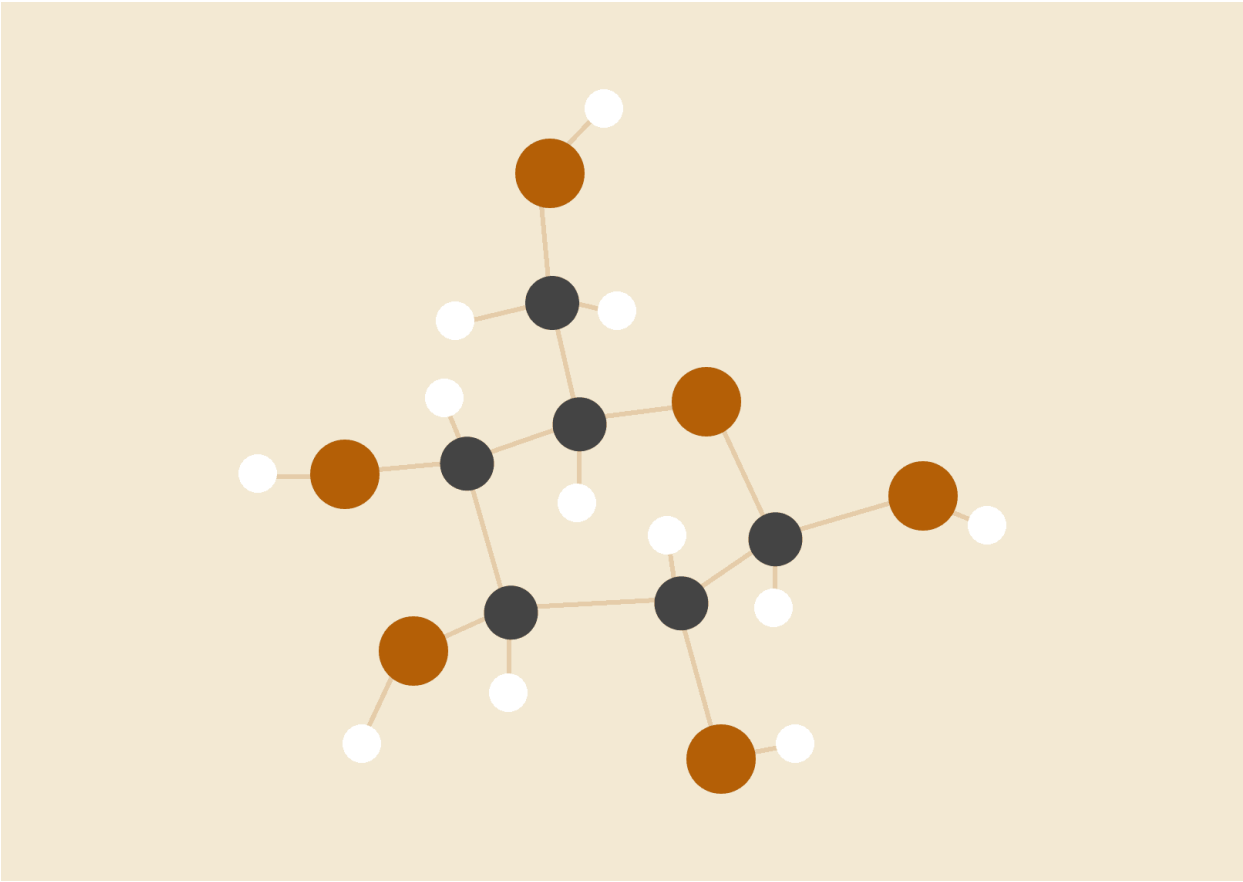
МОДУЛЬНАЯ РАБОТАЯ



Карнаушенко Михаил

Цыплаков Александр

Ермаков Семен

2023

ВВЕДЕНИЕ

В данном отчете представлены результаты работы по домашнему заданию №2 по майнору "Прикладной статистический анализ" в рамках дисциплины "Классификация данных". Целью данного задания было применение различных методов классификации данных, анализ базы данных и получение результатов, позволяющих сделать выводы о качестве проведенного анализа.

Первая часть работы включала формирование собственной базы данных, определение проблематики исследования и включение необходимых переменных. В нашей базе данных было рассмотрено более 50000 наблюдений. В результате была сформулирована задача классификации, где одна из переменных является целевой.

