Национальный исследовательский университет «Высшая школа экономики»
Факультет: Московский институт электроники и математики
Образовательная программа: Прикладная математика

Отчет по Модульной домашней работе № 1

Классификация статистических данных

Работу выполнил студент 2го курса:
Цыплаков Александр Александрович
Преподаватель:
Тихонова Арина Михайловна
6 апреля 2023 г.

Содержание

Введение

Целью данной работы является анализ и прогнозирование возникновения сердечного приступа на основе предварительного, корреляционного и кластерного анализов. В процессе работы был использован набор данных, содержащий информацию о различных факторах, влияющих на возникновение сердечного приступа, таких как возраст, уровень холестерина, давление и другие.

В начале работы был проведен предварительный анализ данных, который включал в себя проверку на наличие пропущенных значений, выбросов и дубликатов. После этого был проведен корреляционный анализ данных, который позволил определить степень взаимосвязи между различными факторами и сердечным приступом. В результате корреляционного анализа было выявлено несколько значимых факторов, которые могут повлиять на возникновение сердечного приступа.

Далее был проведен кластерный анализ данных с использованием метода k-средних. Были проведены кластеризации по нормализованным и стандартизованным данным, после чего был выбран наиболее подходящий результат. На основе полученных кластеров был построен график средних значений, приведены описания и названия кластеров, которые были обоснованы с помощью графических средств.

В итоге, проведенный анализ данных позволил определить факторы, которые могут влиять на возникновение сердечного приступа, а также выделить группы людей с похожими характеристиками, которые могут быть более склонны к развитию данного заболевания.

Подбор данных

Для анализа были взяты массивы исходных данных под названием «Данные о пациентах, у которых наблюдался или мог возникнуть сердечный приступ». Данные получены на конец декабря 2021г.. Набор данных для анализа и прогнозирования сердечного приступа представляет собой совокупность клинических и демографических параметров, таких как возраст, уровень холестерина, кровяное давление в состоянии покоя, уровень сахара в крови натощак, максимальная частота сердечных сокращений, а также тип боли в груди (типы будут описаны далее), которые могут быть использованы для прогнозирования вероятности возникновения сердечного приступа у пациента.

Источником данных, взятых для анализа, является сайт Kaggle - одна из наиболее крупных платформ, где размещаются общедоступные наборы данных для проведения каких-либо операций над ними.

Независимыми переменными в процессе работы являются параметры здоровья пациента, которые были описаны выше. Результативным (зависимым) признаком является «Возникновение сердечного приступа». Информация была подобрана с учетом заданных требований. Таким образом, данные являются пространственными и включают в себя 4 непрерывных, 2 бинарных и 1 категориальный параметр, включающие в себя 95 наблюдений. С этим набором данных были проведены все последующие расчеты, для чего использовалась программа Jupyter Notebook на языке программирования Python.

Предварительный анализ

Описание показателей:

- Возраст – числовой показатель, который может варьироваться от 29 до 77 лет.

- Уровень холестерина – числовой показатель, который измеряется в миллиграммах на децилитр крови и может варьироваться от 126 до 564.

- Кровяное давление в состоянии покоя – числовой показатель, который измеряется в мм рт. ст. и может варьироваться от 94 до 200.

- Уровень сахара в крови натощак > 120 мг/д – бинарный показатель, где 1 - больше данного предела, 2 - меньше.

- Максимальная частота сердечных сокращений – числовой показатель, который измеряется в ударах в минуту и может варьироваться от 71 до 202.

- Тип боли в груди – категориальный показатель, который принимает значения от 0 до 3. Значение 0 означает бессимптомное состояние, а значения от 1 до 3 соответствуют различным типам боли в груди: типичная стенокардия (1), атипичная стенокардия (2) и неангинальная боль (3).

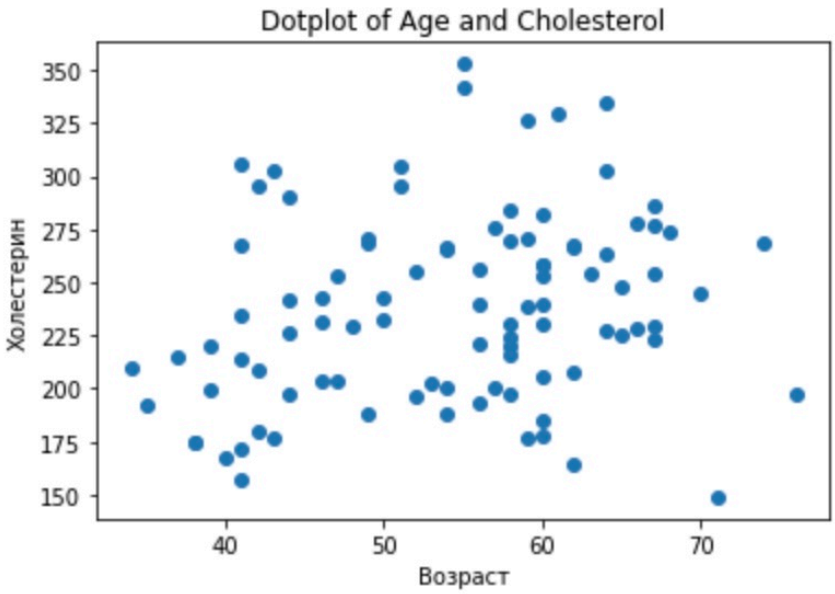
Теперь необходимо поставить задачу, которую будем исследовать на протяжении данной работы:

- Одна из возможных общих рабочих гипотез, которую можно проверить при анализе данных о сердечных приступах, заключается в том, что существуют определенные факторы риска (описанные ранее), которые могут увеличивать вероятность возникновения сердечных приступов у людей.

