные. Проведение экспериментов в больших масштабах весьма затруднительно и затратно. Данные медицинских карт в большинстве своём содержат общую информацию о физическом состоянии матери. Опросы — это довольно простой и удобный способ сбора данных и о физическом, и о психическом состоянии матери, а проведение их в большом количестве значительно менее затратно, чем химические или биологические эксперименты, для которых необходимо специальное оборудование и условия.

В социологии опросы разделяют на две основных группы: интервью и само-заполнение анкеты.

При проведении интервью обязательным условием является контакт интервьюера с респондентом. Вопросы могут задаваться лицом к лицу при очной встрече, по телефону, с помощью аудио или видео связи через интернет.

Само-заполнение анкеты не всегда требует наличия контакта между интервьюером и респондентом. От интервью оно отличается тем, что ответы на вопросы заполняются самими опрашиваемы, когда при проведении интервью ответы записывает сам интервьюер.

Выделяют несколько видов само-заполнения анкет:

• раздаточное анкетирование подразумевает, что анкета раздаётся респондентам в определённом месте, затем спустя определённый промежуток времени заполненный опросник собирают оговоренным образом;

• во время проведения аудиторного анкетирования опросный лист раздаётся в помещении, интервьюер рассказывает о правилах заполнения и отвечает на вопросы респондентов, затем собирает заполненные бланки и проверяет полноту ответов;

• онлайн опрос моно проводить двумя способами:

- опрос веб-панели. В данном случае опрос проводят несколько раз через некоторые промежутки времени с помощью одной методики и одной выборки (набора респондентов) - панели.

- опрос по объявлениям на сайтах со ссылкой на анкету. При таком способе задать необходимые требованиям к характеристикам респондентов трудно, так как перейти по ссылке может любой желающий. Можно включить вопросы о требуемых характеристиках в анкету и произвести фильтрацию, однако есть риск не получить выборку даже минимально необходимого размера [8].

У метода само-заполнения анкет есть неоспоримые преимущества:

— возможность проведения исследования с большой выборкой;

— скорость получения результата;

— экономия трудовых и материальных затрат при подготовке, проведении и обработке;

— отсутствие предубеждений по отношению к опрашиваемому;

— отсутствие необходимости в обучении и поиске интервьюера.

## Примеры применения методов машинного обучения для прогнозирования по опросам

Зачастую для анализа социологических опросов используют методы многомерной статистики, которые требуют наличия нормального распределения в наборе данных или наличие линейной зависимости между переменными.

Методы обработки больших данных позволяют снять наложенные на данные ограничения, что является их главным преимуществом.

Наиболее часто для быстрого анализа результатов анкетирования используют простые визуализации в виде круговых, столбчатых диаграмм и графиков.

Среди методов математической статистики наиболее часто используются следующие:

• расчёт показателей описательной статистики:

-среднее значение (среднее арифметическое, медиана, мода);

- дисперсия;

- стандартное среднеквадратическое отклонение;

- доверительный интервал

• ранжирование;

• шкалирование

• выявление корреляционной зависимости между различными показателями.

Классические методы довольно просты в реализации и для их использования не нужно дополнительное программное обеспечение. Данные, полученные с помощью этих методов легко интерпретировать.

Метод кросс-табуляции заключается в построении таблицы распределения значений зависимой переменной при каждом значении независимой. Проще всего принцип кросс-табуляции рассмотреть на примере.

Анкета опроса кофейни содержала вопросы о количестве выпитых за день чешек кофе с вариантами ответа: 1-2, 3-5, больше 5, а также о поле (мужской или женский).

В результате анализа и подсчётов полученных данных была получена таблица кросс-табуляции, представленная в таблице 1.1.

1. Таблица кросс-табуляции об употреблении кофе различным полом

|  |  |  |  |
| --- | --- | --- | --- |
| Количество выпитого кофе | Пол | | Всего |
| мужской | женский |
| 1-2 чашки | 18 | 36 | 54 |
| 3-5 чашек | 47 | 57 | 104 |
| Больше 5 чашек | 25 | 28 | 53 |
| Всего | 90 | 121 | 211 |

В каждой ячейке таблицы 1 показано количество респондентов, имеющих определённые признаки. Например, 47 опрошенных являются мужчинами, которые выпивают в день от 3 до 5 чашек кофе.

Проводить анализ представленных выше данных довольно затруднительно. Имеет смысл исключить влияние численности респондентов разного пола и рассчитать доли участников анкетирования, потребляющих различное количество кофе, в процентах от общего числа для каждого пола. Полученные данные представлены в таблице 1.2.

1. Таблица кросс-табуляции об употреблении кофе различным полом с исключением влияния численности респондентов разного пола

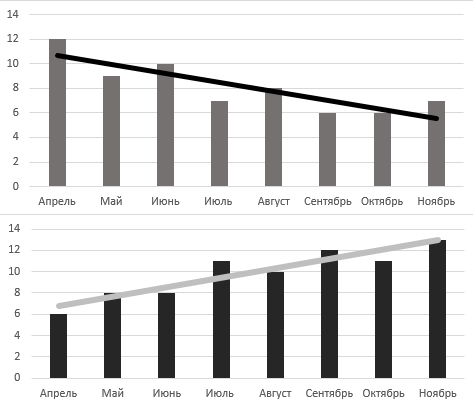
|  |  |  |  |
| --- | --- | --- | --- |
| Количество выпитого кофе | Пол | | Всего |
| мужской | женский |
| 1-2 чашки | 20 | 30 | 26 |
| 3-5 чашек | 52 | 47 | 49 |
| Больше 5 чашек | 28 | 23 | 25 |
| Всего | 100 | 100 | 100 |

Таблица 1.2 показывает, что мужчины больше пьют кофе нежели женщины.

Кросс-табуляция даёт возможность оценить корреляцию между переменными. Она помогает избежать путаницы, которая появляется при просмотре необработанных данных.

Суть метода анализа тенденций заключается в построении графика агрегированных данных об ответах за промежуток времени. Для реализации этого метода опрос должен проводиться несколько раз с определённой периодичностью.

Например, имеются данные за несколько месяцев ответов на вопрос «Пьёте ли вы кофе?» с вариантами ответов да или нет. Также за те же месяцы имеются данные на вопрос «Пьёте ли вы молоко?». Для анализа данных представим их виде столбчатых диаграмм и добавим на них линию тренда (рисунок 1.3).



1. График употребления респондентами кофе (сверху) и молока (снизу)

На рисунке 1.3 видим, что больше респондентов стали употреблять молоко и меньше кофе.

Анализ тенденций путём построения графика и добавления на него линии тренда позволяет оценить динамику изменения конкретного показателя.

Нейронные сети позволяют решать большое количество задач:

• распознавание текстов;

• контекстная реклама;

• игра на бирже;

• фильтрация спама;

• проверка проведения подозрительных операций по банковским картам.

В последнее время идёт активное внедрение нейронных сетей в инструментарий социологов. Их используют для обработки и подготовки исходных данных. В частности, многослойный персептрон с обратным распространением ошибки показал свою эффективность в заполнении отсутствующих данных [9].

В 2020 году для формирования групп респондентов в зависимости от их мотивации к участию в благотворительности был использован аппарат самоорганизующихся карт признаков Кохонена [10]. В этой работе база данных с результатами опросов содержала более 140 вопросов различного вида (категориальные, логические, интервальные, порядковые) и более 5000 записей.

На первом этапе решалась задача кластеризации с использованием самоорганизующихся карт Кохонена, в результате решения которой было выделено 6 кластеров. Описание каждого из кластеров производилось с помощью анализа карт Кохонена.

Совместно с картами Кохонена был использован аппарат деревьев классификации для формулировки классификационных правил. Построение дерева производилось с помощью метода CART. Условием прекращения ветвления была прямая останова по методу FACT.

В 2022 году в качестве модели нейронной сети для системы помощи абитуриентам в выборе специальности использовался многослойный персептрон [11]. Опрос для разработки системы состоял из 20 пунктов. На него ответили 309 абитуриентов. В результате работы удалось добиться точности в 92 %.

Эффективность в анализе результатов опросов показал метод дерево решений. У этого метода есть несколько преимуществ:

• быстрая и ясная интерпретация правил классификации;

• лёгкая визуализация;

• можно сделать прогноз для конкретного субъекта (путь в дереве);

• скорость;

• поддержка как категориальных, так и числовых признаков.

В 2021 году метод дерева решений был применён для анализа поведения школьников при выборе вуза [12]. Тогда Институтом социально-экономических исследований был проведён опрос учащихся 10 и 11 классов образовательных учреждений республики Башкортостан. В опросе приняли участие 6126 человек. Анкета состояла из нескольких блоков, касающихся личных характеристик школьника, причин выбора учебного заведения для дальнейшего обучения и причин миграции. В качестве входных параметров использовались пол, уровень знаний и выбранные предметы для сдачи ЕГЭ. В результате работы с данными было получено дерево решений из 29 правил.

В том же году было произведено исследование взаимосвязи между восприятием населением коррупции и уровнем социального доверия c помощью дерева решений [13]. Изучались данные социологического опроса в Алтайском крае. Было опрошено 1200 респондентов возрастом от 18 до 70 лет. Оценка генерализованного и межличностного доверия осуществлялась по десятибалльной шкале. В результате применения дерева решений были получены данные о наиболее значимых факторах, влияющих на негативную оценку роли коррупции на различные сферы общества.

Существуют различные области применения ассоциативных правил:

- анализ рыночной корзины;

- представление рекомендованных покупок в интернет-магазинах и т.п.;

- поиск ошибок в базах данных

- медицинская диагностика;

- анализ белковых последовательностей;

- анализ данных переписи населения;

- анализ рождаемости;

- анализ погодных явлений;

- анализ показателей жизнедеятельности человека (по фитнес-трекерам и т.п.).

В 2019 году на конференции «Минские научные чтения — 2019» был представлен доклад, описывающий опыт разработки ассоциативных правил в области студенческого чтения [14]. В анкету входили вопросы, связанные с предпочитаемыми жанрами, отношением к классической и массовой литературе, целями чтения, отношением к чтению. Выборка состояла из 149 респондентов в возрасте от 19 до 21 года. В результате было сформировано 6 ассоциативных правил с уровнем поддержки более 50 % и о достоверностью более 80 %.

В 2020 году ассоциативные правила был одним из методов анализа цветовых предпочтений [15]. В ходе опроса 50 испытуемым показывали цветовую палитру из 27 оттенков и просили выбрать из неё те цвета, которые по их мнению, лучше всего подходили для интерьера каждого из типов помещений. Список помещений был следующим:

• гостиная;

• прихожая;

• спальня;

• ванна;

• туалет;

• кухня;

• коридор.

В итоге было получено 18 ассоциативных правил, позволяющих выявить цветовые предпочтения респондентов.

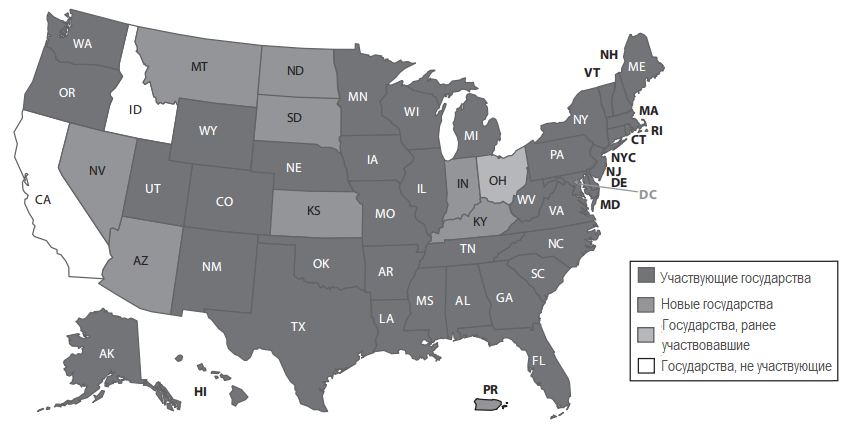
## Обзор набора данных

Данные для анализа были опубликованы Центром по контролю и профилактике заболеваний (CDC) на платформе Kaggle. Эти данные были получены с помощью системы мониторинга оценки риска беременности (PRAMS) — постоянно действующей государственной системы наблюдения за поведением, отношением и переживаниями матерей до, во время и вскоре после беременности.

PRAMS собирает данные с помощью опросов в смешанном режиме: по почте и телефону. Ежегодные размеры выборки по штату варьируются примерно от 1000 до 3000 женщин.

Система PRAMS является частью Центров по контролю и профилактике заболеваний (CDC), призванных снизить младенческую смертность и низкий вес при рождении и содействовать безопасному материнству. Эта система была внедрена 1987 году в связи с высокими показателями младенческой смертности, преждевременных родов, а также смертей и заболеваний, связанных с беременностью.

С момента создания системы число участвующих государств и областей (на схеме именуемых государствами) увеличилось от 6 до 51, включая 47 штатов, округ Колумбия, Нью-Йорк, Пуэрто-Рико и Совет по здравоохранению председателя племени Великих равнин. Географическое распространение системы представлено на арте на рисунке 1.4. Данные актуальные на 2018 год.



1. География внедрения системы PRAMS

В настоящее время служба наблюдения PRAMS охватывает примерно 83% всех родов в США [16].

В настоящее время анкета PRAMS находится в восьмой версии, и её пересмотры происходят примерно каждые 3-5 лет.

Данные PRAMS часто используются для оценки программы общественного здравоохранения и политики как на государственном, так и на национальном уровнях. Например, данные PRAMS были использованы для оценки влияния наличия государственных страховых полисов на количество незапланированных и несвоевременных родов.

Многообещающие исследования также изучили возможность распространения PRAMS на отцов недавно родившихся младенцев, чтобы улучшить качество жизни, понять влияние отцов на здоровье матерей и младенцев, а также опыт перехода к отцовству.

Используемый набор данных — данные системы PRAMS за 2011 год, которые были опубликованы только в 2018 году. Опрос состоял из 196 вопросов, которые были заданы матерям не только во время ведения беременности, но и через несколько недель, 3 месяца, 6 месяцев и год после родов. Вопросы предполагали односложный или краткий ответ и охватывали следующие темы:

• жестокое обращение

• употребление алкоголя

• контрацепция

• грудное вскармливание

• психическое здоровье

• ожирение и др.

Основным способом сбора данных является отправка анкет по почте с последующим отслеживанием по телефону для респондентов, не получающих почту. В рамках каждого режима предпринимается несколько попыток связаться с отобранными женщинами. Телефонный контакт начинается в течение недели после последней рассылки опроса. На каждый рабочий телефонный номер совершается до 15 попыток дозвона, распределенных в разное время суток и в разные дни недели.

Информация о столбцах, из которых состоит набор данных, представлена в таблице 1.3.

1. Описание столбцов в наборе данных

|  |  |  |
| --- | --- | --- |
| **Название столбца** | **Описание столбца** | **Тип данных** |
| Year | Год проведения опроса. | Целое число |
| Location Abbr | Аббревиатура обследуемого штата или местоположения. | Строка |
| Location Desc | Назввание обследуемого штата или местоположения. | Строка |
| During any of your prenatal care visits did a doctor nurse or other health care worker talk with you about what to do if you feel depressed during pregnancy or after your baby is born? | Ответы на вопросы для опроса. (196 колонок). Полный список вопросов представлен в ПРИЛОЖЕНИИ А | Строка |
| … |
| Indicator of whether an HIV test was refused among women offered the test |
| Geolocation | Геолокация опроса. | Строка |

Рассмотренный набор данных позволяет минимизировать погрешности и противоречия, которые возникают во многих исследованиях, благодаря большому размеру выборки и отсутствию этапа отбора участников исследования.

## Предварительная постановка цели и задач

Необходимо разработать модель, позволяющую предсказать поведение матери по её ответам в анкете.

Цель работы — выявление причин, приводящих к ухудшению состояния матери, и факторов, положительно влияющих на материнское здоровье, с помощью модели машинного обучения для прогнозирования материнского поведения.

Достигнуть поставленной цели позволит выполнение следующих задач:

1. Проанализировать особенности набора данных;
2. Выбрать алгоритмы машинного обучения для построения модели;
3. Осуществить предварительную обработку данных, чтобы они были пригодны для работы выбранных алгоритмов;
4. Построить модель;
5. Определить наиболее значимые факторы;
6. Рассмотреть модель на предмет выявления причинно-следственных связей, позволяющих сделать достоверные выводы.

**Вывод**

В данной главе были рассмотрены прогнозирование материнского проведения в различных областях науки, источники данных для этих прогнозов, а также примеры применения методов, которые позволяют строить прогнозы по опросам (источнику данных для данного исследования).

# ТЕОРЕТИЧЕСКАЯ ЧАСТЬ

## Выбор технологии для прогнозирования по опросам

Статистический анализ и визуализация — самые простые и распространённые технологии для анализа опросов и прогнозирования по ним. Результаты таких исследований легко получить и интерпретировать. Суть этих технологий состоит в подсчёте некоторых значений данных по формулам и выявление в этих данных тенденций, сходств и закономерностей с помощью представления результатов в виде удобных и понятных графиков и схем.

Исследования с использованием статистического анализа и визуализации проводят в основном на небольших выборках, в которых мало параметров, в данном случае вопросов. Полученные результаты можно использовать для принятия решений без какой-либо дальнейшей обработки ввиду наглядности итогов.

В описанном ранее датасете опрос состоит из 196 вопросов. Полный список вопросов представлен в ПРИЛОЖЕНИИ А. При проведении статистического анализа для такого набора данных будет сложно сопоставлять показатели, интерпретировать результат и на его основании принимать решения.