Далее был проведен линейный дискриминантный анализ на основе сформированной базы. В ходе анализа были выделены наблюдения, подлежащие дискриминации, и проведен сам анализ, включая оценку значимости дискриминантной функции, определение вклада каждой переменной и классификацию объектов с указанием вероятностей принадлежности к группам.

Третья часть работы включала построение дерева классификации с использованием метода CRT. Были выбраны зависимая переменная и различные независимые переменные, построены деревья, выбрано оптимальное дерево, проведена визуализация результатов и интерпретация.

В заключительной части работы была выполнена декомпозиция смеси распределений на основе предоставленных данных. Каждое распределение, входящее в состав смеси, было описано и визуализировано, а также определено, к какому распределению относятся определенные наблюдения.

Анализ полученных результатов и проведенных методов классификации позволил сделать выводы о качестве проведенного исследования и о его целесообразности для рассматриваемых данных.

Результаты работы приведены в следующих разделах отчета, включая описание методов, использованных при анализе данных, промежуточные и итоговые результаты, а также их интерпретацию.

ОПИСАНИЕ ДАННЫХ

В качестве данных для исследования послужила открытая база National Health Insurance Corporation - Южнокорейской здравоохранительной организации, находящейся в сотрудничестве с государством и являющейся ключевой в этой области в рамках страны.

Это анонимная база данных резидентов, содержащая более 50 тысяч наблюдений. Она содержит достаточно большое количество визуальных и первичных лабораторных данных о каждом человеке. Среди них интересующая нас бинарная target переменная - факт о том, курит человек или нет.
Исследовательская часть работы посвящена возможности классификации наличия такой вредной привычки у испытуемого на основе так называемых “сигналов тела” - то есть физико-химических данных. Работы в таком направлении крайне актуальны, так как все больше ресурсов вкладывается в использование новых технологий в сфере медицины, начиная от аппаратных решений, и заканчивая нейросетями, обрабатывающими данные о больном. Так, например, по результатам данной работы можно будет судить о валидности систем распознавания, работающих с относительным минимумом данных.

Итак, какие же показатели имеются

Показатель	Количество непустых наблюдений	Тип показателя
gender	55692	binary
age	55692	float
height	55692	float
weight	55692	float
systolic	55692	float
fasting blood sugar	55692	float
Cholesterol	55692	float
triglyceride	55692	float
hemoglobin	55692	float
Urine protein	55692	float
serum creatinine	55692	float
dental caries	55692	binary
smoking	55692	binary

Более подробно про каждый показатель.

systolic - верхнее артериальное давление. Никотин, содержащийся в в сигаретах в момент каждого попадания в кровь вызывает н-холиномиметическое воздействие, тем самым повышая АД. С другой стороны, постоянное воздействие никотина вырабатывает резистентность к такого рода воздействиям, что влечет достаточно индивидуальный результат в различных случаях. Но, конечно, есть общий принцип - никотин действует на сосуды, они влияют на давление. Гибкость сосудов снижается - возникает гипертония/гипотония.

fasting blood sugar - сахар в крови. В общем случае курение повышает количество сахара в крови.

Cholesterol - холестерин. Исследования по воздействию курения на холестерина продолжаются до сих пор. Воздействие никотина сказывается на возможности организма выводить “плохой холестерин”.

triglyceride - триглицерид. Это молекулы, высокое или низкое содержание которых является достаточно распространенной проблемой. При их недостатке или повышенном содержании в организме кратно повышается резистентность к инсулину, а также риск сердечно-сосудистых заболеваний.

hemoglobin - гемоглобин. У курильщиков, как правило, повышенный гемоглобин. Это связано с пониженным количеством поступающего в кровь кислорода, и, вследствие, повышенного выработка гемоглобина.

Urine protein - содержание белка в моче. Повышенное значение этого показателя может быть признаком заболевания почек.

serum creatinine - содержание креатинина. Его показатели также могут сказать о возможных заболеваниях почек, печени и сердца. Никотин оказывает влияние не нейромедиаторы. Так, например, он может повлиять на выработку адреналина.

dental caries - зубной кариес. Смолы, мутагены и канцерогены, содержащиеся в сигаретах, оказывают сильное влияние на полость рта.

gender, age, height, weight - показатели пола, возраста, роста и веса соответственно.

smoking - показатель того, курит человек, или же нет.