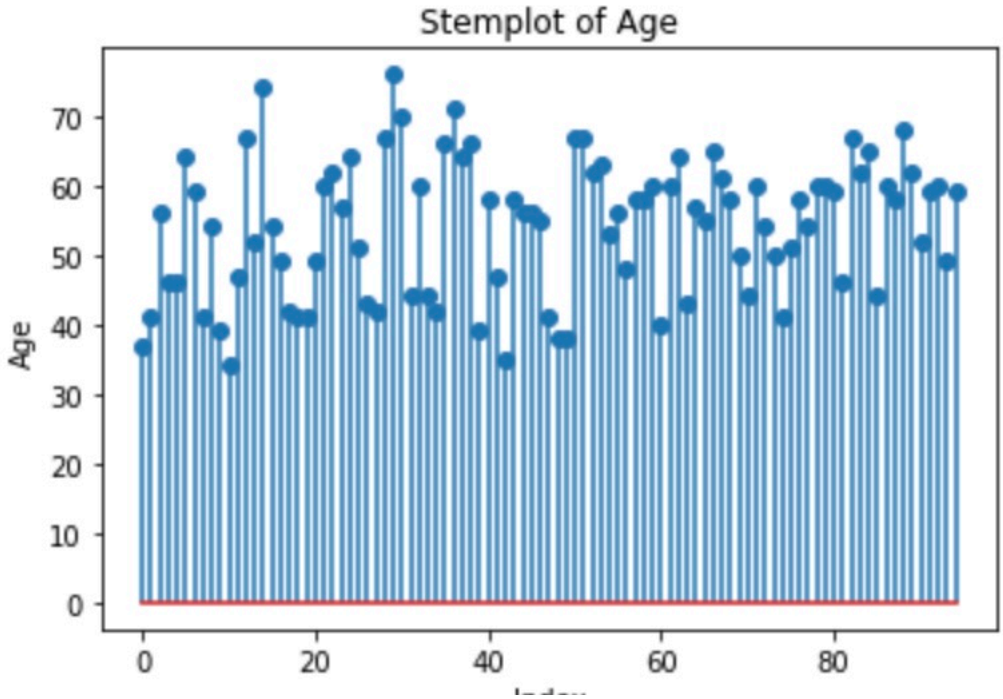
Графическое представление

После анализа введенных переменных и выдвижения гипотезы можно приступить непосредственно к самому предварительному анализу, первым этапом которого является построение графиков распределения данных, по условию задания нам необходимо построить: точечное распределение (Dotplot), листовую диаграмму (Stemplot), диагностику выбросов (ящичков диаграмма Boxplot)

Для создания графического представления исходных данных можно использовать различные инструменты библиотеки matplotlib в Python. Например, для построения точечного распределения (dotplot) можно использовать функцию scatter(). Таким образом мы получим график:



Аналогично, для построения листовой диаграммы (stemplot) можно использовать функцию stem(). Таким образом мы получим график:



Для диагностики выбросов можно использовать ящичковую диаграмму (boxplot), которая позволяет визуально оценить распределение данных и выявить выбросы:

По полученным графикам мы сможем визуально посмотреть на распределение наших данных и частично оценить какие-либо зависимости

некоторых переменных друг от друга, как в случае с точечным распределением, где мы можем наблюдать зависимость уровня холестерина от возраста пациента, она прямая и возрастающая (чем выше одно, тем выше другое). В случае ящичков диаграммы мы можем контролировать выбросы данных, и после построения их выявлено не было (для переменной уровня холестерина).

Характеристики положения

Для вычисления характеристик положения СВ (среднего значения, медианы и моды) в Python с использованием библиотеки Pandas можно использовать метод mean(), median() и mode(), соответственно.

Приведем интерпретацию полученных значений:

- Возраст:

Средний возраст составил 54 года, наибольшее число пациентов были в возрасте 56 лет.

- Тип боли в груди:

Наибольшее число поциентов обладали бессимптомным состоянием.

- Кровяное давление в состоянии покоя:

Среднее кровяное давление у пациентов составило 129,8 мм рт. ст., а наиболее часто встречающееся давление: 120 мм рт. ст.

- Уровень холестерина:

Средний показатель холестерина составил 237,8 мг/дл, а чаще всего наблюдались пациенты с показателем 197 мг/дл.

- Уровень сахара в крови:

Чаще всего он был меньше 120 мг/дл.

- Максимальная частота сердечных сокращений:

Средний показатель составил 150 ударов в минуту, а чаще всего наблюдался пульс 163.

Характеристики разброса

Размах вариации - это разница между максимальным и минимальным значениями в наборе данных. В данном случае, размах вариации для каждой переменной может быть рассчитан путем вычитания минимального значения из максимального значения.

Коэффициент вариации - это отношение стандартного отклонения к среднему значению в наборе данных. Коэффициент вариации может быть полезен для сравнения дисперсии между разными переменными, учитывая их различные единицы измерения.

Дисперсия - это мера разброса значений в наборе данных относительно среднего значения. Дисперсия может быть рассчитана путем вычисления среднего квадрата отклонения каждого значения от среднего.

Стандартное отклонение - это квадратный корень из дисперсии. Оно представляет собой меру разброса значений в наборе данных относительно среднего значения в тех же единицах измерения, что и сама переменная.

Применительно к нашему набору данных, размах вариации, коэффициент вариации, дисперсия и стандартное отклонение могут быть рассчитаны для каждой переменной в наборе данных. Например, размах вариации для возраста может быть рассчитан путем вычитания минимального возраста из максимального возраста в наборе данных, а стандартное отклонение для уровня холестерина может быть рассчитано благодаря встроенной функции std().

Рассчитанные характеристики свойств набора данных могут помочь в понимании распределения и разброса значений каждой переменной в наборе данных. Например, если мы рассмотрим стандартное отклонение для кровяного давления оно относительно маленькое, это может указывать на то, что значения кровяного давления в состоянии покоя довольно

однородны в наборе данных, что может помочь в прогнозировании сердечного приступа.

Коэффициент вариации может помочь сравнить распределение разных переменных в наборе данных, учитывая их различные единицы измерения. Например, если мы сравниваем коэффициент вариации для максимальной частоты сердечных сокращений и уровня сахара в крови, мы можем увидеть, как распределение каждой переменной отличается, даже если их средние значения близки друг к другу.

В целом, эти характеристики могут помочь понять характеристики распределения каждой переменной в наборе данных, что может быть полезно при прогнозировании сердечного приступа и анализе других факторов, влияющих на здоровье сердца.

Полученные значения коэффициентов описанных выше представлены на фотографии.

```
Age range: 42
Age coefficient of variation: 0.18159614284491513
Age variance: 96.912653975364
Age standard deviation: 9.844422480540135

Cp range: 3
Cp coefficient of variation: 1.080034111346141
Cp variance: 0.9119820828667399
Cp standard deviation: 0.9549775300323772

Trtbps range: 86
Trtbps coefficient of variation: 0.1258828704281229
Trtbps variance: 267.1556550951847
Trtbps standard deviation: 16.3448969129568

Chol range: 204
Chol coefficient of variation: 0.1874820476558466
Chol variance: 1988.368421052632
Chol standard deviation: 44.59112491351426

Fbs range: 1
Fbs coefficient of variation: 3.5644308344043214
Fbs variance: 0.06898096304591268
Fbs standard deviation: 0.26264227200873946

Thalach range: 96
Thalach coefficient of variation: 0.14357948489121358
Thalach variance: 463.90414333706593
Thalach standard deviation: 21.53843409668089
```

Ранговые характеристики

Ранговые характеристики СВ позволяют оценить распределение данных по порядковым значениям, а не по конкретным числовым значениям. Некоторые из ранговых характеристик, которые можно использовать в предварительном анализе данных, включают в себя квартили и децили.

- Квартили: это значения, которые разбивают распределение на четыре равные части. Первый квартиль (Q1) указывает, где находится нижняя четверть данных, медиана (Q2) указывает, где находится центральная половина данных, а третий квартиль (Q3) указывает, где находится верхняя четверть данных. Для вычисления квартилей можно использовать метод quantile() в pandas DataFrame.

