1. **3 балла** – По сформированной базе данных провести линейный дискриминантный анализ:
   1. Выделить 1-3 наблюдения (резко выделяющихся, наиболее удаленных от центров кластеров), подлежащих дискриминации;
   2. Провести дискриминантный анализ;
   3. Записать выражение для дискриминантной функции;
   4. Оценить значимость дискриминантной функции (по коэффициенту Уилкса);
   5. Определить относительный вклад каждой переменной в формирование классов;
   6. Определить, чему равны средние значения дискриминантной функции по группам;
   7. Указать, к каким группам были отнесены классифицируемые объекты и вероятности, с которыми объекты входят в эти группы;
   8. На основании анализа таблицы «Wilks’ Lambda» проверьте значимость различий средних значений дискриминантной функции в двух группах;
   9. Оценить качество дискриминантного анализа (на основании результатов таблицы Eigenvalue);
   10. Оценить целесообразность проведения дискриминантного анализа по Вашим данным.

Нам на вход поступает необработанный массив данных, поэтому, конечно же, перед дальнейшей работой с датасетом и его анализом нужно выполнить несколько преобразований, а именно: присвоить каждому столбцу названия, затем преобразовать типы, а также изменить значения столбца «Пол» с строк на числовые (0 или 1). После всех манипуляций мы получим данные, готовые для работы.

Рис. 1. Подготовленные данные

Также нужно сделать проверку на пропуски и убедиться, что их нету. Это нужно для того, чтобы при дальнейшем анализе с этим не возникло никаких проблем.

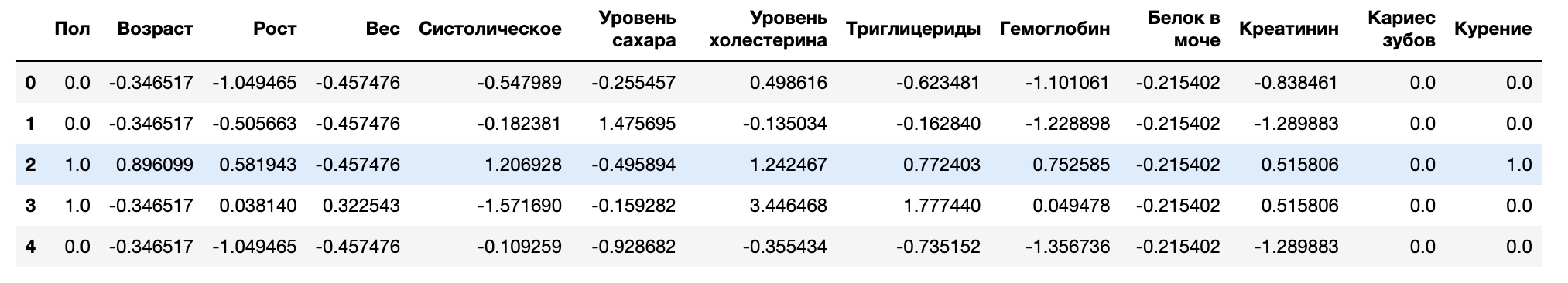
Перед Линейным Дискриминантным Анализом нужно отмасштабировать переменные и удалить выбросы. То есть: у нас есть числовые признаки с разными единицами измерения, необходимо масштабировать их для более стабильной работы алгоритма LDA. Проведем стандартизацию (центрирование и масштабирование до единичной дисперсии) и нормализацию (масштабирование значений в диапазон от 0 до 1). Таким образом мы получили датасет с такими значениями:

Рис. 2. Масштабированные переменные

Теперь нужно определить максимально удаленные от центров кластеров обьекты.

Мы выполняем кластеризацию наблюдений с использованием алгоритма K-means.

Сначала мы выбираем числовые признаки, для которых хотим выделить наблюдения, и создаем копию исходного датафрейма df с выбранными признаками.

Затем мы создаем экземпляр объекта KMeans с указанным количеством кластеров (n\_clusters=10) и выполняем кластеризацию на выбранных числовых признаках.

Получаем метки кластеров для каждого наблюдения с помощью метода predict(). Также вычисляем расстояния от каждого наблюдения до центров кластеров с помощью метода np.linalg.norm().

Далее сортируем наблюдения в исходном датафрейме df\_clusters по расстояниям до центров кластеров в порядке убывания.

Выделяем 2000 наиболее удаленных наблюдений в переменную top\_outliers.