ЛИНЕЙНЫЙ ДИСКРИМИНАНТНЫЙ АНАЛИЗ

1. Обработка данных

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.

	Пол	Возраст	Рост	Вес	Систолическое	Уровень сахара	Уровень холестерина	Триглицериды	Гемоглобин	Белок в моче	Креатинин	Кариес зубов	Курение
0	0.0	40.0	155.0	60.0	114.0	94.0	215.0	82.0	12.9	1.0	0.7	0.0	0.0
1	0.0	40.0	160.0	60.0	119.0	130.0	192.0	115.0	12.7	1.0	0.6	0.0	0.0
2	1.0	55.0	170.0	60.0	138.0	89.0	242.0	182.0	15.8	1.0	1.0	0.0	1.0
3	1.0	40.0	165.0	70.0	100.0	96.0	322.0	254.0	14.7	1.0	1.0	0.0	0.0
4	0.0	40.0	155.0	60.0	120.0	80.0	184.0	74.0	12.5	1.0	0.6	0.0	0.0

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.

Вот несколько графиков визуализации наших данных:

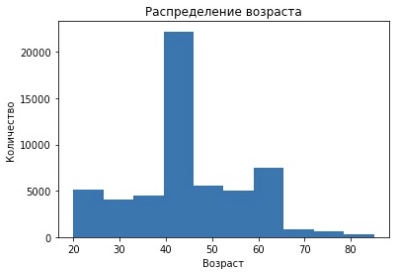


Рис. 2. График распределения возраста

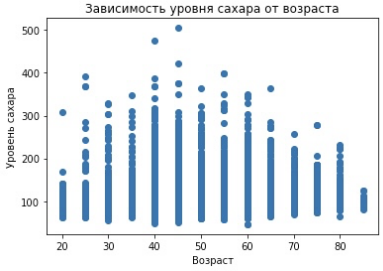


Рис. 3. График зависимости уровня сахара от возраста

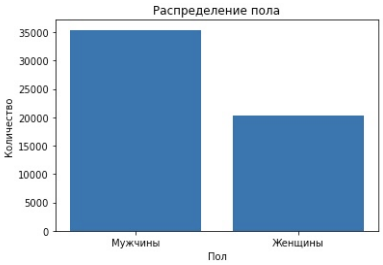


Рис. 4. График распределения пола

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).

Таким образом мы получили датасет с такими значениями:

	Пол	Возраст	Рост	Вес	Систолическое	Уровень сахара	Уровень холестерина	Триглицериды	Гемоглобин	Белок в моче	Креатинин	Кариес зубов	Курение
0	0.0	-0.346517	-1.049465	-0.457476	-0.547989	-0.255457	0.498616	-0.623481	-1.101061	-0.215402	-0.838461	0.0	0.0
1	0.0	-0.346517	-0.505663	-0.457476	-0.182381	1.475695	-0.135034	-0.162840	-1.228898	-0.215402	-1.289883	0.0	0.0
2	1.0	0.896099	0.581943	-0.457476	1.206928	-0.495894	1.242467	0.772403	0.752585	-0.215402	0.515806	0.0	1.0
3	1.0	-0.346517	0.038140	0.322543	-1.571690	-0.159282	3.446468	1.777440	0.049478	-0.215402	0.515806	0.0	0.0
4	0.0	-0.346517	-1.049465	-0.457476	-0.109259	-0.928682	-0.355434	-0.735152	-1.356736	-0.215402	-1.289883	0.0	0.0