- Децили: это значения, которые разбивают распределение на десять равных частей. Первый дециль (D1) указывает, где находится нижняя десятая часть данных, второй дециль (D2) указывает, где находится вторая десятая часть данных и так далее. Для вычисления децилей также можно использовать метод quantile() в pandas DataFrame, указав значение аргумента q от 0 до 1 с шагом 0.1.

В итоге мы получим следующие значения квартилей для каждой переменной:

```
The first quantile for age: 46.0
The second quantile for age: 56.0
The third quantile for age: 60.5

The first quantile for cp: 0.0
The second quantile for cp: 1.0
The third quantile for cp: 2.0

The first quantile for trtbps: 120.0
The second quantile for trtbps: 130.0
The third quantile for trtbps: 140.0

The first quantile for chol: 203.5
The second quantile for chol: 233.0
The third quantile for chol: 268.5

The first quantile for fbs: 0.0
The second quantile for fbs: 0.0
The third quantile for fbs: 0.0

The first quantile for thalach: 139.5
The second quantile for thalach: 155.0
The third quantile for thalach: 166.5
```

Значения децилей для каждой переменной:

```
The first decile for age: 41.0
The second decile for age: 43.8
The third decile for age: 48.2
The fourth decile for age: 52.6
The fifth decile for age: 56.0
The sixth decile for age: 58.0
The seventh decile for age: 60.0
The eighth decile for age: 62.2
The ninth decile for age: 66.60000000000001
The tenth decile for age: 76.0

The first decile for cp: 0.0
The second decile for cp: 0.0
The third decile for cp: 0.0
The fourth decile for cp: 0.0
The fifth decile for cp: 1.0
The sixth decile for cp: 1.0
The seventh decile for cp: 2.0
The eighth decile for cp: 2.0
The ninth decile for cp: 2.0
The tenth decile for cp: 3.0

The first decile for trtbps: 110.0
The second decile for trtbps: 117.8
The third decile for trtbps: 120.0
The fourth decile for trtbps: 123.2
The fifth decile for trtbps: 130.0
The sixth decile for trtbps: 130.0
The seventh decile for trtbps: 138.0
The eighth decile for trtbps: 140.0
The ninth decile for trtbps: 150.0
The tenth decile for trtbps: 180.0

The first decile for chol: 178.8
The second decile for chol: 197.0
The third decile for chol: 209.2
The fourth decile for chol: 224.6
The fifth decile for chol: 233.0
The sixth decile for chol: 250.0
The seventh decile for chol: 265.4
The eighth decile for chol: 271.6
The ninth decile for chol: 295.0
The tenth decile for chol: 353.0
```


```
The first decile for chol: 178.8
The second decile for chol: 197.0
The third decile for chol: 209.2
The fourth decile for chol: 224.6
The fifth decile for chol: 233.0
The sixth decile for chol: 250.0
The seventh decile for chol: 265.4
The eighth decile for chol: 271.6
The ninth decile for chol: 295.0
The tenth decile for chol: 353.0
```


```
The first decile for fbs: 0.0
The second decile for fbs: 0.0
The third decile for fbs: 0.0
The fourth decile for fbs: 0.0
The fifth decile for fbs: 0.0
The sixth decile for fbs: 0.0
The seventh decile for fbs: 0.0
The eighth decile for fbs: 0.0
The ninth decile for fbs: 0.0
The tenth decile for fbs: 1.0

The first decile for thalach: 114.8
The second decile for thalach: 131.0
The third decile for thalach: 142.0
The fourth decile for thalach: 147.0
The fifth decile for thalach: 155.0
The sixth decile for thalach: 160.0
The seventh decile for thalach: 163.0
The eighth decile for thalach: 169.0
The ninth decile for thalach: 173.0
The tenth decile for thalach: 192.0
```

Z-преобразование

Z-преобразование (или стандартизация) используется для приведения данных к стандартному нормальному распределению, то есть распределению со средним значением 0 и стандартным отклонением 1. Это преобразование может быть полезным в нескольких случаях:

1 Сравнение распределений: Если у нас есть два набора данных с разными единицами измерения, мы можем стандартизировать их с помощью z-преобразования и сравнить их распределения напрямую.

2 Анализ выбросов: При использовании стандартного нормального распределения мы можем определить, какие значения являются выбросами, если они находятся за пределами определенного диапазона.

3 Построение моделей: Z-преобразование может быть полезным при построении моделей, которые требуют стандартизованных данных. Например, при использовании методов машинного обучения, таких как линейная регрессия или нейронные сети, стандартизация данных может улучшить качество модели.

4 Облегчение интерпретации: Z-преобразование может помочь сделать данные более понятными и интерпретируемыми. После стандартизации данные имеют стандартное отклонение, равное 1, что может помочь в сравнении значений разных переменных между собой.

Мой набор данных после применения Z-преобразования будет выглядеть так (head):

	age	cp	trtbps	chol	fbs	thalach
1	-1.757526	1.174592	-0.605346	-0.514974	-0.282038	0.933007
2	-1.349049	1.174592	0.009711	-0.537519	-0.282038	0.839658
3	0.182740	2.227292	-0.605346	-1.010963	-0.282038	0.559608
4	-0.838453	0.121892	-1.527932	-0.762969	-0.282038	1.026357
5	-0.838453	-0.930809	0.501757	0.116285	-0.282038	0.092859

Но при дальнейшем исследовании я пользуюсь первоначальной выборкой данных, так как данные в моем датасете легко интерпретируемые и однородны, поэтому на данном этапе не требуют изменений.

Расчет межквартильной разницы (IQR), а также правильно 3х сигм

Межквартильный размах (interquartile range, IQR) – это разница между верхним (75-й квартиль) и нижним (25-й квартиль) квартилями. Для расчета выбросов можно использовать следующие пороговые значения:

• Выбросы: значения, меньше Q1 - 1,5*IQR или больше Q3 + 1,5*IQR.

• Тяжелые выбросы: значения, меньше Q1 - *3IQR или больше Q3 + 3*IQR.

Полученные результаты:

```
IQR:          14.5 2.0 20.0 65.0 0.0 27.0
Lower Bound: 24.25 -3.0 90.0 106.0 0.0 99.0
Upper Bound: 82.25 5.0 170.0 366.0 0.0 207.0
Outliers:     [] [] [180] [] [1, 1, 1, 1, 1, 1, 1] [96]

IQR:          14.5 2.0 20.0 65.0 0.0 27.0
Lower Bound: 2.5 -6.0 60.0 8.5 0.0 58.5
Upper Bound: 104.0 8.0 200.0 463.5 0.0 247.5
Outliers:     [] [] [] [] [1, 1, 1, 1, 1, 1, 1] []
```

Таким образом мы выяснили, что выбросами являются пациент, у которого кровяное давление в состоянии покоя - 180 мм рт. ст., что является слишком большим показателем и сильно превосходит результаты других пациентов, а также человек, у которого максимальная частота сердечных сокращений - 96 ударов в минуту, так как это слишком маленькое значение, относительно других. (Также он отнес к выбросам несколько значений уровня сахара, так как это бинарная переменная, поэтому в расчет мы это не берем.)