В данном случае нужны более сложные технологии, такие как машинное обучение. Однако для толкования результатов работы алгоритмов машинного обучения необходимо использовать визуализацию.

Машинное обучение и визуализация лежат в основе мультидисциплинарной технологии Data Mining, суть которой состоит в поиске неочевидных, объективных и полезных закономерностей в данных большого объёма [17]. Эта технология используется для поддержки принятия решений.

Задачи Data Mining могут быть описательными или предсказательными. К описательным задачам обычно относятся поиск ассоциативных правил и кластеризация, к предсказательным — классификация и регрессия, отчасти – поиск правил. Основные задачи Data Mining представлены на рисунке 2.1.



1. Задачи Data Mining

В данной работе прогнозирование поведения матери заключается в предсказании ответа на вопрос, исходя из других ответов. Например, если женщина на вопрос «Осматривал ли вашего ребёнка врач, медсестра или другой поставщик медицинских услуг в течение первой недели после того, как он или она выписались из больницы?» — «да» (yes), то на вопрос «В течение вашей последней беременности причинял ли вам кто-либо физическую боль каким-либо образом?» она ответит «нет» (no). Для выявления подобных закономерностей используют поиск ассоциативных правил.

Отличительной особенностью ассоциации от других задач является то, что поиск взаимосвязей производится между происходящими одновременно событиями.

Матери — это люди с определённым набором характеристик, который отличается у каждого из них. Поэтому ассоциативные правила, полученные даже для отдельного штата, не будут достоверными. Для решения данной проблемы следует разделить этих женщин на несколько групп, то есть выполнить задачу кластеризации.

Кластеризация позволяет разбить объекты, матерей, с не предопределённым классом на группы, при этом выделять зависимую переменную не требуется. В результате кластеризации выделяются группы, которые имеют два признака — внутреннюю однородность и внешнюю изолированность.

Чтобы предсказывать поведение новых участников анкетирования, необходимо определить его в один из заранее определённых кластеров, что относится к задаче классификации.

Решение задачи классификации позволяет обнаружить признаки или, в данном случае, вопросы, которые характеризуют группу объектов исследуемого набора данных, то есть респондентов-матерей.

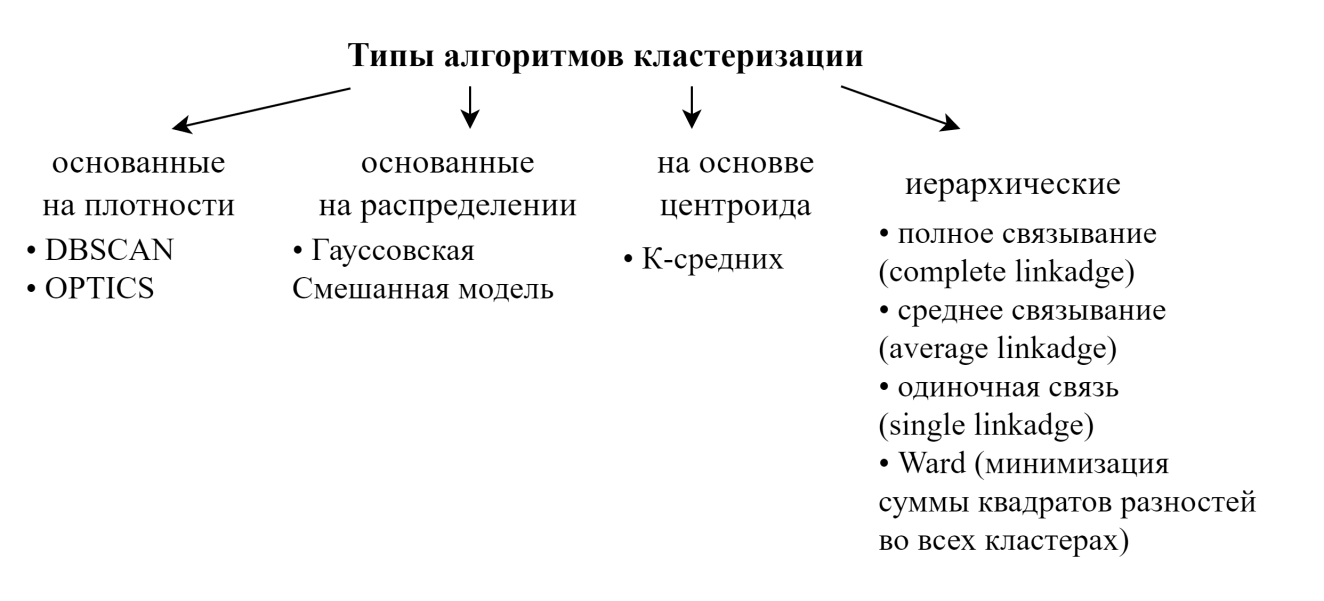
Таким образом, для прогнозирования материнского поведения необходимо выполнить три задачи Data Mining: кластеризация, классификация и поиск ассоциативных правил.

## Обзор и выбор методов для прогнозирования материнского поведения

Для выполнения каждой из трёх определённых ранее задач необходимы свои методы.

### Сравнение методов кластеризации и выбор наиболее подходящего для анализируемого датасета

Для решения задачи кластеризации существует большой ряд методов. Эти методы делятся на 4 типа, которые представлены на рисунке 2.2.



1. Типы алгоритмов кластеризации

Также для кластеризации объектов можно использовать нейронные сети, например, самоорганизующуюся карту Кохонена.

Кластеризация на основе плотности формирует кластеры на основе областей, имеющих высокую концентрацию точек данных и окружённых низкой плотностью точек. Такой принцип кластеризации игнорирует выбросы, что приводит к появлению шумовых точек в данных. Этот тип кластеризации не подходит для анализируемого набора данных, так как группы респондентов могут быть сильно приближены друг к другу или, наоборот, удалены. Оба случая не позволят сформировать кластеры с помощью данного метода. Также количество шумовых точек может быть велико, и прогнозирование на основании лишь части респондентов приведёт к потере знаний.

Принцип подхода кластеризации на основе распределения состоит в том, что точка принадлежит кластеру на основании вероятности. Суть в том, что есть центральная точка и по мере увеличения расстояния от неё точки данных вероятность принадлежности этой точки к кластеру уменьшается. Для использования методов данного типа нужно знать распределение данных, которое найти для анализируемого набора данных опросов матерей достаточно трудоёмко.

Кластеризация на основе центроидов достаточно быстра и эффективна. В алгоритмах данного типа производится разделение по группам на основании центров тяжести этих групп (центроидов). Каждому кластеру присваивается точка данных на основе квадрата расстояния от центра тяжести. Такой подход используют чаще всего. В результате его работы шумовых точек, то есть точек, не определённых к кластеру, не образуется.

Идея иерархической кластеризации состоит в том, что объекты соединяются друг с другом с помощью специальных инструкций. В процессе последовательного слияния образуется уникальный кластер, который собирает все образцы. Формирование кластеров визуализируют с помощью [дендрограммы](https://alphapedia.ru/w/Dendrogram), по которой определяют выбор количества кластеров.

Для набора данных опросов матерей не подходят методы, основанные на распределении и плотности, так как точки данных могут быть сильно удалены друг от друга, и есть вероятность потери данных при поиске ассоциативных правил.

Дальнейшее сравнение будет производится между следующими методами:

- К-средних;

- Агломеративная кластеризация;

- Самоорганизующаяся карта Кохонена.

Чтобы выбрать наиболее подходящий для рассматриваемого набора данных метод, целесообразно оценить качество кластеризации каждого из них. Для этого из анализируемого набора данных было взято 5 тысяч случайных записей одного из штатов, что составляет около 20 % от общего количества респондентов данного штата, и произведена кластеризация данных и осуществлена оценка силуэта при различном числе кластеров. О том, как производилась обработка ответов для кластеризации представлено в первом пункте третьей главы.

В основе идеи коэффициента силуэта лежит вычисление коэффициентов, которые присваиваются каждому объекту в кластере. Значения коэффициента изменяются от -1 до 1. Значение, равное 1, говорит о том, что объект похож на другие объекты кластера и не похож на объекты из других кластеров. Если в кластере много объектов с отрицательным значением коэффициента силуэта, то кластерная структура имеет плохое соответствие с естественными группами, это означает, что кластеров либо слишком много, либо слишком мало [18].

Для расчёта коэффициента силуэта сначала производится вычисление — меры «непохожести» объекта на остальные объекты этого кластера по формуле (2.1).

(2.1)

где — мощность I-го кластера (то есть число попавших в него объектов), — расстояние между объектами i и j кластера I. Усреднение производят по -1 объектам, так как нет смысла учитывать .

Следом определяется — среднее расстояние от объекта i кластера CI до объектов некоторого другого кластера CJ (CJ≠CI) по формуле 2.2.

(2.2)

где — объект, содержащийся в любом кластере , отличном от .

Зная коэффициенты и , коэффициента силуэта рассчитывается по формуле (2.3).

(2.3)

Расчёт среднего значения позволяет оценить, насколько кластерная структуры соответствует группировке данных.

Для быстрого получения результата в библиотеке scikit-learn языка программирования python есть готовый метод sklearn.metrics.silhouette\_score, который реализует подсчёт коэффициента силуэта.

Алгоритм К-средних делит набор N образцов X на K непересекающихся кластеров C таким образом, чтобы минимизировать сумму квадратов, т.е. евклидово расстояние, внутри кластера. Критерий суммы квадратов считается по формуле (2.4).

(2.4)

где — среднее значение точек в C кластере, — элемент набора X, n — количество наблюдений.

Также как и для коэффициента силуэта в библиотеке scikit-learn реализован алгоритм К-средних в виде класса sklearn.cluster.KMeans, которая принимает параметры:

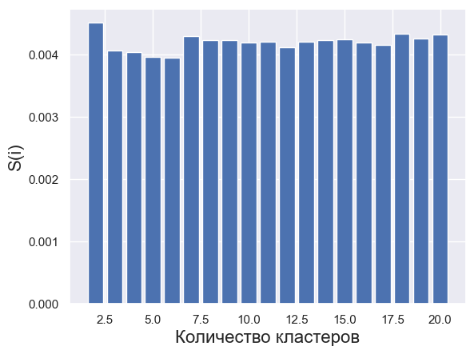
- n\_clusters — количество кластеров;

- n\_init — количество раз, при которых алгоритм k-средних запускается с разными начальными значениями центроида;

- max\_iter — максимальное количество итераций алгоритма k-средних для одного запуска;

- random\_state — определяет генерацию случайного числа для инициализации центроида. Код для формирования столбчатой диаграммы оценок силуэта при разном количестве кластеров для разных методов представлен в ПРИЛОЖЕНИИ Б.

Оценки силуэта для количества кластеров от 2 до 20 и метода К-средних представлены на рисунке 2.3.



1. Оценки силуэта для метода K-средних при различном количестве кластеров

Из рисунка 2.3 видно, что набор случайных записей можно разделить на 2 кластера, т.к. оценка силуэта для данного количества максимальна.

В агломеративной кластеризации выполняется иерархическая кластеризация объектов с использованием подхода снизу вверх: каждое наблюдение — один кластер, затем кластеры последовательно объединяются вместе.

Агломеративная кластеризация для оценки расстояния между точками данных может использовать различные подходы:

• евклидово расстояние — измеряет расстояние по прямой между двумя точками в пространстве;

• манхэттенское расстояние — измеряет сумму абсолютных различий между двумя точками/векторами во всех измерениях;

• косинусное сходство — измеряет угловое косинусное сходство между двумя векторами;

• L1 рассчитывает сумму абсолютных значений вектора;

• L2 рассчитывает квадратный корень из суммы квадратов векторных значений.

Для выбора наиболее подходящего подхода также была произведена оценка силуэта, однако количество кластеров было взято фиксированное и равное 5.

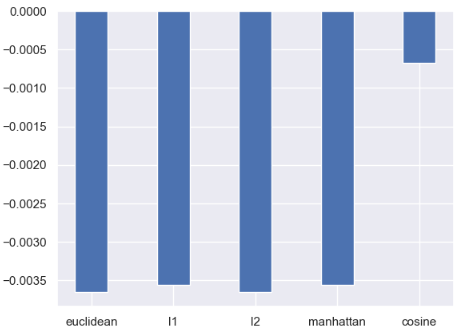
В библиотеке scikit-learn метод агломеративной кластеризации моно реализовать с помощью класса sklearn.cluster.AgglomerativeClustering, которая принимает параметры:

- n\_clusters — количество кластеров;

- affinity — метрика, используемая при вычислении расстояния между экземплярами в массиве объектов (cosine, euclidean и т.д.);

- linkage — критерий привязки (ward, average), определяет, какое расстояние использовать между наборами наблюдений

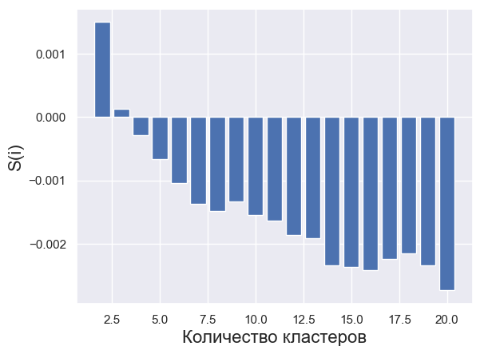
Оценки силуэта для 5 метрик представлены на рисунке 2.4.



1. Оценки силуэта для различных метрик агломеративной фильтрации

На рисунке 2.4 можно заметить плохое соответствие с естественными группами, однако наибольшая схожесть объектов внутри кластера и отличие от объектов других кластеров наблюдается при использовании косинусного сходства в качестве метрики для вычисления расстояния.

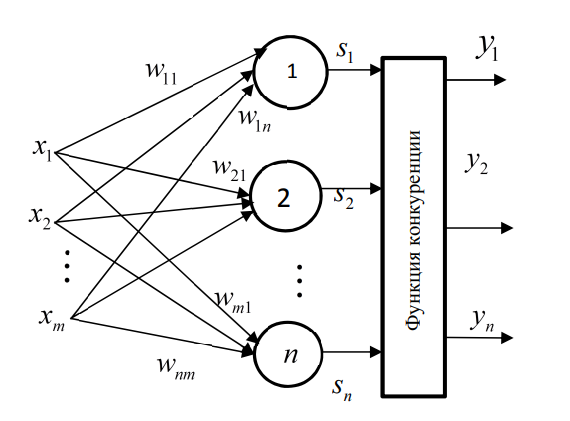
Коэффициенты силуэта при различном количестве кластеров при использовании агломеративной фильтрации на случайном наборе данных респондентов штата показаны на рисунке 2.5. Код для оценки качества кластеризации для данного метода представлен в ПРИЛОЖЕНИИ Б.



1. Оценки силуэта для метода агломеративной фильтрации при различном количестве кластеров

Данные на рисунке 2.5 показывают, что выборку можно разделить на 2 группы, однако метод агломеративной фильтрации показал более худшие оценки силуэта в сравнении с методом К-средних.

Карта Кохонена построена из нейронов вида WTM (Winner Takes Most — победитель получает больше). Структура сети Кохонена представлена на рисунке 2.6.



1. Структура сети Кохонена

Самоорганизующиеся карты используют неконтролируемое обучение. Этот тип обучения также называется конкурентным обучением.

Алгоритм обучения карт Кохонена для кластеризации опросов матерей:

1. Определение параметров всех весовых коэффициентов для соединений. Далее берётся случайная выборка качестве входных данных, после чего с помощью евклидова расстояния определяются веса нейрона-победителя.
2. Корректировка веса нейрона-победителя в процессе обучения, чтобы сделать его ближе к входным данным.
3. Корректировка весов соседей нейрона-победителя по такому же принципу, что и нейрона-победителя. Для этого необходимо произвести расчёт радиуса, значение которого уменьшается с каждой эпохой. Этот радиус определяет соседей, которые будут обновлены.
4. Фактическое значение радиуса вычисляется по формуле (2.5).

(2.5)

где t - текущая итерация, - радиус карты. Значение λ в формуле определяется с помощью формулы (2.6).

=k/ (2.6)

где k - количество итераций. Эта формула использует экспоненциальное затухание, уменьшая радиус по мере продолжения обучения.

1. Корректировка всех весов в области вычисляемого радиуса с помощью правила Кохонена (2.7).

(2.7)

где , — скорость обучения, — мера, позволяющая изменить вес нейрона тем больше, чем ближе этот нейрон к нейрону-победителю. Величины и рассчитываются по формулам (2.8)

(2.8)

где distBMU — расстояние до нейрона-победителя [19].

Алгоритм обучения самоорганизующихся карт Кохонена реализован в проекте sklearn-som для языка программирования python. В этом проекте класс sklearn\_som.som.SOM принимает параметры:

rn\_som.som.**SOM**sklearn\_som.som.**SOM**

- m — форма вдоль измерения 0 (по вертикали);

- n — форма вдоль измерения 1 (по горизонтали);

- dim — размерность (количество объектов) входного пространства;

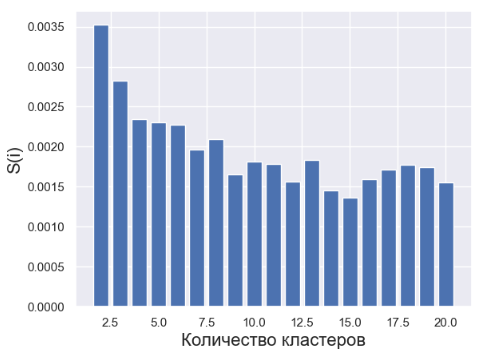
- lr — начальный размер шага для обновления весов карты Кохонена

- sigma — параметр для величины изменения каждого веса. Не обновляется во время обучения (как и скорость обучения);

- max\_iter — параметр для прекращения обучения при достижении этого количества взаимодействий

- random\_state — значение для генератора случайных чисел для инициализации веса.