Рис. 5. Стандартизированные переменные

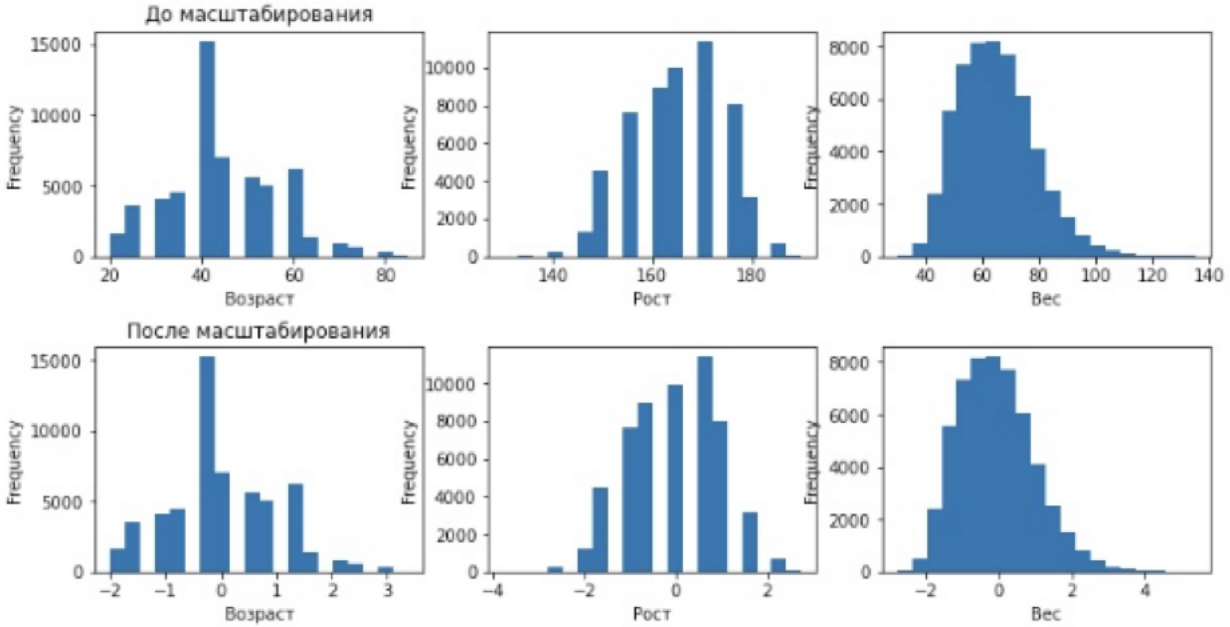


Рис. 6. Гистограммы по данным до и после масштабирования

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: возраст, рост, вес, кровяное давление, уровень сахара, уровень холестерина, триглицериды, гемоглобин, белок в моче и креатинин - выбрали такой набор переменных, потому что они являются непрерывными, в отличие от переменных пол и кариес зубов, которые являются бинарными 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().

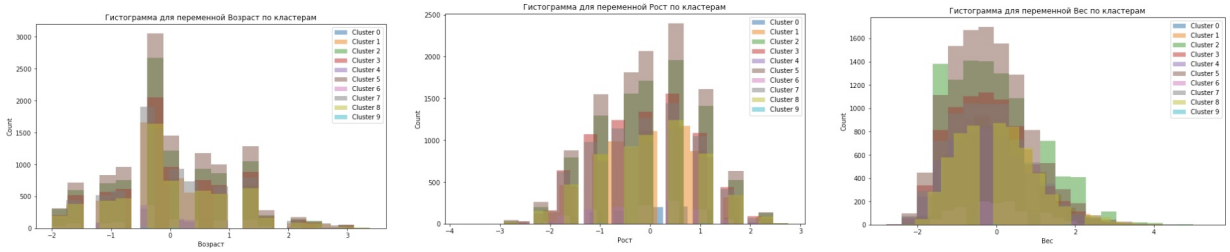
Далее сортируем наблюдения в исходном датафрейма df_clusters по расстояниям до центров кластеров в порядке убывания.

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.

Вот несколько примеров разбиения переменных на кластеры:



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, а именно целевая переменная «курение» содержит всю доступную информацию для разделения классов на «курящих» и «некурящих».

- "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.

Таким образом мы получим график рассеивания:

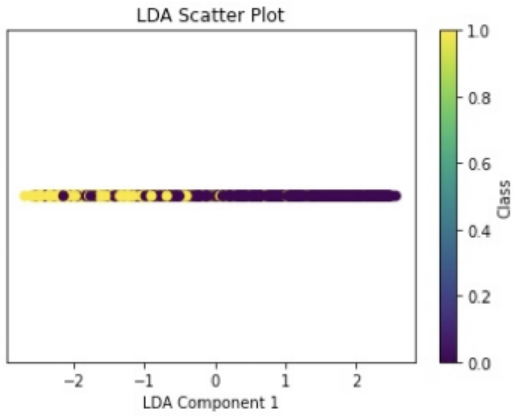


График показывает разделение классов, обозначенных цветами(некурящие - фиолетовый, курящие - желтый). Разделимость: классы хорошо разделимы на основе значения компоненты LDA, можно сделать вывод о хорошей разделимости классов в исходных данных. В таком случае, модель LDA успешно обнаружила линейную комбинацию признаков, которая разделяет классы.