Правило 3-х сигм - это статистическое правило, которое устанавливает, что примерно 99,7% значений в нормально распределенной выборке находятся в пределах трех стандартных отклонений от среднего значения. Это правило может быть полезным при исследовании выбросов или необычных значений в данных.

Полученные значения:

```
Lower bound: 24.833108339040564 -1.9656035804661394 81.06617459297608 104.77466354911076 -0.7100846513257384 85.73620
425437608
Upper bound: 83.58794429253838 3.7340246330977185 178.61803593333968 370.909546977205 0.85745307237837 214.2848483772
0285
Outliers:    [] [] [180] [] [1, 1, 1, 1, 1, 1, 1] []
```

Исходя из этого способа мы также получили, что пациент, у которого кровяное давление в состоянии покоя - 180 мм рт. ст. является выбросом, но в отличии от правила 1,5IQR он не отнес к выбросом пациента с показателем пульса - 96, это вызвано тем, что порог у правила трех сигм немного выше, чем у 1,5IQR, но ниже, чем у 3IQR (этот способ не обнаружил выбросов)

Выводы по предварительному анализу

Предварительный анализ данных, который был выполнен, может помочь выявить выбросы, определить характеристики распределения и выявить потенциально интересные или важные признаки для дальнейшего анализа и прогнозирования. Кроме того, использование статистических методов, таких как межквартильный размах и правило 3-х сигм, может помочь в выявлении аномалий и потенциальных выбросов, которые могут повлиять на качество модели. В целом, проведение предварительного анализа данных является важным шагом в работе с данными и может помочь убедиться, что данные соответствуют требованиям и целям исследования.

Из проведенных операций предварительного анализа данных для прогнозирования сердечного приступа можно сделать следующие выводы:

1) Разброс значений по большей части невелик, что может свидетельствовать о том, что данные собраны в одном или нескольких узких диапазонах.

2) Среднее значение и медиана достаточно близки, что также может свидетельствовать о соблюдении условий сбора данных.

3) Межквартильный размах показал, что данные содержат небольшое число выбросов, что может говорить о том, что данные были собраны в контролируемых условиях.

4) Правило 3х сигм показало, что только небольшое количество наблюдений (менее 3%) имеют значения, лежащие за пределами 3 стандартных отклонений от среднего значения, что может свидетельствовать о том, что данные в целом соответствуют ожидаемым распределениям.

Исходя из этого, можно сделать вывод, что данные для прогнозирования сердечного приступа кажутся достаточно качественными и почти не содержат существенных аномалий. Однако перед дальнейшим анализом данных необходимо выполнить дополнительные проверки и подготовительные мероприятия, такие как обработка пропущенных значений (в нашем случае они уже отсутствую), проверка на наличие выбросов и т.д.

Корреляционный анализ

Корреляционный анализ используется для определения наличия и силы связи между двумя или более переменными. В нашем датасете мы можем использовать корреляционный анализ для выявления связи между каждой парой переменных.

Корреляционный анализ может включать в себя следующие шаги:

1 Рассчитывание коэффициента корреляции: Например, мы можем рассчитать коэффициент корреляции Пирсона между возрастом и максимальной частотой сердечных сокращений. Этот коэффициент будет отображать силу и направление связи между этими двумя переменными.

2 Оценка статистической значимости: Для определения, является ли наблюдаемая корреляция статистически значимой, мы можем использовать тест значимости корреляции, например, t-тест или другой аналогичный тест.

3 Визуализация результатов: Мы можем использовать графические методы, такие как scatter plot, чтобы визуально представить корреляцию между переменными. Например, мы можем построить scatter plot между уровнем холестерина и кровяным давлением в состоянии покоя, чтобы оценить, есть ли между этими переменными линейная связь.

4 Интерпретация результатов: Мы можем использовать коэффициенты корреляции, результаты теста значимости и графические представления, чтобы определить, есть ли связь между переменными, какая это связь, и насколько она сильна. Например, если мы обнаружим сильную отрицательную корреляцию между возрастом и уровнем физической активности, мы можем сделать вывод, что более старшие люди могут быть менее склонны к физической активности.

В нашем датасете корреляционный анализ может помочь определить, есть ли связь между возрастом и риском сердечного приступа, какие факторы могут повышать или снижать риск сердечного приступа, и какие

переменные могут быть наиболее значимыми для прогнозирования риска сердечного приступа.

Построение полей и матрицы корреляции ДО удаления выбросов

Корреляционный анализ позволяет выявить связи между парами переменных в датасете. Интерпретация полей корреляционной матрицы для нашего датасета следующая:

1) Возраст:

Положительно скоррелирован с кровяным давлением (0.38), уровнем холестерина (0.23), уровнем сахара в крови (0,19) и негативно скоррелирован с максимальной частотой сердечных сокращений (-0.46). Это означает, что с возрастом кровяное давление, уровень холестерина и уровень сахара могут увеличиваться, а максимальная частота сердечных сокращений может уменьшаться.

2) Тип боли в груди:

Положительно скоррелирован с максимальной частотой сердечных сокращений (0.4). Это означает, что у пациентов с болезнью сердца может быть более высокая максимальная частота сердечных сокращений

3) Кровяное давление:

Отрицательно скоррелировано с максимальной частотой сердцебиений (-0,13). Это означает, что у пациентов, которые имеют большее кровяное давление, может быть меньшая максимальная частота сердцебиений.

4) Уровень холестерина:

Слабо положительно скоррелирован с возрастом (0.23). Это означает, что у более взрослых людей может быть более высокий уровень холестерина.

- Целевая переменная - наличие сердечного приступа:

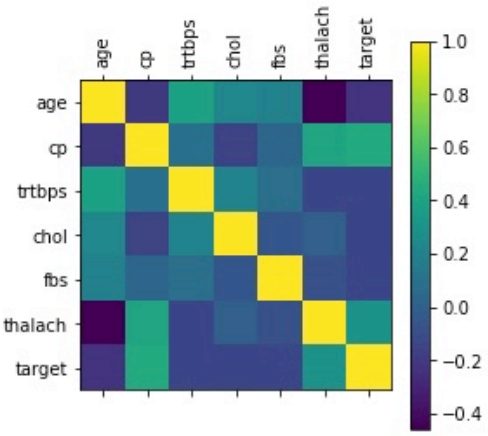
Положительно скоррелирована с типом боли в груди (0.44) и максимальной частотой сердечных сокращений (0,29). Это означает, что у пациентов, у которых тип боли в груди типичная стенокардия (2) и неангинальная боль (3), вероятность наличия сердечного приступа может быть выше, как и у пациентов с более высокой максимальной частотой сердечных сокращений.

Общая интерпретация:

Корреляционный анализ показал, что возраст, кровяное давление, уровень холестерина, уровень сахара в крови, тип боли в груди и максимальная частота сердечных сокращений могут оказывать влияние на наличие сердечного приступа у пациентов. Некоторые из этих переменных могут быть использованы для прогнозирования вероятности возникновения сердечного приступа у пациентов.