Коэффициенты силуэта для различного количества кластеров представлены на рисунке 2.7. Код для формирования данных о коэффициентах силуэта представлен в ПРИЛОЖЕНИИ Б.



1. Оценки силуэта для самоорганизующихся карт Кохонена при различном количестве кластеров

Как и в двух предыдущих методах случайная выборка респондентов разделяется на 2 кластера. Результаты обучения карт Кохонена лучше, чем применение агломеративной кластеризации. Однако показатели метода К-средних превосходят оба этих метода, поэтому для кластеризации респондентов целесообразнее использовать К-средних.

### Анализ методов классификации

Также, как и для задачи кластеризации, для решения задачи классификации существует множество методов, например:

• дерево решений;

• случайный лес;

• искусственная нейронная сеть.

Деревья решений просты в понимании и их можно визуализировать, они также не требуют большой подготовки данных. Недостатком деревьев является то, что они могут быть велики и нестабильны, так как даже упрощённое изменение данных может нарушить всю структуру дерева.

Метод случайного леса имеет большую точность в отличие от деревьев решений, из-за сокращения чрезмерной подгонки. Однако, при большом объёме данных у техники может не хватать вычислительной мощности для выполнения данного метода. Также результат, получаемый с помощью дерева решений сильно сложнее визуализировать.

Нейронные сети обладают высокой устойчивостью к зашумлённым данным и способны классифицировать неподготовленные шаблоны, они лучше работают с входными и выходными данными с непрерывным значением. Недостаток искусственных нейронных сетей состоит в том, что они плохо интерпретируются по сравнению с другими моделями.

Так как неизвестно, хватит ли используемой для работы вычислительной мощности персонального компьютера, метод случайного леса было решено не использовать. В данной работе интерпретация модели важна, потому что это даёт возможность увидеть вопросы, которые являются решающими при определении респондентов в группу. Ввиду этого для решения задачи классификации были выбраны деревья решений.

Деревья решений могут быть реализованы различными алгоритмами. В рассматриваемых ранее исследованиях, где производился анализ результатов анкетирования о мотивации к участию в благотворительности применялся алгоритм CART. Также алгоритм CART устойчив к выбросам и [аномальным значениям](https://wiki.loginom.ru/articles/outlier.html) и имеет высокую скорость работы.

Дерево в алгоритме CART состоит из узла, в котором выполняется правило и двух потомков: правого (правило выполняется) и левого (правило не выполняется). При классификации респондентов сравниваются численные ответы матерей. И для нахождения оптимального правила используется оценочная функция, которая основана на интуитивной идее уменьшения неопределённости в узле.

Идея оценочной функции реализована в индексе Gini. Если набор данных Т содержит данные n классов, тогда индекс Gini определяется с помощью формулы (2.9).

(2.9)

где pi — относительная частота (вероятность) класса i в T.

Отличительной чертой данного алгоритма является механизм отсечения дерева, который состоит в построении дерева оптимального размера и получении точной оценки вероятности ошибочной классификации [20].

В библиотеке sklearn алгоритм CART реализован в классе sklearn.tree.DecisionTreeClassifier.

### Рассмотрение алгоритма для поиска ассоциативных правил

Для поиска ассоциативных правил наиболее часто используют алгоритм Apriori, который обладает свойством анти-монотонности, что позволяет обрабатывать большие объёмы данных за разумное время. Время обработки важно, поэтому для прогнозирования материнского поведения был выбран он.

Алгоритм Apriori для поиска причинно-следственных связей в поведении матерей состоит из следующих шагов:

1. Подсчёт ответов на вопросы (в формате «№вопросаОтвет»), которые наиболее часто встречаются. Для них производят расчёт поддержки, т.е. количество их появления в данных.
2. Генерация кандидатов Ck, т.е. наборов ответов, потенциально являющимися часто встречающимися, и также производят расчёт поддержки для кандидатов. Формирование кандидата Ck осуществляется путём расширения часто встречающегося набора ответов размера (k−1).
3. В соответствии со свойством анти-монотонности устраняются все наборы c Ck, если имеется хотя бы одно из его (k−1) подмножеств, которое не является часто встречающимся.
4. Подсчёт поддержки для каждого кандидата. Построение хэш-дерева с набором кандидатов. Узлы внутри дерева хранят хэш-таблицы с указателями на потомков, а листья – на кандидатов. Такой подход позволяет легко рассчитать поддержку.
5. Проверка преодоления значения поддержки кандидатов выше минимального порога. При преодолении кандидатов переносят в разряд часто встречающихся.
6. Генерация правил.

Чтобы извлечь правило из набора ответов F, находятся все его непустые подмножества. И для каждого подмножества ищутся правила s в соответствии с 2.10.

(2.10)

где conf() — достоверность правила, — правило, supp(F) — поддержка набора F, supp(s) — поддержка подмножества s, minconf — минимально допустимое значение достоверности.

Для извлечения правила использует рекурсивную процедуру. Правило, составленное из часто встречающегося набора ответов, должно включать все элементы набора. Например, если набор состоит из ответов {A,B,C}, то правило A⇒B не должно рассматриваться [21].

1. Главными характеристиками ассоциативных правил являются поддержка и достоверность. Поддержка – это число транзакций, которые содержат как условие, так и следствие, которое рассчитывается по правилу 2.11.

(2.11)

где — поддержка правила , — вероятность того, что оба ответа — А и В — будут в опросе респондента.

Достоверность — мера точности правила, рассчитывается по формуле 2.12.

(2.12)

где — достоверность, — вероятность ответа A при условии, что ответ B произошёл.

Также для оценки значимости ассоциативного правила используют метрику лифт, рассчитываемую по формуле 2.13

(2.13)

где — поддержка встречаемости ответа A или ответа B в наборе ответов респондента, — поддержка встречаемости ответа A, — поддержка встречаемости ответа B.

Высокое значение поддержки говорит об очевидности правила. Низкое — необоснованности, но такие правила зачастую новые и неочевидные.

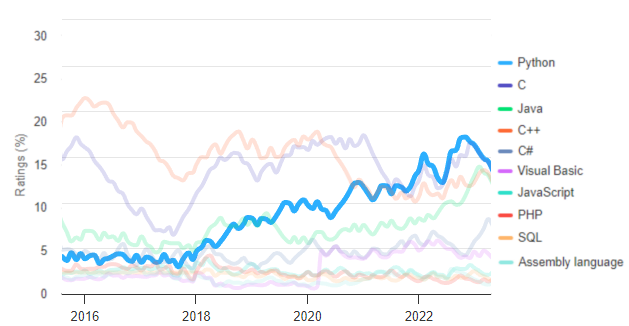
На основании значения достоверности правилу присваивается (высокая достоверность) или не присваивается (низкая достоверность) ценность.

1. Минимальный порог для достоверности и поддержки необходимо подобрать в соответствии с предметной областью. В психологии важен каждый человек, но рая для каждого не построить. Для муниципалитета и производителей важно большинство. Поэтому значения достоверности должны быть высокими. Высокий минимальный порог поддержки может привести к не учёту многих факторов, поэтому его значение должно быть достаточно малым.

## Инструменты для реализации методов

### Язык программирования Python

Согласно рейтингу TIOBE Index в мае 2023 года самым популярным языком программирования является Python [22]. Лидирующую позицию этот язык удерживает с конца 2021 года, что можно увидеть на графике изменения индекса популярности языков программирования от TIOBE Index с 2016 года, представленном на рисунке 2.8. На этом графике Python выделен ярким оттенком.



1. Изменение индекса TIOBE Index. 10 наиболее популярных в мире языков программирования с 2016 года

Популярность Python — это не только большое сообщество программистов, которое помогает в устранении ошибок и разработке, но и востребованность среди работодателей.

Для работы с большими данными выделяют несколько наиболее подходящих языков: [R,](https://blog.skillfactory.ru/9-yazykov-dlya-raboty-s-big-data/" \l "r--%D0%B4%D0%BB%D1%8F-%D0%BB%D1%8E%D0%B1%D0%B8%D1%82%D0%B5%D0%BB%D0%B5%D0%B9-%D1%81%D1%82%D0%B0%D1%82%D0%B8%D1%81%D1%82%D0%B8%D0%BA%D0%B8) [Python,](https://blog.skillfactory.ru/9-yazykov-dlya-raboty-s-big-data/" \l "python--%D0%BF%D0%BE%D0%BF%D1%83%D0%BB%D1%8F%D1%80%D0%BD%D1%8B%D0%B9-%D0%B8-%D0%BF%D0%BE%D0%BD%D1%8F%D1%82%D0%BD%D1%8B%D0%B9) [Java,](https://blog.skillfactory.ru/9-yazykov-dlya-raboty-s-big-data/" \l "java--%D1%81%D0%B0%D0%BC%D1%8B%D0%B9-%D1%83%D0%BD%D0%B8%D0%B2%D0%B5%D1%80%D1%81%D0%B0%D0%BB%D1%8C%D0%BD%D1%8B%D0%B9) [Scala,](https://blog.skillfactory.ru/9-yazykov-dlya-raboty-s-big-data/" \l "scala--%D1%81%D0%B0%D0%BC%D1%8B%D0%B9-%D0%BD%D0%B5%D0%B4%D0%BE%D0%BE%D1%86%D0%B5%D0%BD%D0%B5%D0%BD%D0%BD%D1%8B%D0%B9) [C++,](https://blog.skillfactory.ru/9-yazykov-dlya-raboty-s-big-data/" \l "c--%D1%81%D0%BB%D0%BE%D0%B6%D0%BD%D1%8B%D0%B9-%D0%BD%D0%BE-%D0%B1%D1%8B%D1%81%D1%82%D1%80%D1%8B%D0%B9) [MATLAB](https://blog.skillfactory.ru/9-yazykov-dlya-raboty-s-big-data/" \l "matlab--%D0%B4%D0%BB%D1%8F-%D0%BB%D1%8E%D0%B1%D0%B8%D1%82%D0%B5%D0%BB%D0%B5%D0%B9-%D0%BD%D0%B0%D1%83%D1%87%D0%BD%D1%8B%D1%85-%D0%BC%D0%B5%D1%82%D0%BE%D0%B4%D0%BE%D0%B2). Сравнение количества вакансий на крупнейшей онлайн-рекрутинговой платформе в России для этих языков представлено в таблице 2.1.

1. Число вакансий для разработчиков всех уровней для используемых в Big Data языков программирования

|  |  |
| --- | --- |
| Язык программирования | Число вакансий |
| Python | 9 926 |
| Go | 6 187 |
| R | 5 961 |
| Java | 5 751 |
| C++ | 3 904 |
| MATLAB | 612 |
| Scala | 473 |

Данные таблицы 2.1 доказывают, что Python — это популярный и востребованный инструмент для разработки.

Python выбирают из-за следующих преимуществ:

• больше количество учебных материалов ввиду популярности;

• возможность сочетать с другими языками программирования, такими как Java, C или C++;

• высокая скорость разработки благодаря большому количеству встроенных функций, таких как динамическая типизация;

• большое количество библиотек для решения различных задач.

На сегодняшний день используется четвёртая версия языка Python — Python 3. Её используют 95 % Python-разработчиков.

Количество Python библиотек превышает 140 000. Они охватывают огромный круг задач и сокращают затрачиваемое на разработку время. Все, рассмотренные в предыдущем пункте методы и алгоритмы имеют реализации либо в виде отдельной библиотеки, либо в составе одной из библиотек.

### Набор Python библиотек для работы с большими данными и реализации алгоритмов машинного обучения

Библиотека NumPy — проект с открытым исходным кодом. NumPy привносит вычислительную мощь таких языков, как C и Fortran, в Python. Виду этого это довольно быстрый инструмент для работы с данными. NumPy поддерживает векторные операции с многомерными массивами, что позволяет не использовать циклы для обхода массива и индексацию в явном виде. Эти операции проводятся “за кулисами” в оптимизированном, предварительно скомпилированном коде на C [23].

Использование векторных операций имеет ряд преимуществ:

• код более краток и понятен и его легче читать;

• уменьшение количества строк кода позволяет совершить меньше ошибок;

• код больше похож на стандартную математическую нотацию, что упрощает написание программы для математических конструкций.

Более сложные вычисления позволяет выполнять библиотека SciPy, которая основана на NumPy. SciPy расширяет возможности интерактивного сеанса Python и предоставляет пользователю высокоуровневые команды и классы для манипулирования данными и их визуализации. С этой библиотекой интерактивный сеанс Python становится средой обработки данных и прототипирования систем, конкурирующей с такими системами, как MATLAB, IDL, Octave, R-Lab и SciLab [24].

Для создания визуализаций на Python используется библиотека matplotlib. Эта библиотека позволяет создать любой графический элемент, который может быть статическим, интерактивным или анимированным. Функционал matplotlib включает в себя следующее:

• создание интерактивных фигурок с возможностью их масштабирования, перемещения и обновления;

• настройка стиля и макета для визуализации;

• возможность экспортировать графику во многих файловых форматах с помощью функции savefig(), поддерживаемые форматы файлов: eps, jpeg, jpg, pdf, pgf, png, ps, raw, rgba, svg, svgz, tif, tiff [25];

• возможность встраивать элементы matplotlib в Jupiter Lab и графические пользовательские интерфейсы;

• большой набор сторонних пакетов, которые расширяют функциональность matplotlib или основываются на ней.

Также для визуализации графиков используют библиотеку graphviz, которая позволит сделать визуализацию деревьев решений при классификации матерей-респондентов более понятной благодаря различным цветам и шрифтам.

Библиотека folium позволяет просто и быстро визуализировать данные на интерактивной карте.

На NumPy, SciPy, и matplotlib основана библиотека scikit-learn, предназначенная для машинного обучения и ориентированная на моделирование данных. Она фокусируется на моделировании, а не на загрузке данных и манипулировании ими [26].

Функции библиотеки scikit-learn:

• обучение с учителем;

• обучение без учителя;

• кластеризация;

• снижение размерности;

• ансамблевые методы;

• кросс-валидация;

• выделение признаков;

• отбор признаков.

В пункте 2.2 были приведены примеры реализаций используемых для обработки данных анкетирования методов в библиотеке scikit-learn. Также для кластеризации с помощью самоорганизующихся карт Кохонена используется библиотека sklearn\_som, а поиска ассоциативных правил — apyori.

Обработку данных часто осуществляют с помощью библиотеки pandas, которая построена на пакете NumPy.

Pandas предлагает разработчику много инструментов:

• быстрый и эффективный объект DataFrame для обработки данных с интегрированной индексацией;

• инструменты для чтения и записи данных между структурами данных в памяти и различными форматами, в том числе CSV и текстовыми файлами;

• гибкое изменение формы и поворота наборов данных;

• агрегирование или преобразование данных с помощью мощного механизма группирования, позволяющего выполнять операции разделения, применения и объединения наборов данных;

• высокопроизводительное объединение наборов данных;

• иерархическая индексация по осям обеспечивает интуитивно понятный способ работы с данными большой размерности в структуре данных меньшей размерности [27].

### Среда разработки Jupyter

Jupyter Notebook - одна из самых популярных IDE для Python. По данным ежегодного опроса разработчиков Python, проводимый в рамках совместных усилий Python Software Foundation и JetBrains в 2021 году, 25 % разработчиков используют Jupyter в качестве основной IDE, что делает её второй популярной IDE для Python, уступающей первое место VS Code, с его 26 % [28].

Jupyter выбирают благодаря следующим преимуществам:

• возможность выполнения кода программы блоками, что позволяет быстрее отслеживать ошибки и сравнивать результаты разных вариантов обработки;

• объединение кода, текста и изображений, интерактивных элементов, графического интерфейса и многих других компонентов в одном документе;

• файлы Jupyter Notebook представлены в формате JSON. Созданными текстовыми файлами можно легко делиться и конвертировать – в Jupyter есть специальный инструмент под названием nbconvert, который позволяет пользователям конвертировать notebook во множество форматов, таких как HTML, PDF и другие. Это также позволяет пользователям создавать документацию вместе с кодом.

• возможность одновременно видеть как код, так и результаты своей работы.

## Математическая постановка задачи

Для прогнозирования материнского поведения необходимо произвести кластеризацию образцов анкетирования, представленных в виде матрицы , где – значение ответа на -ый () вопрос -ого () респондента, на k кластеров. В результате будет получена матрица , где — вектор ответов -го респондента, — номер кластера. Требуется построить решающее дерево, наилучшим образом предсказывающее целевую функцию (группу, к которой относится респондент) в зависимости от (ответов участника анкетирования). Для каждого кластера нужно получить набор ассоциативных правил A→B, имеющих значение поддержки — 20 % и достоверности от 50 % до 90 %.

**Вывод**

В данном разделе были рассмотрены алгоритмы для прогнозирования материнского поведения на основе данных анкетирования и среди них выбраны наиболее подходящие, которые будут использоваться для дальнейшей обработки данных опросов матерей:

• К-средних;

• алгоритм CART;

• алгоритм Apriori.

Также были рассмотрены инструменты для реализации выбранных алгоритмов.

Произведена математическая постановка задачи.

# ПРОЕКТНАЯ ЧАСТЬ

## Анализ набора данных и его подготовка для работы

Для начала следует определить, данными каких штатов располагает датасет, так как данные о том, в каких штатах была внедрена система PRAMS в 2011 году (в датасете данные только этого года), отсутствуют. Также есть вероятность, что Центром по контролю и профилактике заболеваний США опубликовал данные не по всем штатам.