3. Дискриминантная функция

Дискриминантная функция - это функция, которая используется в линейном дискриминантном анализе (ЛДА) для классификации объектов на основе их признаков. Цель ЛДА состоит в поиске линейной комбинации признаков, которая максимально разделяет классы и минимизирует разброс внутри каждого класса.

Эта функция вычисляет значения для каждого класса и на основе этих значений принимает решение о принадлежности объекта к определенному классу. Для двух классов, дискриминантная функция представляет собой линию (в двумерном случае) или гиперплоскость (в многомерном случае), которая разделяет классы.

Дискриминантная функция может использоваться для прогнозирования классов новых наблюдений, которые не были использованы при обучении модели

Теперь перейдем от теории к практике.

В нашем случае она принимает следующий вид:

```
Дискриминантная функция:
(2.6586696649037935 * X[0]) + (-0.015227774026650782 * X[1]) + (0.2150758
6919302232 * X[2]) + (-0.16884788220634356 * X[3]) + (-0.1014591342967249
4 * X[4]) + (0.10579242118935972 * X[5]) + (-0.063689581828526 * X[6]) +
(0.4353467726466337 * X[7]) + (0.21153066538165072 * X[8]) + (0.012118481
766257431 * X[9]) + (-0.2464156408100226 * X[10]) + (0.4241455394003204 *
X[11]) + (-2.556460612059554)
```

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

4. Оценка значимости дискриминантной функции (по коэффициенту Уилкинса)

Коэффициент Уилкса - это мера значимости дискриминантной функции в линейном дискриминантном анализе (ЛДА). Он используется для оценки, насколько хорошо дискриминантная функция разделяет классы и какое количество дисперсии в данных объясняется этой функцией.

Коэффициент Уилкса вычисляется путем сравнения дисперсии внутри классов (дисперсия остатков) с общей дисперсией в данных.

Формула для вычисления коэффициента Уилкса:

Wilks_lambda = (П(1 / (1 + λi)))^(C - 1)

Где:

- П обозначает произведение
- λi - собственное значение, полученное при выполнении ЛДА
- C - количество классов

Интерпретация значения коэффициента Уилкса:

Значение коэффициента Уилкса находится в диапазоне от 0 до 1. Чем ближе значение к 0, тем более значима дискриминантная функция. Если коэффициент Уилкса равен 1, это означает, что дискриминантная функция не даёт никакой информации о разделении классов.

В вашем случае, полученное значение коэффициента Уилкса равно 0.615203876748385. Это говорит о том, что дискриминантная функция в значительной мере объясняет дисперсию в данных и является значимой для

разделения классов.

5. Определение относительного вклад каждой переменной в формирование классов

Относительный вклад каждой переменной в формирование классов позволяет оценить важность каждой переменной для разделения классов в модели. Чем больше значение вклада, тем большую роль играет соответствующая переменная в разделении классов.

```
Variable: Пол, Contribution: 0.5712300402426426
Variable: Возраст, Contribution: 0.0013179634461739597
Variable: Рост, Contribution: 0.045423154592219524
Variable: Вес, Contribution: 0.036106047219881775
Variable: Систолическое, Contribution: 0.021959046077327668
Variable: Уровень сахара, Contribution: 0.022183281159955968
Variable: Уровень холестерина, Contribution: 0.013476536575996192
Variable: Триглицериды, Contribution: 0.09557626158712416
Variable: Гемоглобин, Contribution: 0.04575685260525544
Variable: Белок в моче, Contribution: 0.0018848647660725485
Variable: Креатинин, Contribution: 0.05352309021613697
Variable: Кариес зубов, Contribution: 0.09156286151121332
```

Интерпретация полученных значений относительного вклада переменных:

- Переменная "Пол" имеет очень высокий вклад в формирование классов (57%).
- Все остальные переменные, такие как: "Возраст", "Рост", "Вес", "Кровяное давление", "Уровень сахара", "Гемоглобин" и "Белок в моче", "Уровень холестерина", "Триглицериды", "Креатинин" имеют небольшой вклад (от 1.32% до 9%).