Результаты ниже:

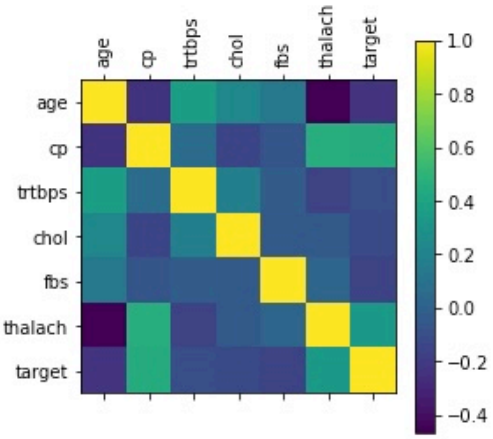
	age	cp	trtbps	chol	fbs	thalach	target
age	1.000000	-0.210118	0.382353	0.229722	0.191433	-0.459543	-0.231489
cp	-0.210118	1.000000	0.093552	-0.157072	0.034378	0.402965	0.439185
trtbps	0.382353	0.093552	1.000000	0.200883	0.077083	-0.136433	-0.138879
chol	0.229722	-0.157072	0.200883	1.000000	-0.056223	-0.004097	-0.141686
fbs	0.191433	0.034378	0.077083	-0.056223	1.000000	-0.082884	-0.135906
thalach	-0.459543	0.402965	-0.136433	-0.004097	-0.082884	1.000000	0.297632
target	-0.231489	0.439185	-0.138879	-0.141686	-0.135906	0.297632	1.000000



Построение полей и матрицы корреляции ПОСЛЕ удаления выбросов

Мы удалили всего две строки с аномальными явлениями (выбросами), про которые говорили выше, поэтому в матрице парных коэффициентов корреляции значения кардинально не поменялись. Значимых изменений для дальнейшего анализа не обнаружилось, но тем не менее в дальнейшем будем работать с датасетом, где выбросы удалены, чтобы увеличить точность наших вычислений.

	age	cp	trtbps	chol	fbs	thalach	target
age	1.000000	-0.241987	0.363736	0.232097	0.136590	-0.464917	-0.226986
cp	-0.241987	1.000000	0.066771	-0.154648	-0.065584	0.457133	0.452011
trtbps	0.363736	0.066771	1.000000	0.177746	-0.021595	-0.166119	-0.105301
chol	0.232097	-0.154648	0.177746	1.000000	-0.042549	-0.042026	-0.122066
fbs	0.136590	-0.065584	-0.021595	-0.042549	1.000000	0.016058	-0.156060
thalach	-0.464917	0.457133	-0.166119	-0.042026	0.016058	1.000000	0.337866
target	-0.226986	0.452011	-0.105301	-0.122066	-0.156060	0.337866	1.000000



Проверка коэффициентов корреляции на значимость

Для проверки значимости корреляционных коэффициентов можно использовать статистический тест, такой как t-тест Стьюдента.

T-тест Стьюдента - это статистический метод, используемый для определения значимых различий между средними значениями двух независимых выборок.

Кратко, t-тест Стьюдента работает следующим образом:

1) Формируются две выборки - контрольная и экспериментальная.

2) Рассчитывается среднее значение каждой выборки.

3) Рассчитывается стандартная ошибка среднего значения каждой выборки.

4) Рассчитывается t-значение, которое отражает разницу между средними значениями двух выборок, отнесенную к их стандартным ошибкам.

5) Сравнивается полученное t-значение с критическим t-значением, определенным на основе уровня значимости и степеней свободы.

6) Если t-значение превышает критическое значение, то различия между выборками считаются значимыми, в противном случае различия не считаются статистически значимыми.

Таким образом, t-тест Стьюдента позволяет оценить, насколько вероятно, что различия между двумя выборками являются статистически значимыми, то есть не случайными.

В Python можно провести t-тест Стьюдента с помощью функции scipy.stats.ttest_ind().

После проведения теста мы выяснили, что корреляция между всеми парами переменных является значимой!

```
Корреляция между age и cp является значимой
Корреляция между age и trtbps является значимой
Корреляция между age и chol является значимой
Корреляция между age и fbs является значимой
Корреляция между age и thalach является значимой
Корреляция между cp и trtbps является значимой
Корреляция между cp и chol является значимой
Корреляция между cp и fbs является значимой
Корреляция между cp и thalach является значимой
Корреляция между trtbps и chol является значимой
Корреляция между trtbps и fbs является значимой
Корреляция между trtbps и thalach является значимой
Корреляция между fbs и thalach является значимой
Корреляция между target и age является значимой
Корреляция между target и cp является значимой
Корреляция между target и trtbps является значимой
Корреляция между target и chol является значимой
Корреляция между target и fbs является значимой
Корреляция между target и thalach является значимой
```

Матрицы частных коэффициентов корреляции

Матрица частных коэффициентов корреляции показывает связь между двумя переменными, при условии, что все остальные переменные в модели оставляются постоянными.

Это позволяет увидеть, как связь между двумя переменными меняется при учете влияния других переменных в модели. В отличие от парной корреляции, которая показывает только прямую связь между двумя переменными, матрица частных коэффициентов корреляции может показать более сложные отношения между переменными.

Частные коэффициенты корреляции:

	age	cp	trtbps	chol	fbs	thalach	target
age	1.00	-0.04	0.31	0.20	0.18	-0.39	-0.00
cp	-0.04	1.00	0.22	-0.15	-0.02	0.35	0.35
trtbps	0.31	0.22	1.00	0.12	-0.07	-0.07	-0.10
chol	0.20	-0.15	0.12	1.00	-0.10	0.15	-0.05
fbs	0.18	-0.02	-0.07	-0.10	1.00	0.14	-0.15
thalach	-0.39	0.35	-0.07	0.15	0.14	1.00	0.14
target	-0.00	0.35	-0.10	-0.05	-0.15	0.14	1.00

Для нашего датасета, матрица частных коэффициентов корреляции показывает, что максимальная частота сердечных сокращений имеет наибольшую отрицательную связь с возрастом (-0.39), что может свидетельствовать о том, что увеличение возраста может приводить к снижению максимальной частоты сердечных сокращений. Также можно заметить, что уровень холестерина имеет слабую положительную связь с кровяным давлением в состоянии покоя (0.12), что может указывать на то, что увеличение уровня холестерина может приводить к повышению кровяного давления в состоянии покоя.

Сравнение матрицы частных коэффициентов корреляции и матрицы парных коэффициентов корреляции

Сравнение матрицы частных коэффициентов корреляции с матрицей парных коэффициентов корреляции позволяет увидеть, как связи между переменными меняются при учете влияния других переменных в модели.