Данные опросов были опубликованы в виде одного csv файла, столбцы которого были описаны в пункте 1.4. Для дальнейшей работы требуется «прочитать» записи в файле, что можно сделать, воспользовавшись функцией библиотеки pandas read\_csv().

Чтобы определить, данными каких штатов располагает датасет, можно воспользоваться функцией unique().

В наборе данных представлены данные опрос респондентов из 26 штатов: Арканзас, Колорадо, Делавэр, Джорджия, Гавайи, Массачусет, Мэриленд, Мичиган, Миннесота, Миссури, Небраска, Нью-Джерси, Нью-Йорк (исключая Нью-Йорк Сити), Висконсин, Вайоминг, Нью-Мексико, Оклахома, Нью-Йорк Сити, Вермонт, Орегон, Пенсильвания, Род-Айленд, Вашингтон, Западная Вирджиния, Мэн, Юта.

Данные о количестве респондентов в каждом штате были получены с помощью функции count() и визуализированы на карте с помощью библиотеки folium. Код расчёта количества респондентов и «отрисовки» карты представлен в ПРИЛОЖЕНИИ В, а сама карта — в ПРИЛОЖЕНИИ Г. Штат, данных которого нет не имеет цвета. Данных из Аляски нет в анализируемом наборе данных, и для более удобного восприятия карты, этот штат был обрезан. Данные о количестве респондентов в каждом штате представлены в таблице 3.1.

1. Число респондентов в каждом штате

| **Штат** | **Число респондентов** | **Штат** | **Число респондентов** |
| --- | --- | --- | --- |
| Оклахома | 32 280 | Нью-Джерси | 22 899 |
| Колорадо | 30 448 | Нью-Йорк Сити | 22 871 |
| Орегон | 28 837 | Миннесота | 22 544 |
| Небраска | 28 339 | Арканзас | 21 665 |
| Мичиган | 27 349 | Род-Айленд | 20 486 |
| Западная Вирджиния | 26 264 | Миссури | 20 450 |
| Массачусет | 26 201 | Вашингтон | 19 511 |
| Висконсин | 25 733 | Вермонт | 17 837 |
| Гавайи | 25 700 | Пенсильвания | 17 651 |
| Джорджия | 24 624 | Делавэр | 17 510 |
| Мэриленд | 24 095 | Нью-Йорк (исключая Нью-Йорк Сити) | 17 274 |
| Юта | 23 590 | Мэн | 17 100 |
| Нью-Мексико | 23 350 | Вайоминг | 11 373 |

На большинство вопросов анкеты ответом является «да» или «нет», однако вариантов написания этих ответов несколько. Чтобы найти все эти варианты необходимо произвести поиск возможных вариантов ответов на все вопросы. Данная процедура проводилась с помощью функции iloc[], которая позволяет выделить столбец в датафрейме, и ранее использовавшейся функции unique(). 3 — 199 — номера столбцов с вопросами.

В результате были получены все вариации ответов «да» (YES, Yes, YES (CHECKED)) и «нет» (NO, No, NO (UNCHECKED)). Также было выявлено 26 вопросов, где ответом является категория, и 2 вопроса, в которых помимо ответов «да» и «нет» есть ответ «Я не знаю» и «ещё в больнице».

Для кластеризации респондентов необходимо привести ответы на вопросы к числовым значениям. В соответствии с имеющимися данными был составлен словарь, в котором вариации ответов «да» и «нет» приведены к единому значению, а ответы «Я не знаю» и «ещё в больнице» — к промежуточным значениям между «да» и «нет».

Чтобы каждый вопрос отделялся от другого, значения ответов были умножены на вопроса. Если ответом на вопрос является категория, то порядковый номер категории умножался на 10, а затем на номер вопроса. Также в процессе замены ответов на числовые значения пополнялся список словарей (resA) с ключами вида «№вопросаОтвет» и числовыми значениями этих ответов. Этот словарь понадобится потом для поиска понятных ассоциативных правил и интерпретации результатов дерева решений. Предполагается, что данная процедура привнесёт различия между респондентами, чтобы получить кластеры похожих между собой объектов внутри кластера и отличающихся от объектов других кластеров и более высокий коэффициент силуэта.

## Выявление причинно-следственных связей материнского поведения в штате США

США — государство, занимающее четвёртое место по площади третье по населению. Этнический и расовый состав в этой стране довольно разнообразен. Законы одного штата Америки могут отличаться от другого, также как и уровень дохода, менталитет, инфраструктура и многое другое. Поэтому рассмотрение всех респондентов в одной группе может привести к большим погрешностям в результатах. В данной работе будет рассмотрено несколько штатов из всех представленных в наборе данных. Не лишним будет сравнить результаты, чтобы оценить масштабы влияния факторов.

Штат Колорадо согласно таблице 3.1 занимает второе место по количеству респондентов среди рассматриваемых штатов. Этот штат не является самым крупным или богатым. Это просто американская провинция.

## Разделение респондентов на группы

Для начала респондентов необходимо разделить на группы с помощью метода К-средних. Чтобы выбрать оптимальное количество групп можно, как и при выборе метода, воспользоваться коэффициентом силуэта. Отбор данных включал выбор всех записей из штата Колорадо и сохранение только столбцов с вопросами. Для стандартизации диапазона функциональных возможностей входного набора данных был использован класс библиотеки sklearn StandardScaler(). Этот класс выполняет стандартизацию поэлементно и использует для этого формулу 3.1.

(3.1)

где z — стандартное значение образца x, x — вектор ответов респондента, u — вектор средних значений, s — стандартное отклонение.

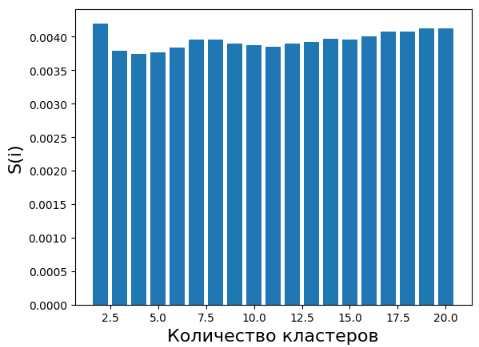
Далее были произведены разбиение набора данных штата на группы от 2 до 20 с помощью метода К-средних, использующего евклидово расстояние из формулы 3.2.

(3.2)

где — евклидово расстояние от центроида до точки ответов респондента, — координаты центроида, — координаты ответов респондента.

Были вычислены оценки силуэта с помощью функции silhouette\_score(). Расчёт коэффициента в данной функции производится по формуле 2.3.

Результаты оценок силуэта были выведены на столбчатую диаграмму, которая представлена на рисунке 3.1.



1. Оценки силуэта для штата Колорадо при различном количестве кластеров

На рисунке 3.1 видно, что при увеличении количества кластеров, коэффициент силуэта растёт, однако при разделении респондентов на 2 группы показатель является наивысшим, поэтому разделение будет производится на 2 кластера.

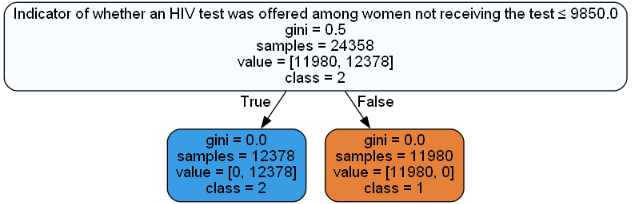
В результате разделения получилось 2 группы в первой — 14927 респондентов, во второй — 15521.

## Построение дерева классификации респондентов

Следующим шагом будет построение дерева классификации. Для этого необходимо добавить данные о принадлежности респондента к группе датафрейм. Этот столбец и будет целевой переменной для классификации.

Для проверки качества классификации выборка по штату была разделена на обучающую, которая составляет 80% от общей и тестовую. Разделение осуществлялось с помощью функции sample() библиотеки pandas, которая отбирает случайным образом записи. Затем обучающая выборка была удалена из общей и получена тестовая.

Для построения дерева использовалась функция DecisionTreeClassifier() библиотеки sklearn, которая реализует алгоритм CART, а для «отрисовки» — export\_graphviz() библиотеки graphviz. Для дальнейшего сравнения результата с другими штатами, визуализация дерева классификации была экспортирована в файл png с помощью функции render() библиотеки graphviz. В результате было получено дерево классификации на рисунке 3.2.



1. Дерево классификации для штата Колорадо

Дерево классификации для штата Колорадо состоит всего из двух слоёв. На пером слое индекс gini равен 0.5, это означает, что элементы равномерно распределены по некоторым классам, на втором слое индекс gini для обеих групп равен 0, что означает принадлежность всех элементов одному классу.

Получается, в штате Колорадо матерей можно разделить на две группы по ответу на один вопрос: Показатель того, предлагалось ли пройти тест на ВИЧ женщине, не прошедшей тест.

Чтобы интерпретировать значение 9850, понадобится созданный ранее лист словарей resA. Вопрос для классификации — это 195-ый вопрос опроса. Его ключи и значения: «NO195»: 0, «YES195»: 19700. Таким образом, если респонденту не предлагали пройти тест на ВИЧ при условии, что она его не прошла, то женщина относится ко второй группе, если предлагали — к первой.

Оценить качество классификации можно с помощью функции библиотеки sklearn accuracy\_score(), которая вычисляет точность, т.е. коэффициент правильно определённых объектов. Рассчитывается значение этой функции как отношение правильно определённых образцов к общему набору образцов, формула 3.3.

(3.3)

где — прогнозируемое значение i-ого образца, — истинное, 1(x)- индикаторная функция, — количество образцов.

Чтобы получить прогнозные значения, использовалась функция predict. Расчёт точности классификации производился на тестовой выборке.

На тестовых данных был получен результат функции accuracy\_score равный 1, т.е. правильно определяется класс всех образцов.

## Поиск ассоциативных правил в каждой отдельной группе

Для поиска ассоциативных правил числовые значения были переведены в текстовые с помощью списка словарей resA для удобства интерпретации. Ассоциативные правила для каждого кластера рассматриваются отдельно. Чтобы получить правила, была использована функция apriori() библиотеки apyori. Минимальная поддержка и достоверность для данной функции были установлены на уровне 10 %, так как неизвестно, удастся ли получить правила с необходимым уровнем поддержки в 20 % и достоверности от 50 % до 90 %.

Данные о правилах, а также о их поддержке, достоверности и лифте были преобразованы в датафрейм и отсортированы по коэффициенту поддержки. Листинг кода для осуществления кластеризации, построения дерева решений, поиска ассоциативных правил и визуализации получаемых данных для штат Колорадо представлен в ПРИЛОЖЕНИИ Д. Прогнозирование материнского поведения в остальных штатах осуществляется аналогичным образом.

Так как правил получается несколько сотен, было решено рассматривать только первые 15. В результате для первой группы респондентов штата Колорадо был получен список правил, представленный в таблице 3.2.

1. Ассоциативные правила для группы 1 штата Колорадо

| **Правило** | **Поддержка** | **Достоверность** | **Лифт** |
| --- | --- | --- | --- |
| YES196 -> YES195 | 0.883969 | 0.883969 | 1.0 |
| NO (UNCHECKED)32 -> YES195 | 0.522141 | 0.522141 | 1.0 |
| 0145 -> YES195 | 0.518389 | 0.518389 | 1.0 |
| NO63 -> YES195 | 0.516581 | 0.516581 | 1.0 |
| YES87 -> YES195 | 0.515442 | 0.515442 | 1.0 |
| NO7 -> YES195 | 0.513901 | 0.513901 | 1.0 |
| NO (UNCHECKED)81 -> YES195 | 0.5137 | 0.5137 | 1.0 |
| YES94 -> YES195 | 0.513298 | 0.513298 | 1.0 |
| YES195 -> NO187 | 0.512896 | 0.512896 | 1.0 |
| YES75 -> YES195 | 0.511958 | 0.511958 | 1.0 |
| YES130 -> YES195 | 0.511824 | 0.511824 | 1.0 |
| YES73 -> YES195 | 0.511757 | 0.511757 | 1.0 |
| YES (CHECKED)92 -> YES195 | 0.511757 | 0.511757 | 1.0 |
| NO128 -> YES195 | 0.511355 | 0.511355 | 1.0 |
| NO (UNCHECKED)110 -> YES195 | 0.511288 | 0.511288 | 1.0 |

Наиболее достоверным и с высоким уровнем поддержки является правило: если женщина, которой предложили тестирование на ВИЧ, отказалась от него, то ей предлагалось, как женщине не прошедшей этот тест, пройти его. Это правило тривиально. Оно не несёт полезной информации. Далее по списку идёт правило: если Medicaid не покрыла предоставление медицинской услуги, то женщине, не прошедшей тест на ВИЧ, предлагалось его пройти. Следующее правило: если женщина не имела предыдущих живорождений, то ей, как не прошедшей тест на ВИЧ предлагалось его пройти. Также условием предложения женщине, не прошедшей тестирование на ВИЧ, пройти его, являются:

• за 12 месяцев до рождения ребёнка она ссорилась с мужем или партнёром чаще, чем обычно;

• использование послеродовой контрацепции;

• за 12 месяцев до рождения ребёнка она переехала на новый адрес;

• деньги от комиссионных, дивидендов или дохода от сдачи в аренду не являются источником дохода домохозяйства;

• неиспользование противозачаточных средств при зачатии при условии, что не пыталась забеременеть;

• младенец, проживает в настоящее время с матерью;

• во время беременности были инфекции почек или мочевого пузыря (мочевыводящих путей);

• зарегистрированы стрессы, связанные с травмой;

• причина неиспользования противозачаточных средств при зачатии - думала, что она или партнёр бесплодны;

• во время беременности не была травмирована в автокатастрофе.

Также если женщине, не прошедшей тест на ВИЧ, предложили его пройти, то за 3 месяца до того, как она забеременела новым ребёнком, у неё не было эпилепсии (припадка).

Поддержка и достоверность для всех правил одинаковы между собой. Их значения довольно высоки, однако лифт для всех правил равен 1, что говорит о том, что правила проявляются в основном как случайное совместное появление. Также стоит отметить, что в 14 правилах из 15 поддержка составляет чуть больше 50 %, т.е. половину группы, и однозначно говорить о факторах, влияющих на предложение о тестировании на ВИЧ, не стоит. Условия необходимо рассматривать в комплексе.

Сравним полученные данные из таблицы 3.2 с данными 2-ой группы, правила для которой находятся в таблице 3.3.

1. Ассоциативные правила для группы 2 штата Колорадо

| **Правило** | **Поддержка** | **Достоверность** | **Лифт** |
| --- | --- | --- | --- |
| YES196 -> NO195 | 0.877585 | 0.877585 | 1.0 |
| NO160 -> NO195 | 0.519039 | 0.519039 | 1.0 |
| YES167 -> NO195 | 0.516075 | 0.516075 | 1.0 |
| NO185 -> NO195 | 0.514786 | 0.514786 | 1.0 |
| YES69 -> NO195 | 0.513756 | 0.513756 | 1.0 |
| NO75 -> NO195 | 0.513111 | 0.513111 | 1.0 |
| NO (UNCHECKED)111 -> NO195 | 0.511694 | 0.511694 | 1.0 |
| NO95 -> NO195 | 0.511114 | 0.511114 | 1.0 |
| Public21 -> NO195 | 0.51105 | 0.51105 | 1.0 |
| NO43 -> NO195 | 0.510534 | 0.510534 | 1.0 |
| YES49 -> NO195 | 0.511178 | 0.511178 | 1.0 |
| YES1 -> NO195 | 0.511178 | 0.511178 | 1.0 |

Продолжение таблицы 3.3

|  |  |  |  |
| --- | --- | --- | --- |
| **Правило** | **Поддержка** | **Достоверность** | **Лифт** |
| NO (UNCHECKED)91 -> NO195 | 0.511178 | 0.511178 | 1.0 |
| NO157 -> NO195 | 0.510921 | 0.510921 | 1.0 |
| NO134 -> NO195 | 0.510599 | 0.510599 | 1.0 |

Ассоциативные правила для второй группы штата Колорадо не повторяют правил для первой группы. Однако, что для первой, то для второй группы правила связаны с тестированием на ВИЧ женщин, которые его не проходили. Связано это с тем, что этот вопрос является решающим при классификации респондентов. Явным различием двух групп является ответ на вопрос. Первой группе предлагали тестирование на ВИЧ, а второй — нет.

Во второй группе условием того, что женщине, не прошедшей тестирование не предлагали его пройти, являются:

• женщина отказывалась от тестирования;

• последние 2 года женщина не курила;

• за 3 месяца до того, как она забеременела новым ребёнком, у неё была депрессия;

• за 3 месяца до того, как она забеременела, у неё не было проблем с сердцем;

• за 12 месяцев до рождения ребёнка она была в физической борьбе;

• младенец в настоящее время не проживает с матерью;

• медицинская услуга оплачена страховкой не через работодателя;

• во время беременности не использовались противозачаточные средства;

• источник дородовой помощи — государственный;

• дородовой уход был получен в кратчайшие сроки;

• во время посещения дородового ухода медицинский работник говорил о том, чтобы сдать анализ крови на ВИЧ;

• медицинский работник говорил о том, что делать, если женщина чувствует депрессию во время беременности или после рождения ребёнка;

• мать не курила в течение трёх месяцев до беременности;

• проблем с плацентой у матери не было.

Поддержка и достоверность для выявленных во второй группе правил также довольно высока. Большинство правил отличается от первой группы.