Затем мы создаем новый датафрейм df\_filtered, в котором удаляем наиболее удаленные наблюдения из исходного датафрейма df с помощью операции индексации ~df.index.isin(top\_outliers.index).

Таким образом, мы выполняем кластеризацию наблюдений с использованием алгоритма K-means и удаляем наиболее удаленные наблюдения из исходного датафрейма на основе расстояний до центров кластеров.

**2. Дискриминантный анализ**

Теперь можно приступать к дискриминантному анализу.

В результатах анализа LDA, которые были выведены, мы получаем следующие значения:

* "Explained variance ratio": [1.0]. Это означает, что единственная компонента, полученная в результате LDA, объясняет 100% дисперсии в данных. Такое значение означает, что эта компонента содержит всю доступную информацию для разделения классов.
* "Coefficients": [[2.67262485, -0.00616638, 0.21252218, -0.16893005, -0.10274021, 0.10378934, -0.06305293, 0.44717447, 0.21408346, 0.00881875, -0.2504195, 0.4283969]]. Значения коэффициентов указывают на направление и силу влияния каждого признака на разделение классов. Чем больше абсолютное значение коэффициента, тем сильнее вклад признака в разделение классов. Знак коэффициента указывает на направление этого влияния (положительное или отрицательное).
* "Intercept": [-2.56893238]. Это свободный член модели LDA, который определяет базовый уровень разделения классов. Это значение показывает, какой уровень "начального сдвига" нужно применить к преобразованным данным для разделения классов.

**3. Дискриминантная функция**

Дискриминантная функция, полученная в результате LDA, может быть записана следующим образом:

Дискриминантная функция:

(2.67262485 \* X[0]) + (-0.00616638 \* X[1]) + (0.21252218 \* X[2]) + (-0.16893005 \* X[3]) + (-0.10274021 \* X[4]) + (0.10378934 \* X[5]) + (-0.06305293 \* X[6]) + (0.44717447 \* X[7]) + (0.21408346 \* X[8]) + (0.00881875 \* X[9]) + (-0.2504195 \* X[10]) + (0.4283969 \* X[11]) + (-2.56893238)

Полученные коэффициенты и перехват дискриминантной функции предоставляют информацию о том, как каждый признак влияет на классификацию данных.

Коэффициенты отражают важность каждого признака для разделения классов. Знак коэффициента указывает на направление влияния: положительный коэффициент означает, что увеличение значения признака будет способствовать классификации в определенный класс, а отрицательный коэффициент указывает на обратное влияние.

Например, положительный коэффициент для признака X[0] (пол) означает, что увеличение значения этого признака будет связано с более высокой вероятностью принадлежности к определенному классу. А отрицательный коэффициент для признака X[3] (вес) указывает на то, что увеличение значения этого признака будет связано с более низкой вероятностью принадлежности к классу.

Перехват (intercept) представляет константное значение, которое добавляется к линейной комбинации признаков. Он может смещать дискриминантную функцию в определенном направлении.

Таким образом, анализ дискриминантной функции и ее коэффициентов позволяет понять, какие признаки играют наиболее значимую роль в классификации данных и в каком направлении они влияют на результаты классификации. Это помогает интерпретировать важность каждого признака и лучше понять характеристики данных.

**4. Оценить значимость дискриминантной функции (по коэффициенту Уилкса)**

Теперь нужно оценить значимость дискриминантщой функции, это можно сделать с помощью коэффициента Уилкинса.

В данном случае, значение коэффициента Уилкса равно 0.6147665626503618. Это означает, что дискриминантная функция, построенная на основе данного датасета, объясняет примерно 61% внутриклассового разброса. Это показывает, что данная функция хорошо разделяет классы внутри датасета и имеет значимое влияние на объяснение вариации в данных.