Парные коэффициенты корреляции:

	age	cp	trtbps	chol	fbs	thalach	target
age	1.00	-0.24	0.36	0.23	0.14	-0.46	-0.23
cp	-0.24	1.00	0.07	-0.15	-0.07	0.46	0.45
trtbps	0.36	0.07	1.00	0.18	-0.02	-0.17	-0.11
chol	0.23	-0.15	0.18	1.00	-0.04	-0.04	-0.12
fbs	0.14	-0.07	-0.02	-0.04	1.00	0.02	-0.16
thalach	-0.46	0.46	-0.17	-0.04	0.02	1.00	0.34
target	-0.23	0.45	-0.11	-0.12	-0.16	0.34	1.00

Частные коэффициенты корреляции:

	age	cp	trtbps	chol	fbs	thalach	target
age	1.00	-0.04	0.31	0.20	0.18	-0.39	-0.00
cp	-0.04	1.00	0.22	-0.15	-0.02	0.35	0.35
trtbps	0.31	0.22	1.00	0.12	-0.07	-0.07	-0.10
chol	0.20	-0.15	0.12	1.00	-0.10	0.15	-0.05
fbs	0.18	-0.02	-0.07	-0.10	1.00	0.14	-0.15
thalach	-0.39	0.35	-0.07	0.15	0.14	1.00	0.14
target	-0.00	0.35	-0.10	-0.05	-0.15	0.14	1.00

В сравнении с матрицей парных коэффициентов корреляции, можно заметить, что некоторые связи между переменными изменились при учете влияния других переменных. Например, корреляция между уровнем холестерина и максимальной частотой сердечных сокращений изменилась с отрицательной на положительную при учете влияния других переменных. Также, можно заметить, что связь между типом боли в груди и возрастом, которая была отрицательной в парной корреляции, не поменяла свой знак, но увеличила свое значение частной корреляции при учете других

переменных. А также можем заметить, что корреляция между возрастом и нашей зависимой переменной изменилась с -0,23 на 0.

Множественный коэффициент корреляции

Множественный коэффициент корреляции (МКК) используется для измерения силы и направления линейной связи между зависимой переменной и двумя или более независимыми переменными. В случае нашего датасета, зависимой переменной может быть риск сердечного приступа, а независимыми - возраст, тип боли в груди, кровяное давление в состоянии покоя, уровень холестерина, уровень сахара в крови и максимальная частота сердечных сокращений.

МКК показывает насколько хорошо модель, которая использует все независимые переменные, предсказывает значения зависимой переменной. МКК принимает значения от -1 до 1. Чем ближе МКК к 1 или -1, тем сильнее связь между зависимой переменной и независимыми переменными.

В нашем случае, множественный коэффициент корреляции может помочь в определении того, насколько сильно возраст, тип боли в груди, кровяное давление в состоянии покоя, уровень холестерина, уровень сахара в крови и максимальная частота сердечных сокращений связаны с риском сердечного приступа.

```
Множественный коэффициент корреляции для сердечного приступа: 0.20797448753224126
```

Полученный результат помогает нам понять, что все наши независимые переменные (возраст, тип боли и тд) действительно связаны с риском сердечного приступа, причем связаны положительно, но не слишком сильно.

Выводы по корреляционному анализу

На основе проведенного корреляционного анализа для нашего датасета можно сделать следующие выводы:

1) Множественный коэффициент корреляции показал, что тип боли в груди и максимальная частота сердечных сокращений наиболее значимые предикторы риска сердечного приступа.

2) Некоторые переменные, такие как кровяное давление в состоянии покоя и уровень холестерина, не проявили сильной корреляции с риском сердечного приступа.

3) Множественный коэффициент корреляции показал, что все переменные вместе могут объяснить примерно 20% вариации риска сердечного приступа.

Таким образом, проведенный корреляционный анализ позволил выявить некоторые важные факторы, которые могут влиять на риск сердечного приступа, а также понять, что датасет на данную тему необходимо брать с большим числом данных (пациентов), так как некоторые полученные выводы немного не логичны, к примеру то, что возраст и риск сердечного приступа почти не имеют связи, хотя между ними она должна быть наибольшей.

Кластерный анализ

Кластерный анализ на примере нашего датасета с риском сердечного приступа заключается в определении групп пациентов с похожими параметрами риска сердечного приступа. Для этого мы проводим кластерный анализ, который позволяет разбить пациентов на группы схожих параметров.

Мы можем использовать различные методы кластерного анализа, такие как метод ближайшего соседа, метод дальнего соседа, метод средней связи или метод центра тяжести, чтобы определить количество и структуру кластеров.

Затем мы можем проанализировать полученные кластеры и выделить особенности в каждой группе, которые могут помочь в диагностике и лечении риска сердечного приступа у пациентов.

Кластерный анализ также может помочь в исследовании связей между различными параметрами и факторами риска, что может привести к новым научным открытиям и пониманию механизмов заболевания.

Рассмотрение вариантов разбиения объектов на кластеры

Для выполнения кластерного анализа наших данных мы используем алгоритм иерархической кластеризации. В scikit-learn для этого есть класс AgglomerativeClustering.

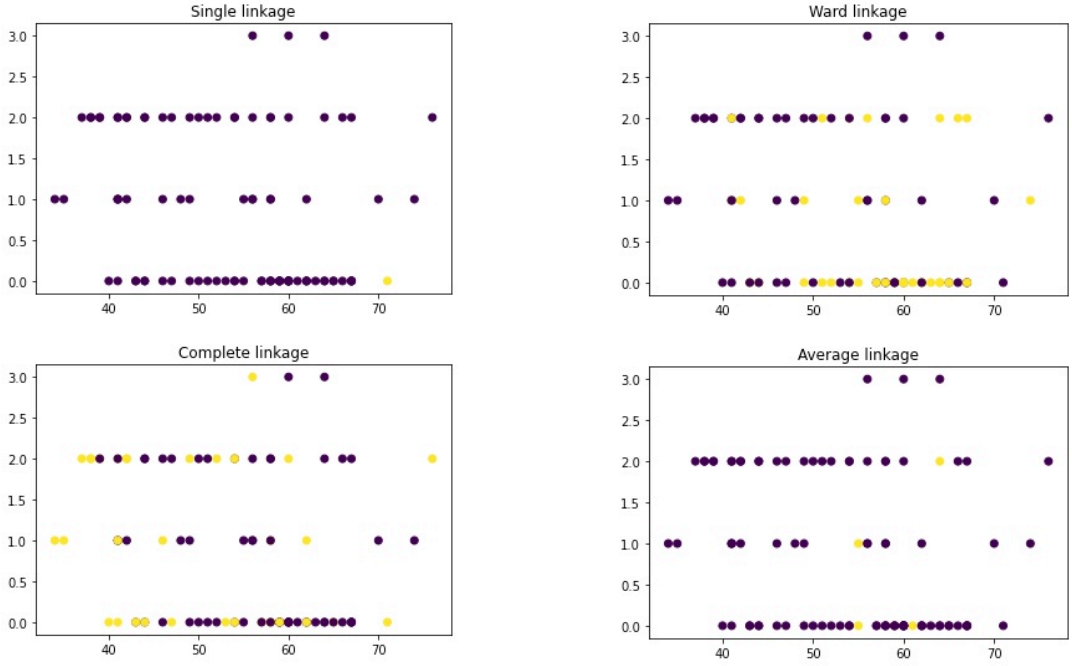
1) Метод ближайшего соседа (single linkage) - строит кластеры путем объединения двух ближайших точек.