Более подробно следует рассмотреть вторую группу респондентов, чтобы узнать выявить другие причинно-следственные связи в поведении матерей. Так как наиболее достоверные правила связаны с вопросом о тестировании на ВИЧ, следует сделать дополнительный отбор правил по интересующему вопросу.

Одним из условий того, что женщине не предложили тестирование — за 3 месяца до того, как она забеременела новым ребёнком, у неё была депрессия. Поэтому следует произвести фильтрацию правил и найти те, в которых следствием является это условие. Это ответ «YES» на вопрос 167.

Отобранные ассоциативные правила также отсортированы по убыванию достоверности. Первые 15 правил представлены в таблице 3.4.

1. Ассоциативные правила для группы 2 штата Колорадо, в которых следствием является наличие у женщины депрессии за 3 месяца до того, как она забеременела

| **Правило** | **Поддержка** | **Достоверность** | **Лифт** |
| --- | --- | --- | --- |
| YES196 -> YES167 | 0.453321 | 0.453321 | 1.0 |
| YES1 -> YES167 | 0.271761 | 0.271761 | 1.0 |
| NO160 -> YES167 | 0.271117 | 0.271117 | 1.0 |
| NO157 -> YES167 | 0.267831 | 0.267831 | 1.0 |
| NO138 -> YES167 | 0.267315 | 0.267315 | 1.0 |
| YES13 -> YES167 | 0.266542 | 0.266542 | 1.0 |
| YES (CHECKED)90 -> YES167 | 0.265962 | 0.265962 | 1.0 |
| NO40 -> YES167 | 0.265833 | 0.265833 | 1.0 |
| VAGINALLY15 -> YES167 | 0.265447 | 0.265447 | 1.0 |
| YES178 -> YES167 | 0.265318 | 0.265318 | 1.0 |
| NO177 -> YES167 | 0.265125 | 0.265125 | 1.0 |
| NO112 -> YES167 | 0.264996 | 0.264996 | 1.0 |
| NOT OBESE164 -> YES167 | 0.264931 | 0.264931 | 1.0 |
| YES10 -> YES167 | 0.264803 | 0.264803 | 1.0 |
| YES181 -> YES167 | 0.264803 | 0.264803 | 1.0 |

Достоверность правил в таблицы 3.4 меньше 50 %, однако для каждого из этих правил как условие, так и следствие присутствует более чем в 25 % выборки, а это около 4 000 человек.

Условиями наличия у женщины депрессии за 3 месяца до беременности являются:

• отказ женщины от тестирования на ВИЧ;

• медицинский работник говорил с женщиной о том, что делать, если она чувствует депрессию во время беременности или после рождения ребёнка;

• женщина не курила сигареты за последние 2 года;

• во время беременности высокое кровяное давление не наблюдалось;

• медицинский работник говорил, как употребление алкоголя во время беременности может повлиять на ребёнка;

• женщина при зачатии не использовала контрацепцию, так как не возражала против беременности;

• медицинский работник не говорил с матерью о грудном вскармливании вашего ребёнка;

• вагинальные роды;

• проверка депрессии до беременности производилась;

• до беременности женщина не занималась спортом 3 или более дней в неделю;

• ребёнок не был выписан из больницы в течение 48 часов после рождения;

• женщина не страдала ожирением до беременности;

• за 12 месяцев до рождения ребёнка муж или партнёр матери потерял работу;

• медицинский работник говорил женщине о том, как подготовиться к здоровой беременности и рождению ребёнка.

Из условий видно, что депрессию диагностируют в основном у физически здоровых женщин при этом они проходят проверку для этой диагностики. Также стоит отметить, что матерей, у которых была депрессия за 3 месяца до того, как они забеременели, консультировали о том, что делать, если они начнут её чувствовать. Также стоит отметить, что у этих женщин беременность была запланированной.

Из полученных данных стоит выделить, что женщины здоровы и заботятся о своём здоровье. Для выявления причин депрессии стоит пересмотреть список вопросов для опросника, а также рассмотреть правила, где следствием является отсутствие депрессии. Для этого необходимо выявить ассоциативные правила с ответом «NO» на вопрос 167, которые представлены в таблице 3.5.

1. Ассоциативные правила для группы 2 штата Колорадо, в которых следствием является отсутствие у женщины депрессии за 3 месяца до того, как она забеременела

| **Правило** | **Поддержка** | **Достоверность** | **Лифт** |
| --- | --- | --- | --- |
| NO195 -> NO167 | 0.483925 | 0.483925 | 1.0 |
| YES196 -> NO167 | 0.424264 | 0.424264 | 1.0 |
| YES37 -> NO167 | 0.252561 | 0.252561 | 1.0 |
| NO22 -> NO167 | 0.252368 | 0.252368 | 1.0 |
| YES136 -> NO167 | 0.251659 | 0.251659 | 1.0 |
| YES70 -> NO167 | 0.25153 | 0.25153 | 1.0 |
| YES69 -> NO167 | 0.250757 | 0.250757 | 1.0 |
| NO (UNCHECKED)111 -> NO167 | 0.250564 | 0.250564 | 1.0 |
| YES179 -> NO167 | 0.250499 | 0.250499 | 1.0 |
| NO185 -> NO167 | 0.250306 | 0.250306 | 1.0 |
| WIC36 -> NO167 | 0.250242 | 0.250242 | 1.0 |
| NO158 -> NO167 | 0.250242 | 0.250242 | 1.0 |
| YES (CHECKED)92 -> NO167 | 0.249984 | 0.249984 | 1.0 |
| NO75 -> NO167 | 0.249855 | 0.249855 | 1.0 |
| NOT OBESE166 -> NO167 | 0.249855 | 0.249855 | 1.0 |

Из таблицы 3.5 выделяются следующие условия для отсутствия депрессии у женщины за 3 месяца до того, как она забеременела:

• не предлагалось пройти тест на ВИЧ при условии, что он не пройден;

• позднее обращение за дородовой помощью или её отсутствие;

• роды производились не в больнице;

• мать сообщала о гестационном диабете;

• за 12 месяцев до рождения ребёнка мать или её муж или партнёр попали в тюрьму;

• за 12 месяцев до рождения ребёнка мать была в физической борьбе;

• медицинская услуга оплачивалась страховкой не через работодателя;

• были обсуждения семейного анамнеза до беременности;

• проблем с сердцем за 3 месяца до беременности не было;

• во время беременности было участие в программе WIC (Специальной программа дополнительного питания для женщин, младенцев и детей в США);

• мать не курила в течение последних трех месяцев беременности;

• женщина не использовала противозачаточные средства при зачатии, так как думала, что она или партнёр бесплодны;

• в настоящее время младенец не проживает с матерью;

• женщина не страдает ожирением.

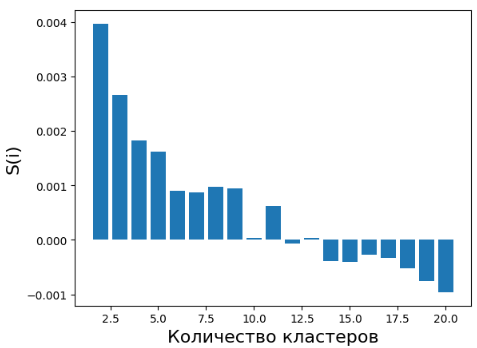
Из сравнения условий наличия и отсутствия депрессии у женщины за 3 месяца до беременности стоит выделить, что отказ от прохождения тестирования на ВИЧ не является причиной депрессии, так как это условие наблюдается у обеих групп правил. Также что женщины с депрессией, что без, не страдают ожирением. Также стоит выделить, что женщины, у которых не было депрессии есть социальная помощь и сложные жизненные ситуации.

## Применение прогнозирования материнского поведения для сравнения матерей двух штатов

В рассматриваемом наборе данных Нью-Йорк представлен двумя выборками: Нью-Йорк Сити и Нью-Йорк без Нью-Йорк Сити. Так территориально Нью-Йорк Сити находится в штате Нью-Йорк и является крупнейшим городом США, рассмотрение результатов для них будет полезно.

Также как и для штата Колорадо, для Нью-Йорк Сити и штата Нью-Йорк нужно определить оптимальное количество групп, на которое буду разделены респонденты. Для этого нужно вычислить оценки силуэта.

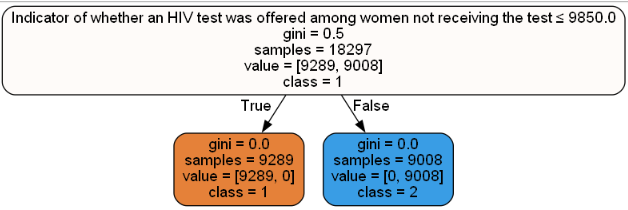
Для Нью-Йорк Сити коэффициенты силуэта для разного количества кластеров представлены на рисунке 3.3, на котором видно, что группу респондентов оптимально разделить на 2.



1. Оценки силуэта для различного количества кластеров Нью-Йорк Сити

В результате разделения респондентов в первый класс попали 11 571 человек, во второй — 11 300.

Дерево классификации (рисунок 3.4) для определения респондентов Нью-Йорк Сити в группу правильно предсказывает группу в 100 % случаях, то было проверено на тестовой выборке (коэффициент, вычисленный функцией accuracy\_score() равен 1).



1. Дерево классификации для Нью-Йорк Сити

Для Нью-Йорк Сити классификационным правилом является ответ на вопрос: «предлагалось ли пройти тест на ВИЧ женщине, не прошедшей тест», также как и для штата Колорадо. Если женщине не предлагалось пройти тест, то её относят к первой группе, если предлагалось — ко второй.

Первые 15 ассоциативных правил с наибольшей поддержкой и достоверностью для первой группы Нью-Йорк Сити представлены в таблице 3.6.

1. Ассоциативные правила для группы 1 Нью-Йорк Сити

| **Правило** | **Поддержка** | **Достоверность** | **Лифт** |
| --- | --- | --- | --- |
| YES196 -> NO195 | 0.954801 | 0.954801 | 1.0 |
| NO (UNCHECKED)33 -> NO195 | 0.530291 | 0.530291 | 1.0 |
| NO150 -> NO195 | 0.520525 | 0.520525 | 1.0 |
| NO (UNCHECKED)194 -> NO195 | 0.517501 | 0.517501 | 1.0 |
| YES5 -> NO195 | 0.517155 | 0.517155 | 1.0 |
| YES12 -> NO195 | 0.515772 | 0.515772 | 1.0 |
| NO61 -> NO195 | 0.514303 | 0.514303 | 1.0 |
| YES51 -> NO195 | 0.513957 | 0.513957 | 1.0 |
| NO52 -> NO195 | 0.51318 | 0.51318 | 1.0 |
| YES158 -> NO195 | 0.513093 | 0.513093 | 1.0 |
| YES135 -> NO195 | 0.512661 | 0.512661 | 1.0 |
| YES (CHECKED)99 -> NO195 | 0.512661 | 0.512661 | 1.0 |
| YES173 -> NO195 | 0.512575 | 0.512575 | 1.0 |
| NO59 -> NO195 | 0.512575 | 0.512575 | 1.0 |
| YES180 -> NO195 | 0.512488 | 0.512488 | 1.0 |

Наиболее достоверное правило для первой группы аналогично наиболее достоверному правилу второй группы штата Колорадо, которое заключается в следующем: женщина не могла согласиться на ВИЧ тестирование, потому что ей его не предлагали.

Условиями того, что первой группе Нью-Йорк Сити не предлагали пройти тестирование на ВИЧ, являются:

• мать не получала страховое покрытие Medicaid для дородового ухода;

• мать не сообщала об употреблении алкогольных напитков в течение последних 3 месяцев беременности;

• женщина отказывалась от теста на ВИЧ, потому что не хотела, чтобы другие думали, что она подвержена риску заражения ВИЧ;

• мать укладывает спать ребёнка на спину;

• в первую неделю после выписки из больницы ребёнка осматривал медицинский работник;

• в течение 12 месяцев до беременности бывший муж или бывший партнёр не причинял физический вред;

• медицинский работник говорил о том, что делать, если произойдут преждевременные роды;

• женщина не получала дородовой уход так рано, как хотела;

• мать курила в течение последних трёх месяцев беременности;

• более чем за 3 недели до родов у матери отошли воды;

• женщина отказалась от использования послеродовой контрацепции, так как не думает, что может забеременеть (стерильна);

• во время беременности стоматолог говорил о том, как ухаживать за зубами и дёснами;

• во время беременности муж или партнёр не причинял физический вред;

• до беременности женщина принимала отпускаемые по рецепту лекарства (не противозачаточные).

Несмотря на высокую поддержку и достоверность, которые удовлетворяют поставленной математической задаче, значение лифта для вышеперечисленных правил равно 1, что говорит о том, то связь между условием и следствием отсутствует.

Ассоциативные правила для второй группы респондентов Нью-Йорк Сити находятся в таблице 3.7.

1. Ассоциативные правила для группы 2 Нью-Йорк Сити

| **Правило** | **Поддержка** | **Достоверность** | **Лифт** |
| --- | --- | --- | --- |
| YES196 -> YES195 | 0.953982 | 0.953982 | 1.0 |
| NO (UNCHECKED)92 -> YES195 | 0.519646 | 0.519646 | 1.0 |
| NO (UNCHECKED)34 -> YES195 | 0.518053 | 0.518053 | 1.0 |
| YES45 -> YES195 | 0.517522 | 0.517522 | 1.0 |
| NO (UNCHECKED)110 -> YES195 | 0.515221 | 0.515221 | 1.0 |
| NO87 -> YES195 | 0.514602 | 0.514602 | 1.0 |
| YES188 -> YES195 | 0.514248 | 0.514248 | 1.0 |
| YES73 -> YES195 | 0.513805 | 0.513805 | 1.0 |
| NO (UNCHECKED)103 -> YES195 | 0.513097 | 0.513097 | 1.0 |
| YES13 -> YES195 | 0.512743 | 0.512743 | 1.0 |
| NO (UNCHECKED)31 -> YES195 | 0.512655 | 0.512655 | 1.0 |
| YES (CHECKED)83 -> YES195 | 0.512389 | 0.512389 | 1.0 |
| YES195 -> YES24 | 0.512035 | 0.512035 | 1.0 |
| YES195 -> YES167 | 0.511947 | 0.511947 | 1.0 |
| YES195 -> YES67 | 0.511504 | 0.511504 | 1.0 |

Вторая группа Нью-Йорк Сити наиболее достоверным правилом похожа на вторую группу штата Колорадо.

Для второй группы условием предложения женщине, не проходившей тестирование на ВИЧ, пройти его являются:

• женщина не использовала противозачаточные средства при зачатии, так как думала, что она или партнёр бесплодны;

• незадолго до того, как женщина забеременела, она была на программе Medicaid;

• медицинский работник говорил о том, какой вес женщина должна набрать во время беременности;

• медицинская услуга оплачена страховкой, приобретённой не напрямую;

• женщина не пользуется послеродовой контрацепцией;

• в течение 3 месяцев до того, как женщина забеременела новым ребёнком, у неё были проблемы с щитовидной железой;

• были зарегистрированы стрессы, связанные с травмой;

• во время посещений дородового наблюдения медицинский работник говорил, что употребление алкоголя во время беременности может повлиять на вашего ребёнка;

• источник дохода домохозяйства – выплаты работникам социальной защиты, ветеранам, пособия или пенсии.

Также во вторую группу попало несколько правил, где тестирование на ВИЧ является не следствием, а условием. Если женщине предлагалось пройти тестирование на ВИЧ, то:

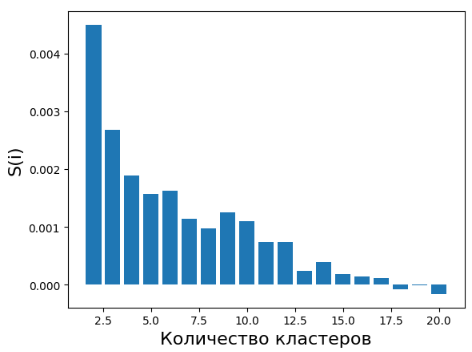
• младенец в настоящее время жив;

• за 3 месяца до того, как она забеременела новым ребёнком, у неё была депрессия;

• у женщины были зарегистрированы эмоциональные стрессоры.

Чтобы произвести сравнение данных Нью-Йорк Сити со штатом Нью-Йорк без Нью-Йорк Сити, необходимо произвести те же процедуры, что и двумя предыдущими штатами.

Оценки силуэта для штата Нью-Йорк (рисунок 3.5) показывают, что респонденты делятся на две группы.

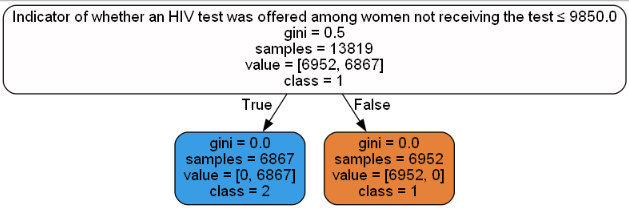


1. Оценки силуэта для различного количества кластеров для штата Нью-Йорк без Нью-Йорк Сити

После кластеризации респондентов в первую группу попало 8 681 мать, во вторую — 8 593.

Дерево классификации, представленное на рисунке 3.6 правильно предугадывает класс 100 % респондентов.

Коэффициент gini, равный 0.5 на первом слое, дерева говорит о том, то объекты равномерно распределены между группами. Значение коэффициента gini равное 0 означает, что данные принадлежат одному классу.



1. Дерево классификации для штата Нью-Йорк без Нью-Йорк Сити

Для двух групп респондентов штата Нью-Йорк без Нью-Йорк Сити также был произведён поиск ассоциативных правил. Для первой группы они расположены в таблице 3.8.