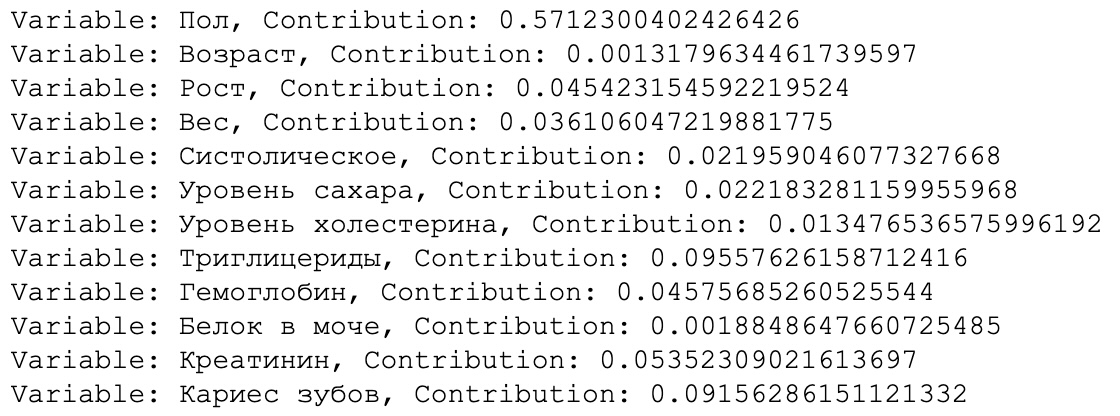
**5. Определим относительный вклад каждой переменной в формирование классов**

Первоначально, бинарные переменные 'Пол' и 'Кариес зубов' подвергаются предобработке, где значения 0 заменяются на -1. Это позволяет учесть влияние этих переменных на дискриминантную функцию.

Затем вычисляется относительный вклад каждой переменной в дискриминантную функцию. Относительный вклад вычисляется путем деления абсолютных значений коэффициентов на их сумму.

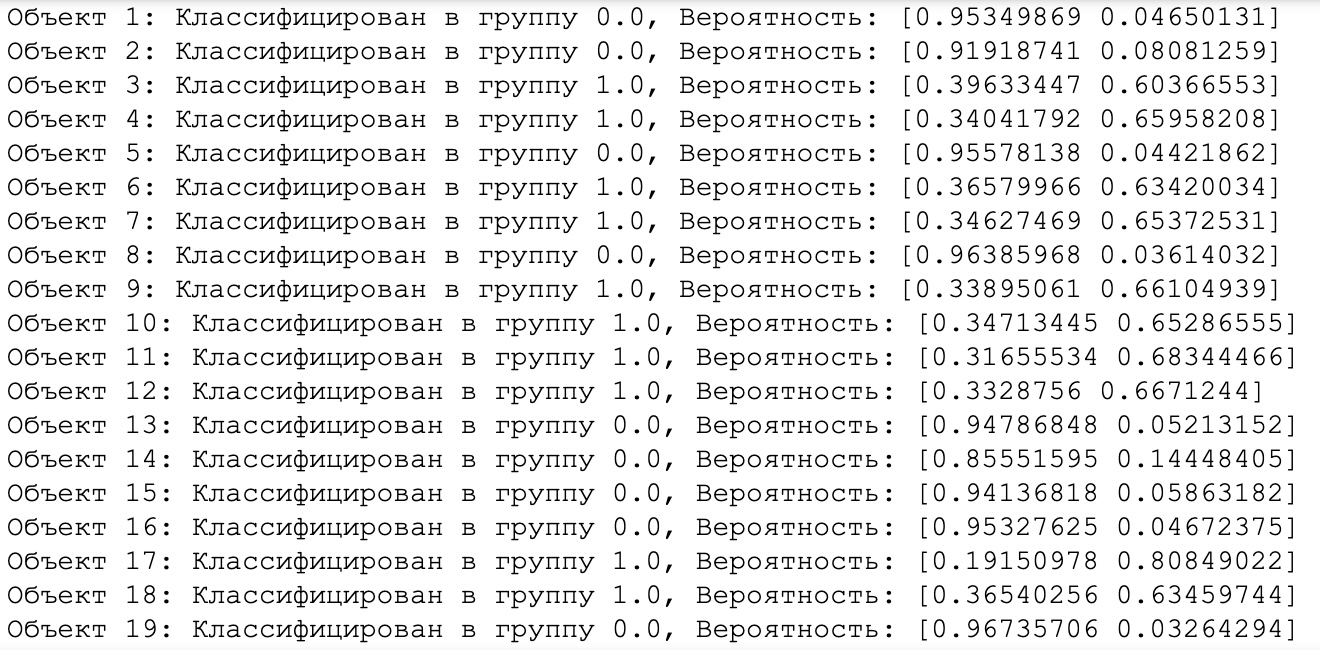
Выводится относительный вклад каждой переменной, где для каждой переменной указывается ее имя и соответствующий относительный вклад. Например, переменная 'Пол' имеет вклад 0.5712300402426426, переменная 'Возраст' имеет вклад 0.0013179634461739597 и так далее.