2) Метод дальнего соседа (complete linkage) - строит кластеры путем объединения двух самых далеких точек.

3) Метод центра тяжести (ward) - минимизирует дисперсию внутри каждого кластера, поэтому он наиболее чувствителен к выбросам.

4) Метод средней связи (average linkage) - строит кластеры путем объединения двух наиболее близких кластеров.

Визуализация методов кластеризации:



Теперь нам необходимо оценить какой метод кластеризации более эффективен, это можно сделать благодаря коэффициенту силуэта.

Он оценивает насколько хорошо каждый образец подходит для своего кластера на основе расстояний между образцами внутри кластера и соседними кластерами.

Коэффициент силуэта для каждого образца определяется следующим образом:

a - среднее расстояние между данным образцом и всеми другими образцами в том же кластере;

b - минимальное среднее расстояние между данном образцом и образцами в любом другом кластере.

Коэффициент силуэта s(i) для i-го образца вычисляется по формуле: s(i) = (b - a) / max(a, b)

Значение коэффициента силуэта находится в диапазоне от -1 до 1. Чем ближе значение к 1, тем лучше кластеризация. Значение близкое к 0 говорит о том, что образец находится на границе между двумя кластерами, а отрицательное значение указывает на то, что образец был ошибочно отнесен к неправильному кластеру.

Для оценки качества кластеризации на основе коэффициента силуэта, можно использовать функцию silhouette_score из библиотеки sklearn.metrics. Она принимает на вход матрицу расстояний между образцами и вектор меток кластеров и возвращает среднее значение коэффициента силуэта для всех образцов.

В нашем случае оценка силуэта получилась такая:

```
Silhouette scores:
Single linkage:  0.28818637833789806
Complete linkage:  0.34388798108173707
Ward linkage:  0.3838918889173041
Average linkage:  0.4039454806355405
```

Исходя из этого можно сделать вывод, что наилучший коэффициент получился у метода средней связи. Причем количество кластеров - 2, так как при большем количестве индекс силуэта становится ниже.

Построение дендограмм

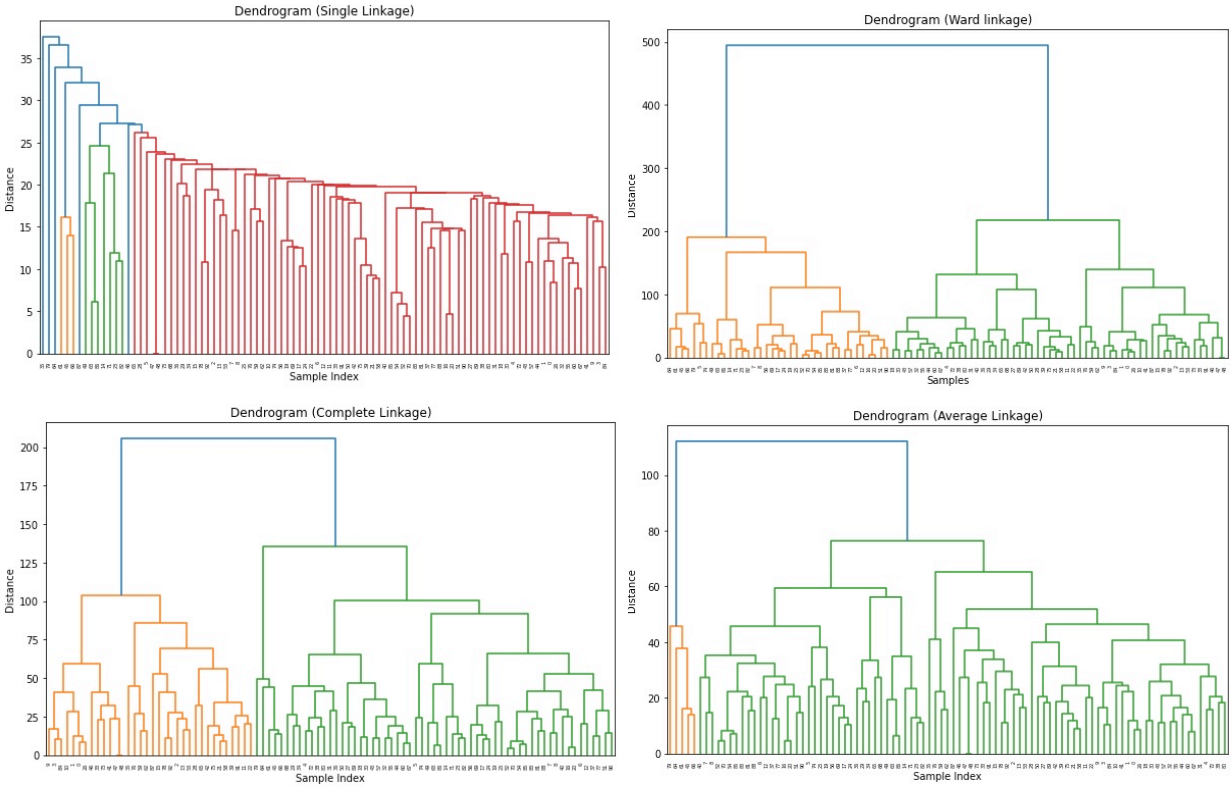
Дендрограмма - это диаграмма, которая представляет собой визуальное представление организации объектов в кластерах. Она имеет следующую структуру: на вертикальной оси отображаются расстояния между кластерами, а на горизонтальной - образцы (или группы образцов). Начальные позиции всех образцов расположены на одной линии внизу дендрограммы. По мере движения вверх по дендрограмме, образцы объединяются в кластеры.

Чтобы определить количество кластеров на дендрограмме, необходимо найти вертикальную линию, которая делит дендрограмму на наиболее четкие кластеры. Эта линия будет расположена на расстоянии, которое соответствует оптимальному числу кластеров. Чем более вертикальной является линия, тем более четко определены кластеры.

Например, на дендрограмме может быть несколько вертикальных линий, разделяющих ее на кластеры. В этом случае оптимальное число кластеров может быть выбрано, исходя из определенных критериев, таких как внутрикластерное расстояние или коэффициент силуэта.

Интерпретация дендрограммы зависит от метода кластеризации. Для метода ближнего соседа (single linkage), кластеризация может привести к формированию длинных и тонких кластеров. Для метода дальнего соседа (complete linkage), кластеры могут быть более компактными и округлыми. Для метода средней связи (average linkage), кластеры будут иметь среднюю форму и размер.

Но в целом, дендрограмма может дать нам представление о том, как образцы объединяются в кластеры и как кластеры соотносятся между собой.