1. Ассоциативные правила для группы 1 штата Нью-Йорк без Нью-Йорк Сити

| **Правило** | **Поддержка** | **Достоверность** | **Лифт** |
| --- | --- | --- | --- |
| YES196 -> YES195 | 0.933648 | 0.933648 | 1.0 |
| NO119 -> YES195 | 0.525285 | 0.525285 | 1.0 |
| NO178 -> YES195 | 0.523442 | 0.523442 | 1.0 |
| YES195 -> YES (CHECKED)192 | 0.521253 | 0.521253 | 1.0 |
| NO (UNCHECKED)31 -> YES195 | 0.520447 | 0.520447 | 1.0 |
| YES13 -> YES195 | 0.518949 | 0.518949 | 1.0 |
| YES195 -> YES71 | 0.518489 | 0.518489 | 1.0 |
| YES87 -> YES195 | 0.518258 | 0.518258 | 1.0 |
| NO57 -> YES195 | 0.517797 | 0.517797 | 1.0 |
| NO165 -> YES195 | 0.516991 | 0.516991 | 1.0 |
| NO (UNCHECKED)30 -> YES195 | 0.51653 | 0.51653 | 1.0 |
| YES (CHECKED)78 -> YES195 | 0.516415 | 0.516415 | 1.0 |
| YES195 -> NO142 | 0.5163 | 0.5163 | 1.0 |
| Unintended18 -> YES195 | 0.514802 | 0.514802 | 1.0 |
| YES49 -> YES195 | 0.514572 | 0.514572 | 1.0 |

По первым 15 ассоциативным правилам первая группа штата Нью-Йорк без Нью-Йорк Сити уже отличается от всех ранее рассматриваемых групп. Однако В большинстве достоверных правил следствием является предложение женщине, не проходившей тестирование на ВИЧ, пройти его. Условием для этого являются:

• женщина чаще всего ребёнка укладывают спать не на живот;

• проверки депрессии до беременности не производилась;

• во время посещений дородового наблюдения медицинский работник говорил о том, как употребление алкоголя во время беременности может повлиять на вашего ребёнка;

• женщина использует послеродовую контрацепцию;

• во время последней беременности никто не причинял женщине физический вред;

• женщина не чистила зубы во время последней беременности;

• в доме есть одна или несколько спален;

• незапланированная беременность;

• во время посещений дородового наблюдения медицинский работник говорил о том, чтобы сдать анализ крови на ВИЧ.

Также если женщине предложили пройти тест на ВИЧ, то:

- были зарегистрированы стрессоры, связанные с партнёром;

- женщина отказалась от тестирования на ВИЧ, так как не думала, что она подвержена риску заражения ВИЧ;

- в течение месяца до того, как женщина забеременела ребёнком, она не принимала ежедневно поливитамины.

Ассоциативные правила для второй группы представлены в таблице 3.9.

1. Ассоциативные правила для группы 2 штата Нью-Йорк без Нью-Йорк Сити

| **Правило** | **Поддержка** | **Достоверность** | **Лифт** |
| --- | --- | --- | --- |
| YES196 -> NO195 | 0.929594 | 0.929594 | 1.0 |
| YES188 -> NO195 | 0.524497 | 0.524497 | 1.0 |
| YES52 -> NO195 | 0.519027 | 0.519027 | 1.0 |
| NO66 -> NO195 | 0.518212 | 0.518212 | 1.0 |
| YES71 -> NO195 | 0.515187 | 0.515187 | 1.0 |
| NO189 -> NO195 | 0.515187 | 0.515187 | 1.0 |
| NO195 -> NO (UNCHECKED)99 | 0.514838 | 0.514838 | 1.0 |
| NO (UNCHECKED)55 -> NO195 | 0.514721 | 0.514721 | 1.0 |
| NO137 -> NO195 | 0.514256 | 0.514256 | 1.0 |
| YES (CHECKED)194 -> NO195 | 0.513907 | 0.513907 | 1.0 |
| 1 or more145 -> NO195 | 0.513208 | 0.513208 | 1.0 |
| NO173 -> NO195 | 0.512976 | 0.512976 | 1.0 |
| NO139 -> NO195 | 0.512859 | 0.512859 | 1.0 |
| NO25 -> NO195 | 0.512627 | 0.512627 | 1.0 |
| YES57 -> NO195 | 0.51251 | 0.51251 | 1.0 |

Для второй группы следствием зачастую является отсутствие предложения пройти тестирование на ВИЧ. Условиями для этого являются:

• в течение 3 месяцев до того, как женщина забеременела, у неё были проблемы с щитовидной железой;

• женщина получала дородовой уход так рано, как хотела;

• за 12 месяцев до рождения ребёнка мать не расходилась и не разводилась со своим мужем или партнёром;

• были зарегистрированы стрессоры, связанные с партнёром;

• с тех пор, как родился ребёнок, женщина не проходила послеродовой осмотр;

• дородовой уход оплачивался за счёт страховки, приобретённой не напрямую;

• мать не сообщала о диабете, развивающемся до недавней беременности;

• отказ женщины от теста на ВИЧ по причине того, то она не хотела, чтобы другие думали, что она подвержена риску заражения ВИЧ;

• у женщины было одно или более живорождений;

• стоматолог не говорил с женщиной о том, как ухаживать за зубами и дёснами;

• женщина не слышала или читала, что приём витамина фолиевой кислоты может помочь предотвратить некоторые врождённые дефекты;

• о финансовых стрессорах не сообщалось;

• во время последней беременности женщине причиняли физический вред.

Также в данной группе отличительным следствием выделяется правило: если женщине, не прошедшей тестирование на ВИЧ, не предлагали его пройти, то причиной отказа её от использования послеродовой контрацепции является то, что она не думает, что может забеременеть (стерильна).

Исходя из следствий полученных ранее ассоциативных правил, логично сравнивать группу 2 штата Нью-Йорк без Нью-Йорк Сити (NY) с группой 1 Нью-Йорк Сити (NYC), так как обеим этим группам не предложили тестирование на ВИЧ, и группу 1 штата Нью-Йорк без Нью-Йорк Сити с группой 2 Нью-Йорк Сити, которым предложили тестирование.

Группа 2 штата Нью-Йорк без Нью-Йорк Сити имеет факторы, влияющие на отсутствие предложения пройти тест на ВИЧ, отличные от группы 1 Нью-Йорк Сити. Однако было замечено существенное различие этих групп. Женщины Нью-Йорк Сити не получили предложение на тестирование, потому то не получали дородовой уход так рано, как хотели. Тогда как матери штата Нью-Йорк без Нью-Йорк Сити получали.

У матерей 2 группы штата NY отмечается наличие нескольких негативных условий, связанных с партнёром: несмотря на стрессоры, связанные с ним и физический вред во время беременности, они не уходили от партнёров. Также эти женщины хоть и получали дородовой уход так рано, как хотели, они меньше пользовались медицинскими услугами: с ними не разговаривал стоматолог об уходе за зубами (возможно они его не посещали), не слышали о приёме фолиевой кислоты для предотвращения врождённых дефектов и не проходили послеродовой осмотр.

Женщины 1 группы NYC хоть и не имеют социальной страховки Medicaid, получали консультации и лечение врачей, а их дети проходили осмотр в первую неделю после выписки.

Также стоит выделить, что обе группы отказывались от тестирования на ВИЧ.

У 2 группы NYC есть схожее правило с 1 группой NY: если женщине говорили, как употребление алкоголя во время беременности может повлиять на вашего ребёнка, то ей предложили пройти тестирование на ВИЧ. Однако жительницы NYC не используют послеродовую контрацепцию, а NY — используют, и независимо от этого им предлагали тестирование на ВИЧ. Также отличительными чертами 2 группы NYC являются источник дохода — пособия или пенсии и участие в социальной программе Medicaid, тогда как жительницы NY имеют в доме одну или несколько спален, однако из-за предложения тестирования на ВИЧ у них были зарегистрированы стрессоры, связанные с партнёром. Эта группа также отказывалась от тестирования на ВИЧ. Консультации у врачей получали обе группы.

В общем можно выделить, что женщины штата Нью-Йорк без Нью-Йорк Сити имеют больше проблем со своими партнёрами. В этом штате матери, которые получали консультации специалистов, также получали предложение о прохождении тестирования. Если женщина на получала этого предложения, то консультаций она тоже не получала.

Женщины Нью-Йорк Сити получали консультации врачей. Женщины, получавшие предложение о тестировании на ВИЧ, имели социальную поддержку: страховку, выплаты.

Для Нью-Йорк Сити штата Нью-Йорк без Нью-Йорк Сити нужно рассматривать разные программы здравоохранения и поддержки матерей, чтобы повысить количество предложений на прохождение ВИЧ тестирования и уменьшить процент отказов от него (в 3-ёх группа их 4 были отмечены отказы).

## Возможности использования прогнозирования материнского поведения в различных сферах

Выявление причинно-следственных связей в материнском поведении позволяет:

• государству выстроить более эффективную социальную политику в отношении матерей (внедрение социальных программ, профилактических мероприятий и т.д.);

• фармацевтическим компаниям разрабатывать маркетинговую стратегию по продаже витаминов;

• производителям детских продуктов (авто кресел, детского питания и т.д.) также более эффективно использовать рекламу;

• производителям контента для беременных и мам улучшить качество продукта.

Используемый в работе датасет довольно старый, однако сравнение результатов прошлых годов с текущими (на сегодняшний день неопубликованными в открытом доступе) даёт возможность оценить результативность социальной политики государства в отношении матерей.

Также полученные данные позволяют выявить проблемы в здравоохранении в различных регионах. Например, было выявлено, что в штате Нью-Йорк и Нью-Йорк Сити женщины отказываются от тестирования на ВИЧ, что может навредить как здоровью матери, так и ребёнка.

Данные 2011 года пригодятся для сравнения материнского поведения детей и родителей, а также эти данные помогут выявить факторы, из-за которых будущее поколение откажется материнства или наоборот, захочет родить детей.

**Вывод**

В этой главе были рассмотрены причинно-следственные связи в поведении матерей трёх штатов. Для всех трёх штатов было определено оптимальное количество кластеров равное 2. Правило классификации респондентов также во всех трёх штатах одинаковое — было ли предложено пройти тестирование на ВИЧ женщинам, которые его не прошли. Для каждой группы респондентов были рассмотрены условия положительного влияния на мать — предложение пройти тестирование на ВИЧ и отрицательного — отсутствие данного предложения. Также для второй группы штата Колорадо были рассмотрены условия наличия и отсутствия у женщины депрессии за 3 месяца до того, как она забеременела.

Было произведено сравнение данных для групп штата Нью-Йорк и Нью-Йорк Сити, чтобы определить, можно ли использовать для этих штатов одну политику в области здравоохранения для дородового ухода.

# ЗАКЛЮЧЕНИЕ

В результате данной работы были выявлены причинно-следственные связи в поведении матерей с помощью данных опросов штата Колорадо и Нью-Йорк вместе с Нью-Йорк Сити. В каждом штате было произведено разделение респондентов-матерей на группы для получения более достоверных ассоциативных правил. Для первой группы штата Колорадо, а также для штата Нью-Йорк и Нью-Йорк Сити было рассмотрено по 15 первых правил с наибольшим значением поддержки и достоверности. Для второй группы штата Колорадо помимо этого были рассмотрены правила, в которых следствием является наличие или отсутствие депрессии у женщины за 3 месяца до беременности.

Цель работы была достигнута. Для всех трёх штатов факторами, положительно влияющими на здоровье матерей, являются условия для предложения тестирования на ВИЧ женщине, которая его не проходила. Причины ухудшения состояния матерей в рассматриваемых штатах — отсутствие предложения пройти тест на ВИЧ. Для второй группы штата Колорадо к положительным факторам добавляются условия для отсутствия депрессии у матери за 3 месяца до беременности, к отрицательным — наличие депрессии.

В рамках данной работы было рассмотрено только три штата из 26. И для двух из них было проанализировано первые 15 наиболее достоверных правил. Для получения более полных данных можно проанализировать больше правил с разными вопросами-следствиями.

Но помимо этого можно использовать другие методики перевода ответов в числовые значения. Вместо умножения на номер вопроса, как это производилось в этой работе, можно каждому вопросу присвоить собственный коэффициент на этапе разработки вопроса. Также можно разбить вопросы на смысловые группы, в каждой из которых подсчитывать балл, и кластеризовать респондентов, используя меньшее количество параметров.

Получаемые в результате применяемого способа знания дают возможность выявить факторы, позволяющие улучшить протекание беременности, родов и грудного вскармливания у матерей.

Способ обработки анализируемого набора данных позволяет получить новые знания не только в области материнского поведения Подобную обработку опросов можно использовать и в политике для анализа политических взглядов, чтобы предотвратить экстремистскую деятельность, и в муниципальном управлении для определения нужд и желаний жителей территориальной единицы различного уровня, чтобы оптимально и разумно расходовать бюджет.

# СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Мицюк Н.А. Проблема материнства в современных зарубежных исторических исследованиях // Вестник Тверского государственного университета. Серия «История». — 2015 — №2. — С. 124-134.
2. Богдашина И.В. Повседневные практики матерей провинциального советского города в 1950-1960-е г.г. // Вестник Костромского государственного университета — 2020 — №2. — С. 42-51.
3. Ермолова Т.В., Иволина Т.В., Дедова О.В., Литвинов А.В. Проблема дисфункционального материнства в новейших зарубежных исследованиях // Современная зарубежная психология. — 2019 — Том 8. № 4. — С. 25–37.
4. Куликов Л.В., Малёнова А.Ю., Потапова Ю.В. Субъективная картина материнства в российских и зарубежных исследованиях // [Вестник Московского университета. Серия 14. Психология](https://cyberleninka.ru/journal/n/vestnik-moskovskogo-universiteta-seriya-14-psihologiya) — 2020 — №4. — С. 135-167.
5. Jones C.W., Esteves K.C., Gray S.A.O., Clarke T.N., Callerame K., Theall K.P., Drury S.S. The transgenerational transmission of maternal adverse childhood experiences (ACEs): Insights from placental aging and infant autonomic nervous system reactivity // Psychoneuroendocrinology — 2019 — Vol.106 — P. 20-27.
6. Orchard E.R., Voigt K., Chopra S., Thapa T., Ward P.G.D., Egan G.F., Jamadar S.D. The maternal brain is more flexible and responsive at rest: effective connectivity of the parental caregiving network in postpartum mothers // Scientific Reports — 2023 — №13 — P. 1-13
7. Xu-Hong Yang. Online Collection and Intelligent Analysis System for Postpartum Depression and Social Support of Second-Born Mothers Based on Big Data // [6th International Conference on Computing Methodologies and Communication (ICCMC)](https://ieeexplore.ieee.org/xpl/conhome/9752772/proceeding) — 2022
8. Шапиро М.Г. Классификация методов опроса в социологии // Социальные исследования — 2017 — №2. — С. 51-59.
9. Сташкова О.В., Шестопал О.В. Использование искусственных нейронных сетей для восстановления пропусков в массиве исходных данных // [Известия высших учебных заведений. Северо-Кавказский регион. Технические науки](https://cyberleninka.ru/journal/n/izvestiya-vysshih-uchebnyh-zavedeniy-severo-kavkazskiy-region-tehnicheskie-nauki) — 2017 — №1. — С. 37-42
10. Фомина Е.Е. Применение аппарата нейронных сетей для анализа результатов анкетирования // Вестник Пермского национального исследовательского политехнического университета. Социально-экономические науки — 2020 — №2. — С. 99-110
11. Корнаухова М.А., Основин С.С., Баскаков Е.В., Савин Е.С., Полусмакова Н.С. Информационная система помощи выбора специальности для абитуриентов на основе нейронной сети // Инженерный вестник Дона — 2022 — №11.
12. Атнабаева А.Р. Анализ поведения школьников при выборе ВУЗа на основе метода дерева решений // Экономика и бизнес: теория и практика — 2021 — №82 — С. 42-44
13. Спирина А.С. Социальное доверие как фактор восприятия коррупции (по результатам социологического опроса в Алтайском крае) // Society and Security Insights — 2021 — №2 — С. 106-114
14. Токарь О. В. Опыт разработки ассоциативных правил в области студенческого чтения // Интеграция и развитие научно-технического и образовательного сотрудничества - взгляд в будущее : сборник статей II Междунар. научно-техн. конф. "Минские научные чтения - 2019" — 2019 — : в 3 т. Т. 1. — С. 129-132.
15. Самойлова Т.А., Грибер Ю.А. Интеллектуальный анализ цветовых предпочтений: поиск ассоциативных правил VS кластерный анализ // Мир науки. Педагогика и психология — 2020 — №6.
16. Shulman H.B., D’Angelo D.V., Harrison L., Smith R., Warner L. Pregnancy Risk Assessment Monitoring System (PRAMS): Overview of Design and Methodology // [American Journal of Public Health](https://www.researchgate.net/journal/American-Journal-of-Public-Health-1541-0048) — 2018 — №108 — P. e1-e9
17. Кубегенова, А. Д. Использование методов и основных алгоритмов в технологии Data Mining// Исследования молодых ученых : материалы IV Междунар. науч. конф. (г. Казань, ноябрь 2019 г.). — 2019 — С. 8-11.
18. Силуэт кластера (Cluster silhouette) [Электронный ресурс].- URL: <https://wiki.loginom.ru/articles/cluster-silhouette.html> (дата обращения: 31.05.2023)
19. Introduction to Self-Organizing Maps [Электронный ресурс].- URL: <https://rubikscode.net/2018/08/20/introduction-to-self-organizing-maps/> (дата обращения: 25.05.2023)
20. Деревья решений — CART математический аппарат. [Электронный ресурс].- URL: <https://basegroup.ru/community/articles/math-cart-part1> (дата обращения: 16.05.2023)
21. Apriori — масштабируемый алгоритм поиска ассоциативных правил [Электронный ресурс].- URL: <https://loginom.ru/blog/apriori#apriori> (дата обращения: 16.05.2023)
22. TIOBE Index for May 2023 [Электронный ресурс].- URL: <https://www.tiobe.com/tiobe-index/> (дата обращения: 14.05.2023)
23. Официальный сайт с документацией библиотеки NumPy [Электронный ресурс].- URL: <https://numpy.org/doc/stable/user/whatisnumpy.html> (дата обращения: 14.05.2023)
24. Официальный сайт с документацией библиотеки SciPy [Электронный ресурс].- URL: <http://scipy.github.io/devdocs/> (дата обращения: 15.05.2023)
25. Руководство пользователя библиотеки matplotlib [Электронный ресурс].- URL: <https://matplotlib.org/stable/users/index.html> (дата обращения: 15.05.2023)
26. Руководство пользователя библиотеки scikit-learn [Электронный ресурс].- URL: <https://scikit-learn.org/stable/user_guide.html> (дата обращения: 15.05.2023)
27. About pandas. Официальный сайт библиотеки pandas с документацией [Электронный ресурс].- URL: <https://pandas.pydata.org/about/> (дата обращения: 15.05.2023)
28. Python Developers Survey 2021 Results [Электронный ресурс].- URL: <https://lp.jetbrains.com/python-developers-survey-2021/#PythonPackaging> (дата обращения: 15.05.2023)