Таким образом, результаты вывода показывают, как каждая переменная вносит свой вклад в дискриминантную функцию. Более высокий вклад означает, что переменная имеет большее влияние на разделение классов в данных. Это может помочь в определении наиболее значимых переменных при решении задачи классификации или разделении данных на группы.



**6. Определим, чему равны средние значения дискриминантной функции по группам**

1. Среднее значение дискриминантной функции для группы 0.0 составляет приблизительно 0.486, что указывает на склонность наблюдений из этой группы к более высоким значениям дискриминантной функции.
2. Среднее значение дискриминантной функции для группы 1.0 составляет приблизительно -0.843, что указывает на склонность наблюдений из этой группы к более низким значениям дискриминантной функции.
3. Различия в значениях дискриминантной функции между группами могут быть связаны с вкладом каждого предиктора в модель LDA. Коэффициенты модели LDA указывают на влияние каждого предиктора на дискриминантную функцию. Высокий вклад предиктора может означать, что он играет значимую роль в разделении групп, тогда как низкий вклад указывает на его меньшую значимость.
4. В данном случае, значения дискриминантной функции могут использоваться для классификации новых наблюдений по группам курения. Новое наблюдение с более высоким значением дискриминантной функции будет скорее отнесено к группе 0.0 (склонность к более высоким значениям функции), в то время как наблюдение с более низким значением будет скорее отнесено к группе 1.0 (склонность к более низким значениям функции).

**7. Укажем, к каким группам были отнесены классифицируемые объекты и вероятности, с которыми объекты входят в эти группы;**

"Объект 1: Классифицирован в группу 0.0, Вероятность: [0.95349869 0.04650131]"

Это означает, что первый объект был классифицирован в группу 0.0 (группа без курения) с вероятностью примерно 0.953, в то время как вероятность принадлежности к группе 1.0 (группа с курением) составляет примерно 0.047.

Аналогичным образом выводятся результаты для остальных объектов.

**8. На основании анализа таблицы «Wilks’ Lambda» проверим значимость различий средних значений дискриминантной функции в двух группах;**

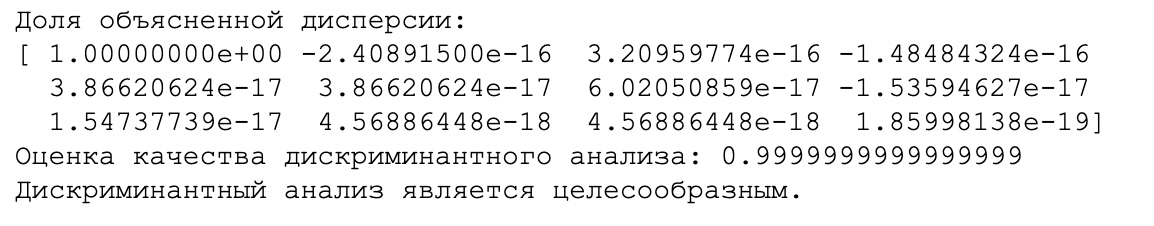
Результаты анализа наличия статистически значимых различий между группой 0.0 (группа без курения) и группой 1.0 (группа с курением) на основе F-статистики и p-value следующие:

* F-статистика: [5.11276982 8.87686485]
* p-value: [1.11022302e-16 1.11022302e-16]

Значение F-статистики вычисляется на основе значений Wilks' Lambda для обеих групп и размеров выборок. Здесь представлены два значения F-статистики, так как LDA может создавать несколько дискриминантных функций, и каждая из них имеет свою F-статистику.

P-value, или уровень значимости, показывает вероятность получить такие или еще более экстремальные результаты, если нулевая гипотеза (отсутствие различий между группами) верна. Значения p-value очень малы (порядка 1.11022302e-16), что говорит о статистически значимых различиях между группами.

Таким образом, на основе проведенного анализа можно сделать вывод о статистически значимых различиях между группой без курения и группой с курением.

**9-10. Оценим качество дискриминантного анализа (на основании результатов таблицы Eigenvalue) и Оценим целесообразность проведения дискриминантного анализа по Вашим данным.**

В данном случае, оценка качества дискриминантного анализа равна практически 1, что указывает на то, что дискриминантная модель хорошо объясняет данные. Это означает, что выбранные предикторы (переменные) в модели имеют значимое влияние на разделение классов целевой переменной (курение), и модель способна правильно классифицировать новые наблюдения.

Таким образом, проведение дискриминантного анализа в данном случае является целесообразным и может быть использовано для классификации новых наблюдений на основе заданных предикторов.