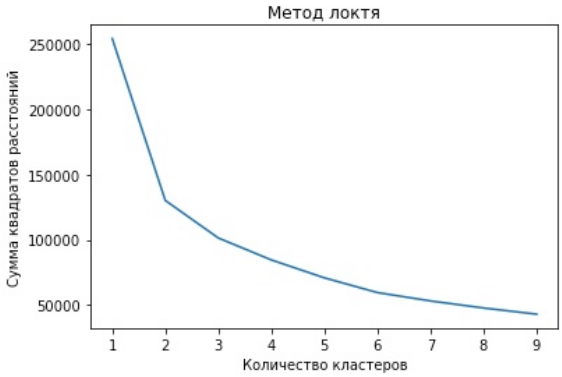


Из дендрограммы можно сделать вывод, что оптимальное количество кластеров для данного набора данных составляет 2. Это количество соответствует вертикальной линии на дендрограмме, которая разделяет ее на три более-менее равных кластера. Также, в качестве подтверждения, мы можем использовать результаты метода локтя, который также указывает на 2 как оптимальное число кластеров.

Метод локтя является одним из способов выбора оптимального числа кластеров при кластеризации методом K-средних. Он основан на анализе суммы квадратов расстояний между каждой точкой и ее ближайшим центром кластера. Чем меньше эта сумма, тем лучше кластеризация.

Для определения оптимального числа кластеров, мы строим график суммы квадратов расстояний в зависимости от числа кластеров. На этом графике мы ищем точку, где изменение суммы квадратов расстояний начинает замедляться (визуально напоминает сгиб на графике, как на изгибе локтя). Это означает, что дальнейшее увеличение числа кластеров

не приведет к существенному уменьшению суммы квадратов расстояний, а



может даже привести к ухудшению качества кластеризации.

На этом графике мы можем выбрать оптимальное число кластеров, где изменение суммы квадратов расстояний начинает замедляться. В данном случае, если мы посмотрим на график, то можно увидеть, что точка перегиба находится при числе кластеров равном 2, что подтверждает результаты метода дендрограммы.

На основе этого выбранного числа кластеров, можно провести кластеризацию методом K-средних, чтобы получить окончательное разбиение данных на кластеры.

Кластеризация методом K-средних

Для проведения кластеризации методом к-средних на понормализованных и стандартизованных данных на Python мы можем использовать библиотеку scikit-learn.

В результате получим значения метрики силуэта для каждого типа предобработки данных:

```
Silhouette score for standardized data:  0.22402927455919058
Silhouette score for normalized data:  0.2861500445431168
```

Значение метрики силуэта для нормализованных данных выше, что говорит о более качественной кластеризации в этом случае. Поэтому, можно сделать вывод, что для данного датасета более подходящим типом предобработки данных является нормализация.

Выбор расстояния

В случае нашего датасета, так как он содержит различные типы переменных (непрерывные, бинарные), было бы целесообразно использовать косинусное расстояние. Это связано с тем, что оно измеряет угол между векторами, а не их абсолютное расстояние. Таким образом, косинусное расстояние более устойчиво к различиям масштабов переменных и типам переменных.

Однако, в методе k-средних, используемом в данном случае, расстояние измеряется в целочисленных единицах, поэтому для данного метода было бы более подходящим использовать евклидово расстояние, которое измеряет расстояние между двумя точками в пространстве. Это особенно важно при использовании метода k-средних, который определяет центроиды как среднее значение точек в кластере.

Построение графика средних значений кластеров

График средних значений для каждого кластера позволяет лучше понять, какие параметры вносят наибольший вклад в формирование кластеров и как они различаются между кластерами.

На основе анализа результатов кластеризации можно предложить следующую интерпретацию:

Кластер 1 - «Низкий риск»

Характеризуется более высокими значениями по параметрам кровяное давление, уровень холестерина, но более низкой максимальной

частоте сердцебиений. Кластер 1 отличается более низким риском возникновения приступа. Данные выводы о том, что это кластер характеризуется более низким риском, обусловлены корреляционным анализом, а именно тем, что в этом кластере переменные с наибольшей корреляцией имеют более низкие значения. Все это то может указывать на то, что у пациентов в этом кластере нет серьезного риска возникновения приступа, и они могут требовать более легкого подхода к лечению и профилактике.

Кластер 2 - «Высокий риск»

Характеризуется более низкими значениями по параметрам кровяное давление, уровень холестерина, но максимальная частота сердцебиений наоборот выше, как и наличие симптомы болей в груди. Кластер 2 имеет более высокий риск возникновения приступа. Это может указывать на то, что у пациентов в этом кластере отсутствуют серьезные заболевания сердечно-сосудистой системы, и они могут требовать более легкого подхода к лечению и профилактике. Это может указывать на то, что у пациентов в этом кластере есть серьезный риск возникновения сердечного приступа и следует принимать меры по их лечению и профилактике.

На основе графика средних значений можно сделать вывод, что параметры кровяное давление, уровень холестерина, максимальная частота сердцебиений и тип боли в груди являются наиболее значимыми при формировании кластеров, и они вносят наибольший вклад в различия между кластерами.

Выводы по кластерному анализу

В данном датасете была проведена кластеризация методом k-средних на два кластера. Были выделены два кластера с разными характеристиками и составами.

Первый кластер, который назвали "Высокий риск", включает пациентов с более высокими показателями максимальной частоты сердцебиений, а также более тяжелым типом грудной боли. Этот кластер имеет более высокий риск возникновения приступа. Он составляет около 56% от общего числа пациентов в датасете.

Второй кластер, названный "Низкий риск", составляет оставшиеся 44% пациентов. Он включает в себя пациентов с более низкими показателями максимальной частоты сердцебиений, а также имеет более низкий риск возникновения приступа.

Данный анализ помог выделить две группы пациентов с разным уровнем риска возникновения приступа. Он может быть полезен для разработки более индивидуальных подходов к лечению пациентов, особенно для тех, кто относится к "Высокому риску».

Выводы по работе

В данной работе был произведен комплексный анализ датасета, состоящего из независимых параметров: возраст, тип грудной боли, кровяное давление, уровень сахара в крови, максимальная частота сердцебиений и уровень холестерина, а также зависимого параметра - риск возникновения приступа.

Предварительный анализ данных позволил установить, что в датасете отсутствуют пропущенные значения, но имеются выбросы, которые были обработаны исходя из контекста задачи. Также были построены графики рассеяния, что позволило визуально оценить распределение и взаимосвязи между переменными.

Корреляционный анализ показал наличие значимой корреляции между переменными. Это позволило более детально изучить взаимосвязи между ними в датасете.

Кластерный анализ позволил выделить две группы пациентов с разными характеристиками и составами. Первый кластер, который был назван "Высокий риск", включал пациентов с более высокими показателями максимальной частоты сердцебиений и более высоким риском возникновения приступа. Второй кластер, названный "Низкий риск", включал пациентов с более низкими показателями данных переменных и более низким риском возникновения приступа.

Исходя из проведенных анализов, можно сделать вывод, что данный датасет имеет значительную долю взаимосвязей между переменными, что необходимо учитывать при их анализе. Кластерный анализ позволил выделить группы пациентов с разным уровнем риска, что может быть полезным для разработки более индивидуальных подходов к лечению