# ПРИЛОЖЕНИЕ А

**Полный список вопросов, заданных матерям**

1. Во время какого-либо из ваших посещений дородового ухода говорил ли с вами медицинский работник о том, что делать, если вы чувствуете депрессию во время беременности или после рождения ребёнка?
2. Во время какого-либо из ваших посещений дородового ухода говорил ли с вами медицинский работник о проведении анализов для выявления врождённых дефектов или болезней, которые распространены в вашей семье?
3. История предыдущих живорождений
4. Адекватность пренатальной помощи (индекс Кесснера)
5. Показатель того, чаще всего ребёнка укладывают спать на спину
6. За 3 месяца до того, как вы забеременели новым ребёнком, у вас была астма?
7. За 12 месяцев до рождения ребёнка вы переехали на новый адрес
8. За 12 месяцев до рождения вашего ребёнка умер кто-то очень близкий вам
9. (\*PCH) Индикатор неиспользования противозачаточных средств при зачатии
10. За 12 месяцев до рождения вашего ребёнка ваш муж или партнёр потерял работу
11. Незадолго до того, как вы забеременели, когда вы хотели забеременеть?
12. Осматривал ли вашего ребёнка медицинский работник в первую неделю после выписки из больницы?
13. Во время какого-либо из ваших посещений дородового наблюдения говорил ли с вами медицинский работник о том, как употребление алкоголя во время беременности может повлиять на вашего ребёнка?
14. Масса тела ребёнка классифицируется как низкая масса тела при рождении (НМТ), если масса тела меньше или равна 2500 г, или нормальная масса тела при рождении (НМТ), если масса тела превышает 2500 г.
15. Как родился ваш новый ребёнок?
16. В доме есть работающая дымовая сигнализация.
17. групп доходов (2004 г. и последующие годы)
18. (\*PCH) Индикатор предполагаемой беременности
19. Индикатор рождения ребёнка в больнице
20. Индикатор отсутствия дородовой помощи
21. Источник дородовой помощи (государственный/частный)
22. Индикатор родов в больнице
23. Подтверждение беременности после первого триместра
24. Индикатор того, что младенец в настоящее время жив
25. Индикатор любых финансовых стрессоров, о которых сообщалось
26. За 12 месяцев до рождения вашего ребёнка вы были бездомной
27. (\*PCH) Индикатор адекватной социальной поддержки после родов
28. Индикатор того, что когда-либо чистили зубы
29. Источник дохода домохозяйства – деньги семьи или друзей
30. Источник дохода домохозяйства – пособие на ребёнка или алименты
31. Источник дохода домохозяйства – другой источник (2004-2011 годы)
32. Индикатор того, оплатила ли Medicaid предоставление медицинской услуги
33. Показатель того, получала ли мать страховое покрытие Medicaid для дородового ухода.
34. Незадолго до того, как вы забеременели, вы были на Medicaid? (2009 - 2011 годы)
35. Показатель того, была ли мать получателем Medicaid в любое время (во время беременности для дородового наблюдения или во время родов).

Продолжение ПРИЛОЖЕНИЯ А

1. Во время беременности вы участвовали в программе WIC (Специальная программа дополнительного питания для женщин, младенцев и детей)?
2. Показатель позднего (после первого триместра) или отсутствия обращения за дородовой помощью
3. Куда вы чаще всего ходили на дородовые визиты?
4. (\*PCH) Показатель охвата медицинским страхованием за месяц до беременности (2009 - 2011 гг.)
5. Во время какого-либо из ваших посещений дородового ухода говорила ли с вами медсестра или другой медицинский работник о грудном вскармливании вашего ребёнка?
6. Во время какого-либо из ваших посещений дородового ухода говорил ли с вами медицинский работник о том, как употребление запрещённых наркотиков может повлиять на вашего ребёнка?
7. Во время какого-либо из ваших посещений дородового ухода говорил ли с вами медицинский работник о лекарствах, которые безопасно принимать во время беременности?
8. Показатель неполучения дородового ухода в кратчайшие сроки среди женщин, начавших дородовой уход поздно или не начавших его вовсе
9. Показатель того, оплачивается ли дородовой уход из других источников
10. Во время какого-либо из ваших посещений дородового ухода говорил ли с вами медицинский работник о том, какой вес вы должны набрать во время беременности?
11. Во время любого из ваших посещений дородового ухода говорил ли с вами медицинский работник о признаках и симптомах преждевременных родов?
12. Во время какого-либо из ваших посещений дородового ухода говорил ли с вами медицинский работник о физическом насилии над женщинами со стороны их мужей или партнёров?
13. Во время любого из ваших посещений дородового ухода говорил ли с вами медицинский работник об использовании ремней безопасности во время беременности?
14. Во время какого-либо из ваших посещений дородового ухода говорил ли с вами медицинский работник о том, чтобы сдать анализ крови на ВИЧ?
15. Во время какого-либо из ваших посещений дородового ухода говорил ли с вами медицинский работник о том, как курение во время беременности может повлиять на вашего ребёнка?
16. Во время какого-либо из ваших посещений дородового ухода говорил ли с вами медицинский работник о том, что делать, если у вас преждевременные роды?
17. Получали ли вы дородовой уход так рано, как хотели? (2009-2011 годы)
18. Показатель того, оплачивался ли дородовой уход за счет страховки через военную службу
19. Индикатор отсутствия страховки для оплаты дородового ухода
20. Показатель того, оплачивался ли дородовой уход за счет страховки, приобретённой напрямую
21. Показатель того, оплачивался ли дородовой уход за счет страховки через работодателя
22. Во время вашей последней беременности причинял ли вам кто-либо физический вред каким-либо образом?
23. (\*PCH) В течение 12 месяцев до того, как вы забеременели, ваш муж или партнёр толкал, пинал, душил или причинял вам физический вред каким-либо другим способом?

Продолжение ПРИЛОЖЕНИЯ А

1. Во время вашей последней беременности ваш муж или партнёр причинял вам физический вред каким-либо образом?
2. За 12 месяцев до того, как вы забеременели, причинял ли вам кто-нибудь физический вред?
3. В течение 12 месяцев до того, как вы забеременели, ваш бывший муж или бывший партнёр душил вас или причинял физический вред каким-либо другим способом?
4. Во время вашей последней беременности бывший муж или бывший партнёр толкал, пинал, душил или физически причинял вам какие-либо другие травмы?
5. За 12 месяцев до рождения вашего ребёнка вы ссорились с мужем или партнёром чаще, чем обычно.
6. За 12 месяцев до рождения вашего ребёнка у вас было много счетов, которые вы не могли оплатить.
7. За 12 месяцев до рождения вашего ребёнка у кого-то из ваших близких были серьёзные проблемы с алкоголем или наркотиками.
8. За 12 месяцев до рождения ребёнка вы разошлись или развелись со своим мужем или партнёром.
9. Индикатор любых зарегистрированных эмоциональных стрессоров
10. За 12 месяцев до рождения вашего ребёнка близкий член семьи был очень болен, и ему пришлось лечь в больницу.
11. За 12 месяцев до рождения вашего ребёнка вы были в физической борьбе
12. За 12 месяцев до рождения вашего ребёнка вы или ваш муж или партнёр попали в тюрьму
13. Индикатор любых зарегистрированных стрессоров, связанных с партнёром
14. Общее количество зарегистрированных стрессоров, сгруппированных как отсутствие 1-2 3-5 и 6-13.
15. Индикатор любых зарегистрированных стрессоров, связанных с травмой
16. За 12 месяцев до рождения вашего ребёнка вы потеряли работу, хотя хотели продолжать работать
17. Индикатор младенца, проживающего в настоящее время с матерью
18. За 12 месяцев до рождения вашего ребёнка ваш муж или партнёр сказал, что не хочет, чтобы вы были беременны.
19. Индикатор того, есть ли в доме одна или несколько ванных комнат
20. Индикатор наличия в доме одной или нескольких спален
21. Индикатор того, есть ли в доме кухня
22. Источник дохода домохозяйства – помощь, такая как пособие TANF или SSI (годы 2009-2011)
23. Источник дохода домохозяйства – деньги от комиссионных, дивидендов или дохода от сдачи в аренду
24. Источник дохода домохозяйства - зарплата или деньги с работы
25. Источник дохода домохозяйства – выплаты работникам социальной защиты ветеранам, пособия или пенсии (2000-2003 гг.)
26. Когда вы забеременели новым ребёнком, вы пытались забеременеть?
27. (\*PCH) Индикатор нежелательной беременности
28. Причина неиспользования противозачаточных средств при зачатии – проблемы с противозачаточными средствами
29. (\*PCH) Индикатор использования послеродовой контрацепции
30. Причина неиспользования противозачаточных средств при зачатии – муж или партнёр не хотел
31. Причина неиспользования противозачаточных средств при зачатии – какая-либо другая причина

Продолжение ПРИЛОЖЕНИЯ А

1. Причина отказа от контрацепции при зачатии – не возражала против беременности
2. Причина неиспользования противозачаточных средств при зачатии - были побочные эффекты от контроля над рождаемостью
3. Причина неиспользования противозачаточных средств при зачатии - думала, что она или партнёр бесплодны
4. Причина, по которой не использовались противозачаточные средства при зачатии – думала, что в то время она не могла забеременеть.
5. (\*PCH) Индикатор неиспользования противозачаточных средств при зачатии для тех, кто не пытается забеременеть
6. Показатель использования противозачаточных средств во время беременности среди женщин с незапланированной беременностью
7. Индикатор использования противозачаточных средств во время беременности (для тех, кто не пытается забеременеть)
8. Причина неиспользования послеродовой контрацепции – муж или партнёр не хочет
9. Причина отказа от послеродовой контрацепции - отсутствие секса
10. Причина отказа от использования послеродовой контрацепции - не думает, что может забеременеть (стерильна)
11. (\*PCH) Индикатор приема каких-либо лекарств от бесплодия или лечения
12. Причина неиспользования послеродовой контрацепции – какая-то другая причина
13. Причина неиспользования послеродовой контрацепции - невозможность оплатить противозачаточные средства
14. Причина неиспользования послеродовой контрацепции - в настоящее время беременна
15. Причина отказа от послеродовой контрацепции – желание забеременеть
16. Причина неиспользования послеродовой контрацепции - не хочет использовать противозачаточные средства
17. (\*PCH) Индикатор приема каких-либо лекарств от бесплодия или лечения (только для тех, кто пытается забеременеть)
18. Показатель того, платили ли другие источники за предоставление медицинских услуг
19. Показатель того, была ли медицинская услуга оплачена страховкой через военную службу
20. Индикатор отсутствия страховки для оплаты медицинских услуг
21. Индикатор того, была ли медицинская услуга оплачена страховкой, приобретенной напрямую
22. Показатель того, была ли медицинская услуга оплачена страховкой через работодателя
23. Индикатор выписки ребёнка из больницы в течение 48 часов после рождения. (только для вагинальных родов)
24. Показатель осмотра младенцев в течение 1 недели после выписки из больницы для младенцев, выписанных в течение 48 часов
25. Как вы чаще всего укладываете ребёнка спать сейчас?
26. Куда вы обычно берете своего ребёнка для осмотра?
27. После рождения вашего ребёнка поместили в отделение интенсивной терапии?
28. Как часто ваш новый ребёнок спит в одной кровати с вами или с кем-либо еще?
29. Малыш всегда или почти всегда ездит в детском автокресле
30. Показатель того, чаще всего ребёнка укладывают спать на живот

Продолжение ПРИЛОЖЕНИЯ А

1. Показатель того, спит ли младенец обычно в общей кровати.
2. (\*PCH) Индикатор того, разрешено ли сейчас курение в доме
3. Индикатор того, кормила ли мать грудью через 4 недели после родов.
4. Индикатор того, кормила ли мать грудью через 8 недель после родов.
5. Кормили ли вы когда-нибудь грудью или сцеживали грудное молоко, чтобы кормить новорожденного после родов?
6. В доме есть заряженные ружья, винтовки или другое огнестрельное оружие.
7. ребёнка привезли домой из больницы в детском автокресле.
8. Были ли у вас вагинальные кровотечения во время беременности?
9. Были ли вы травмированы в автокатастрофе во время беременности?
10. Приходилось ли зашивать шейку матки (несостоятельный серкляж шейки матки) во время беременности?
11. Были ли у вас инфекции почек или мочевого пузыря (мочевыводящих путей) во время беременности?
12. Были ли у вас преждевременные или ранние роды во время беременности?
13. Были ли у вас сильная тошнота, рвота или обезвоживание во время беременности?
14. Вам делали переливание крови во время беременности?
15. Были ли у вас проблемы с плацентой (например, отслойка плаценты и предлежание плаценты) во время беременности?
16. Отошли ли воды более чем за 3 недели до родов (преждевременный разрыв плодных оболочек PROM) во время беременности?
17. Показатель того, сообщала ли мать о гестационном диабете (2009–2011 годы)
18. (\*PCH) Показатель того, сообщала ли мать о диабете, развившемся до недавней беременности (2009–2011 годы)
19. Было ли у вас высокое кровяное давление во время беременности?
20. Вы когда-нибудь слышали или читали, что прием витамина фолиевой кислоты может помочь предотвратить некоторые врожденные дефекты?
21. За месяц до того, как вы забеременели новым ребёнком, сколько раз в неделю вы принимали поливитамины?
22. Показатель того, принимала ли мать витамины более 4 раз в неделю в течение месяца до беременности
23. (\*PCH) В течение месяца до того, как вы забеременели новым ребёнком, принимали ли вы ежедневно поливитамины?
24. (\*PCH) Индикатор предыдущих преждевременных родов
25. (\*PCH) Индикатор смерти плода в анамнезе, мертворождения или выкидыша
26. Количество предыдущих живорождений (паритет матери классифицируется как отсутствие живорождений в прошлом или одно или несколько живорождений в прошлом).
27. Показатель употребления алкоголя в течение последних двух лет
28. (\*PCH) Показатель пьянства (4+ порции) в течение 3 месяцев до беременности
29. (\*PCH) Показатель употребления алкоголя за три месяца до беременности
30. Сколько сигарет вы выкуриваете в среднем в день сейчас?
31. Показатель того, сообщала ли мать об употреблении алкогольных напитков в течение последних 3 месяцев беременности.
32. Изменения в употреблении алкоголя между тремя месяцами до беременности и последними тремя месяцами беременности
33. За 3 месяца до того, как вы забеременели, сколько сигарет вы выкуривали в среднем в день?

Продолжение ПРИЛОЖЕНИЯ А

1. Сколько сигарет вы выкуривали в среднем в день за последние 3 месяца беременности?
2. Изменение курения с последних трех месяцев беременности к послеродовому периоду
3. Изменение курения за три месяца до беременности до послеродового периода
4. Изменения в курении с трех месяцев до беременности до последних трех месяцев беременности
5. (\*PCH) Показатель того, курила ли мать в течение трех месяцев до беременности
6. Показатель того, курила ли мать в течение последних трех месяцев беременности
7. Индикатор того, курит ли мать в настоящее время
8. Курили ли Вы сигареты за последние 2 года? (2009-2011 годы)
9. Показатель отказа матери от курения во время беременности (только для курильщиков)
10. (\*PCH) Индекс массы тела матери до беременности, классифицированный как недостаточная масса тела, нормальная избыточная масса тела или ожирение.
11. Индикатор проверки диабета до беременности
12. Индекс массы тела матери до беременности, классифицированный как страдающий ожирением (ИМТ >= 30) или без ожирения (ИМТ < 30)
13. Чистили ли вы зубы во время последней беременности?
14. Индекс массы тела матери до беременности, классифицированный как страдающий ожирением (ИМТ > 29) или без ожирения (ИМТ <= 29)
15. За 3 месяца до того, как вы забеременели новым ребёнком, у вас была депрессия?
16. (\*PCH) Показатель того, сообщала ли мать о частых симптомах послеродовой депрессии (2009–2011 годы)
17. Испытывали ли вы тревогу за 3 месяца до того, как вы забеременели?
18. Говорил ли с вами медицинский работник о детской хандре или послеродовой депрессии во время беременности или после родов?
19. (\*PCH) Показатель того, что мать чистила зубы за 12 месяцев до беременности (2009 - 2011 гг.)
20. Во время последней беременности вам приходилось обращаться к стоматологу по поводу проблемы?
21. Во время вашей последней беременности стоматолог или работник стоматолога говорил с вами о том, как ухаживать за зубами и деснами?
22. Во время последней беременности вы посещали стоматолога или стоматологическую клинику?
23. Индикатор диеты до беременности для похудения
24. Индикатор проверки артериального давления до беременности
25. Индикатор занятий спортом до беременности 3 или более дней в неделю
26. (\*PCH) Индикатор проверки депрессии до беременности
27. Индикатор обсуждения семейного анамнеза до беременности
28. Показатель приёма до беременности отпускаемых по рецепту лекарств, кроме противозачаточных
29. До того, как вы забеременели, говорила ли с вами медсестра или другой медицинский работник о том, как подготовиться к здоровой беременности и рождению ребёнка?
30. Причина отказа от теста на ВИЧ - боязнь получить результаты теста

Продолжение ПРИЛОЖЕНИЯ А

1. (\*PCH) В течение 3 месяцев до того, как вы забеременели новым ребёнком, было ли у вас высокое кровяное давление (гипертония)?
2. (\*PCH) Показатель получения консультации до зачатия (2009–2011 годы)
3. Были ли у вас проблемы с сердцем за 3 месяца до того, как вы забеременели?
4. За 3 месяца до того, как вы забеременели новым ребёнком, у вас была анемия (плохая кровь с низким содержанием железа)?
5. За 3 месяца до того, как вы забеременели новым ребёнком, у вас была эпилепсия (припадки)?
6. В течение 3 месяцев до того, как вы забеременели новым ребёнком, были ли у вас проблемы со щитовидной железой?
7. (\*PCH) С тех пор, как родился ваш новый ребёнок, проходили ли вы послеродовой осмотр?
8. Причина отказа от тестирования на ВИЧ - ранее тестирование на ВИЧ
9. Причина отказа от теста на ВИЧ – другая причина
10. Причина отказа от теста на ВИЧ - не думала, что она подвержена риску заражения ВИЧ
11. В какой-либо момент во время вашей последней беременности или родов вы проходили тест на ВИЧ?
12. Причина отказа от теста на ВИЧ - не хотела, чтобы другие думали, что она подвержена риску заражения ВИЧ
13. Показатель того, предлагалось ли пройти тест на ВИЧ женщинам, не прошедшим тест
14. Показатель того, было ли отказано в тестировании на ВИЧ среди женщин, которым предложили тест

# ПРИЛОЖЕНИЕ Б

**Листинг кода вывода графиков оценок силуэта для рассматриваемых в главе 2 методов кластеризации**

from sklearn.preprocessing import StandardScaler

import matplotlib.pyplot as plt

import numpy as np

import pandas as pd

from sklearn import preprocessing

from sklearn.cluster import AgglomerativeClustering

from sklearn.metrics import silhouette\_score

from sklearn.cluster import KMeans

from sklearn\_som.som import SOM

#df— записи набора данных

df\_state\_check = df.sample (n=5000 ,random\_state= 0 )

def agg\_clustering(data, num\_clusters, metric):

cluster\_model = AgglomerativeClustering(n\_clusters=num\_clusters,

affinity=metric,

linkage='average')

clusters = cluster\_model.fit\_predict(data)

score = silhouette\_score(data,

cluster\_model.labels\_,

metric='euclidean')

return clusters, score

scaler = preprocessing.StandardScaler()

df\_state\_check= scaler.fit\_transform(df\_state\_check)

y\_hd\_euclidean, euclidean\_score\_hd = agg\_clustering(df\_state\_check, 5, 'euclidean')

y\_hd\_l1, l1\_score\_hd = agg\_clustering(df\_state\_check, 5, 'l1')

y\_hd\_l2, l2\_score\_hd = agg\_clustering(df\_state\_check, 5, 'l2')

y\_hd\_manhattan, manhattan\_score\_hd = agg\_clustering(df\_state\_check, 5, 'manhattan')

y\_hd\_cosine, cosine\_score\_hd = agg\_clustering(df\_state\_check, 5, 'cosine')

silhouette\_scores\_hd = {'euclidean': euclidean\_score\_hd,

'l1': l1\_score\_hd,

'l2': l2\_score\_hd,

'manhattan': manhattan\_score\_hd,

'cosine': cosine\_score\_hd}

plt.bar(list(silhouette\_scores\_hd.keys()),

list(silhouette\_scores\_hd.values()),

width=0.4)

#Оценки силуэта для агломеративной кластеризации

k = [2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20]

silhouette\_scores\_ag = []

for i in range(0,len(k)):

ac = AgglomerativeClustering(n\_clusters = k[i], linkage='average', affinity='cosine')

silhouette\_scores\_ag.append(

silhouette\_score(df\_state\_check, ac.fit\_predict(df\_state\_check)))

# Вывод столбчатой диаграммы для различного количества кластеров

plt.bar(k, silhouette\_scores\_ag)

plt.xlabel('Количество кластеров', fontsize = 16)

plt.ylabel('S(i)', fontsize = 16)

plt.show()

Продолжение ПРИЛОЖЕНИЯ Б

#Оценки силуэта для метода К-средних

silhouette\_scores\_kmeans = []

for i in range(0,len(k)):

kmeans = KMeans(n\_clusters=k[i], random\_state=0, n\_init=10)

silhouette\_scores\_kmeans.append(

silhouette\_score(df\_state\_check, kmeans.fit\_predict(df\_state\_check)))

# Вывод столбчатой диаграммы для различного количества кластеров

plt.bar(k, silhouette\_scores\_kmeans)

plt.xlabel('Количество кластеров', fontsize = 16)

plt.ylabel('S(i)', fontsize = 16)

plt.show()

#Оценки силуэта для самоорганизующеся карты Кохонена

silhouette\_scores\_som = []

for i in range(0,len(k)):

som = SOM(m=k[i], n=1, dim=196, lr=1, sigma=1, max\_iter=3000)

som.fit(df\_state\_check, epochs=100, shuffle=True)

silhouette\_scores\_som.append(

silhouette\_score(df\_state\_check, som.fit\_predict(df\_state\_check)))

# Вывод столбчатой диаграммы для различного количества кластеров

plt.bar(k, silhouette\_scores\_som)

plt.xlabel('Количество кластеров', fontsize = 16)

plt.ylabel('S(i)', fontsize = 16)

plt.show()

# ПРИЛОЖЕНИЕ В

**Листинг кода для отображения карты, показывающей географию проведения опроса**

import numpy as np

import pandas as pd

import folium

df = pd.read\_csv('CDC\_PRAMStat\_Data\_2011.csv', delimiter=',', nrows = None)

df.dataframeName = 'CDC\_PRAMStat\_Data\_2011'

states=["AL","AK","AZ","AR","CA","CO","CT","DE","FL","GA","HI","ID","IL","IN","IA","KS","KY","LA","ME","MD","MA","MI","MN","MS","MO","MT","NE","NV","NH","NJ","NM","NY","NC","ND","OH","OK","OR","PA","RI","SC","SD","TN","TX","UT","VT","VA","WA","WV","WI","WY"]

count\_states =[]

for i in range(0,len(states)):

if df["LocationAbbr"].isin([states[i]]).any():

count\_states.append((df[df['LocationAbbr']==states[i]]).count()[0])

else:

count\_states.append(0)

#Реализация фрейма pandas из списковв

data = []

data.append(states)

data.append(count\_states)

state\_data = pd.DataFrame(data). transpose ()

state\_data.columns =['State', 'Respondents']

m = folium.Map(location = [40, -95],

zoom\_start = 4)

state\_geo = "us-states.json"

folium.Choropleth(

# geographical locations

geo\_data = state\_geo,

name = "choropleth",

# the data set we are using

data = state\_data,

columns = ["State", "Respondents"],

# YlGn refers to yellow and green

fill\_color = "YlGn",

fill\_opacity = 0.7,

line\_opacity = .1,

key\_on = "feature.id",

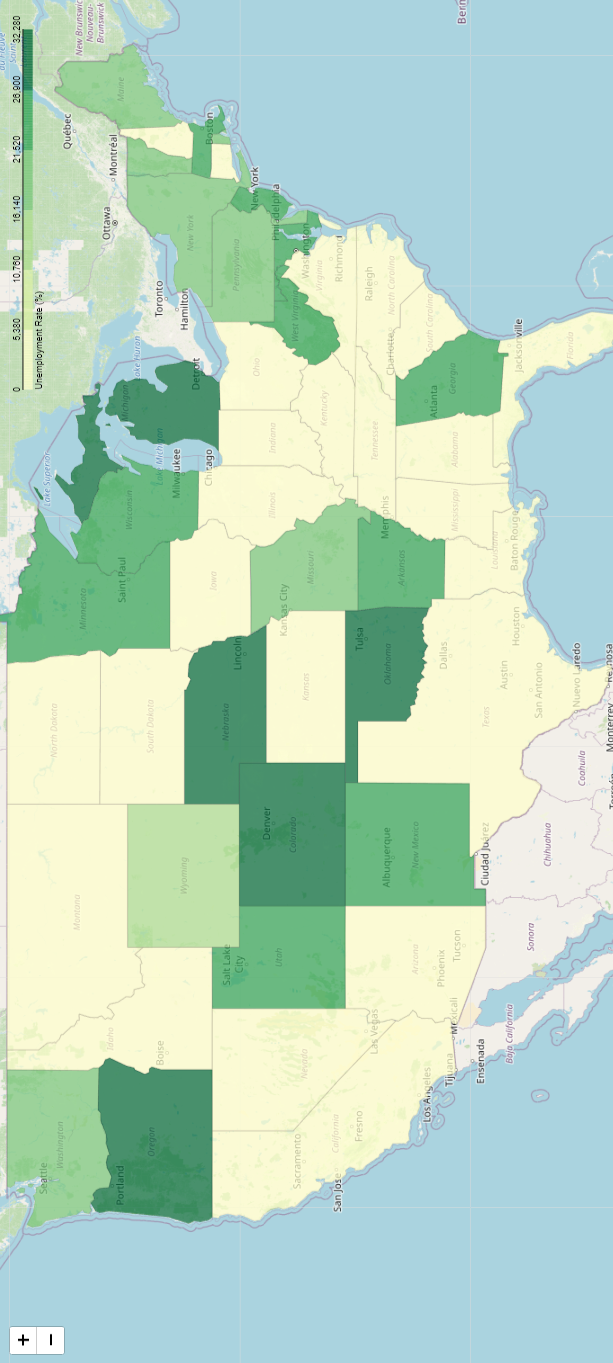
legend\_name = "Количество респондентов",

).add\_to(m)

m.save('states\_map.html')

# ПРИЛОЖЕНИЕ Г

**Отражение числа респондентов в каждом штате на карте**



# ПРИЛОЖЕНИЕ Д

**Листинг кода для разбиения респондентов-матерей штата Колорадо на группы, построения дерева классификации для этих групп и поиска ассоциативных правил в первой группе респондентов штата Колорадо**

import matplotlib.pyplot as plt

import numpy as np

import pandas as pd

from sklearn import preprocessing

from sklearn.metrics import silhouette\_score

from sklearn.metrics import accuracy\_score

from sklearn.cluster import KMeans

from collections import Counter

from sklearn import tree

import graphviz

import apyori

from apyori import apriori

import re

df = pd.read\_csv('CDC\_PRAMStat\_Data\_2011.csv', delimiter=',', nrows = None)

df.dataframeName = 'CDC\_PRAMStat\_Data\_2011'

respons\_yn={'Yes':100,

'YES':100,

'YES (CHECKED)':100,

'I do not know':50,

'NO':0,

'No':0,

'NO (UNCHECKED)':0,

'STILL HOSP':50}

resA=[]

for i in range(3,199):

qs = list(df.iloc[:, i].unique())

respons={}

resA\_part={}

for j in range(0,len(qs)):

key\_res=str(qs[j])+str((i-2))

if qs[j] in respons\_yn.keys():

respons[qs[j]]=respons\_yn[qs[j]]\*i

resA\_part[key\_res]=respons\_yn[qs[j]]\*i

else:

respons[qs[j]]=(j\*10)\*i

resA\_part[key\_res]=(j\*10)\*i

df.iloc[:, i] = [respons[item] for item in df.iloc[:, i]]

resA.append(resA\_part)

states = df['LocationDesc'].unique()

df\_state=df[df['LocationDesc']=='Colorado']

df\_state = df\_state.iloc [:, 3:199]

scaler = preprocessing.StandardScaler()

df\_state\_claster = scaler.fit\_transform(df\_state)

#Блок подбора наилучшего количества кластеров

k = [2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20]

silhouette\_scores = []

for i in range(0,len(k)):

kmeans = KMeans(n\_clusters=k[i], random\_state=0, n\_init=10)#.fit(df\_state\_check)

# Appending the silhouette scores of the different models to the list

Продолжение ПРИЛОЖЕНИЯ Д

silhouette\_scores.append(

silhouette\_score(df\_state, kmeans.fit\_predict(df\_state\_claster)))

plt.bar(k, silhouette\_scores)

plt.xlabel('Количество кластеров', fontsize = 16)

plt.ylabel('S(i)', fontsize = 16)

plt.show()

#Получение списка вопросов

q=list(df.columns.values)

questions=[]

for i in range(3,199):

questions.append(q[i])

#Разбиение на кластеры

number\_cl=2

cl = KMeans(n\_clusters=number\_cl, random\_state=0, n\_init=10)

predictions = cl.fit\_predict(df\_state)

#Для подсчё того того, сколько респондентов попало в какой кластер

c = Counter(predictions)

print(c)

#Добавление меток кластера в датафрейм

df\_state['Cluster'] = predictions. tolist ()

df\_state\_train = df\_state.sample (frac= 0.8 ,random\_state= 0 )

df\_state\_test = df\_state.drop (df\_state\_train. index )

#Построение дерева классификации

y = df\_state\_train.iloc[:,196].to\_numpy()

x = df\_state\_train.iloc[:,:196].to\_numpy()

clf = tree.DecisionTreeClassifier()

clf = clf.fit(x, y)

#Отрисовка дерева классификации

dot\_data = tree.export\_graphviz(clf, out\_file=None,

feature\_names=questions,

class\_names=["1","2"],

filled=True, rounded=True,

special\_characters=True)

graph = graphviz.Source(dot\_data)

fie\_ext = 'png'

temp\_img = 'state\_Colorado\_tree'

graph.render(temp\_img,format=fie\_ext, view=False)

Xp=df\_state\_test.iloc[:,:196].to\_numpy()

Yp=df\_state\_test.iloc[:,196].to\_numpy()

#Расчёт точности классификации

Ytree=clf.predict(Xp)

accuracy\_score(Yp,Ytree)

#Выделение датафрема, содержащего 1-ую группу штата

df\_state\_cl=df\_state[df\_state['Cluster']==0]

#Замена числовых значений на текстовые ответы с указанием № вопроса

for i in range(0,196):

rs = {

value: key for key, value in resA[i].items()}

df\_state\_cl.iloc[:, i] = [rs[item] for item in df\_state\_cl.iloc[:, i]]

#Поиск ассоциатиных правил

Продолжение ПРИЛОЖЕНИЯ Д

a = df\_state\_cl.iloc[:,:196].to\_numpy()

rules = apriori(a, min\_support = 0.1, min\_confidence = 0.1, min\_lift = 0, max\_length = 2, target = "rules")

association\_results = list(rules)

#Формирование датафрейма ассоциативных правил

rules=[]

support=[]

confidence=[]

lift=[]

for item in association\_results:

pair = item[0]

items = [x for x in pair]

if len(items) >= 2:

rules.append(items[0] + " -> " + items[1])

support.append(item[1])

confidence.append(item[2][0][2])

lift.append(item[2][0][3])

data\_rules = []

data\_rules .append(rules)

data\_rules .append(support)

data\_rules .append(confidence)

data\_rules .append(lift)

state\_data = pd.DataFrame(data\_rules). transpose ()

state\_data.columns =['Rule', 'Support', 'Confidence', 'Lift']

#Сортировка ассоциативных правил по поддержке

df\_rules=state\_data.sort\_values(by=['Support'], ascending=False)

display (df\_rules.head(n=15))

#Получение правил со следствием отсутствие депрессии

df\_rules\_copy = df\_rules.copy().reset\_index(drop=True)

keep= ["-> NO167"]

df\_rules167 = df\_rules\_copy[df\_rules\_copy.Rule.str.contains('| '.join(keep))]

display (df\_rules167.head(n=15))

|  |
| --- |
| МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ |
| ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ |
| **«МОСКОВСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»** |
| **(МОСКОВСКИЙ ПОЛИТЕХ)** |
| Факультет информационных технологий |

Кафедра «Прикладная информатика»

Форма обучения: очная

**ИЛЛЮСТРАТИВНЫЙ МАТЕРИАЛ**

**К ВЫПУСКНОЙ КВАЛИФИКАЦИОННОЙ РАБОТЕ**

по направлению 01.03.02 «Прикладная математика и информатика»

на тему «Прогнозирование материнского поведения в регионе по данным опросов на основе технологий больших данных»

Студентка Олеся Михайловна Репина

Руководитель работы

доцент, к.э.н. Александр Евгеньевич Рабинович

ДОПУСКАЕТСЯ К ЗАЩИТЕ

Заведующий кафедрой

профессор, к.э.н. Станислав Вадимович Суворов

МОСКВА 2023