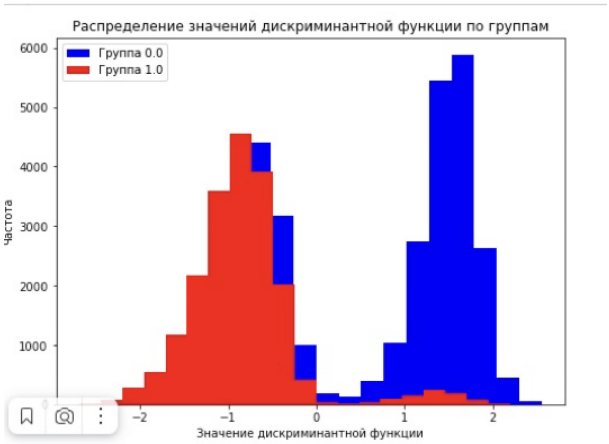
6. Определение относительного вклад каждой переменной в формирование классов

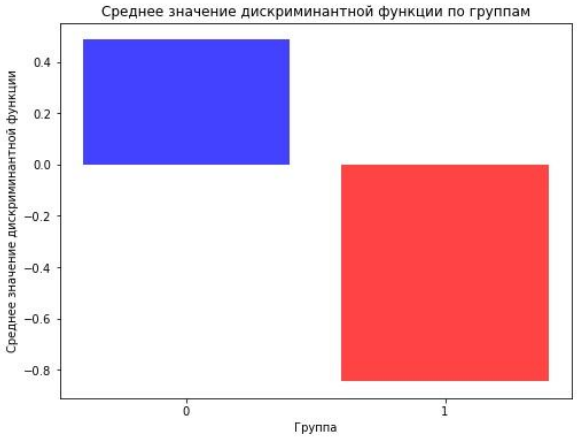
Среднее значение дискриминантной функции для группы 0.0 составляет приблизительно 0.486, что указывает на склонность наблюдений из этой группы к более высоким значениям дискриминантной функции.

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.

Различия в значениях дискриминантной функции между группами могут быть связаны с вкладом каждого предиктора в модель LDA. Высокий вклад предиктора может означать, что он играет значимую роль в разделении групп, тогда как низкий вклад указывает на его меньшую значимость.

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. Принадлежность объектов к группам

```
Объект 1: Классифицирован в группу 0.0, Вероятность: [0.95349869 0.04650131]
Объект 2: Классифицирован в группу 0.0, Вероятность: [0.91918741 0.08081259]
Объект 3: Классифицирован в группу 1.0, Вероятность: [0.39633447 0.60366553]
Объект 4: Классифицирован в группу 1.0, Вероятность: [0.34041792 0.65958208]
Объект 5: Классифицирован в группу 0.0, Вероятность: [0.95578138 0.04421862]
Объект 6: Классифицирован в группу 1.0, Вероятность: [0.36579966 0.63420034]
Объект 7: Классифицирован в группу 1.0, Вероятность: [0.34627469 0.65372531]
Объект 8: Классифицирован в группу 0.0, Вероятность: [0.96385968 0.03614032]
Объект 9: Классифицирован в группу 1.0, Вероятность: [0.33895061 0.66104939]
Объект 10: Классифицирован в группу 1.0, Вероятность: [0.34713445 0.65286555]
Объект 11: Классифицирован в группу 1.0, Вероятность: [0.31655534 0.68344466]
Объект 12: Классифицирован в группу 1.0, Вероятность: [0.3328756 0.6671244]
Объект 13: Классифицирован в группу 0.0, Вероятность: [0.94786848 0.05213152]
Объект 14: Классифицирован в группу 0.0, Вероятность: [0.85551595 0.14448405]
Объект 15: Классифицирован в группу 0.0, Вероятность: [0.94136818 0.05863182]
Объект 16: Классифицирован в группу 0.0, Вероятность: [0.95327625 0.04672375]
Объект 17: Классифицирован в группу 1.0, Вероятность: [0.19150978 0.80849022]
Объект 18: Классифицирован в группу 1.0, Вероятность: [0.36540256 0.63459744]
```

"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. Проверка значимости средних значений дискриминантной функции в двух группах на основании анализа таблицы “Wilks’ Lambda”

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. Оценка качества дискриминантного анализа (на основании результатов таблицы Eigenvalue) и целесообразности проведения дискриминантного анализа по нашим данным.

```
Оценка качества дискриминантного анализа: 0.9999999999999999
Дискриминантный анализ является целесообразным.
```

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.

Приложение “LDA и Деревья”

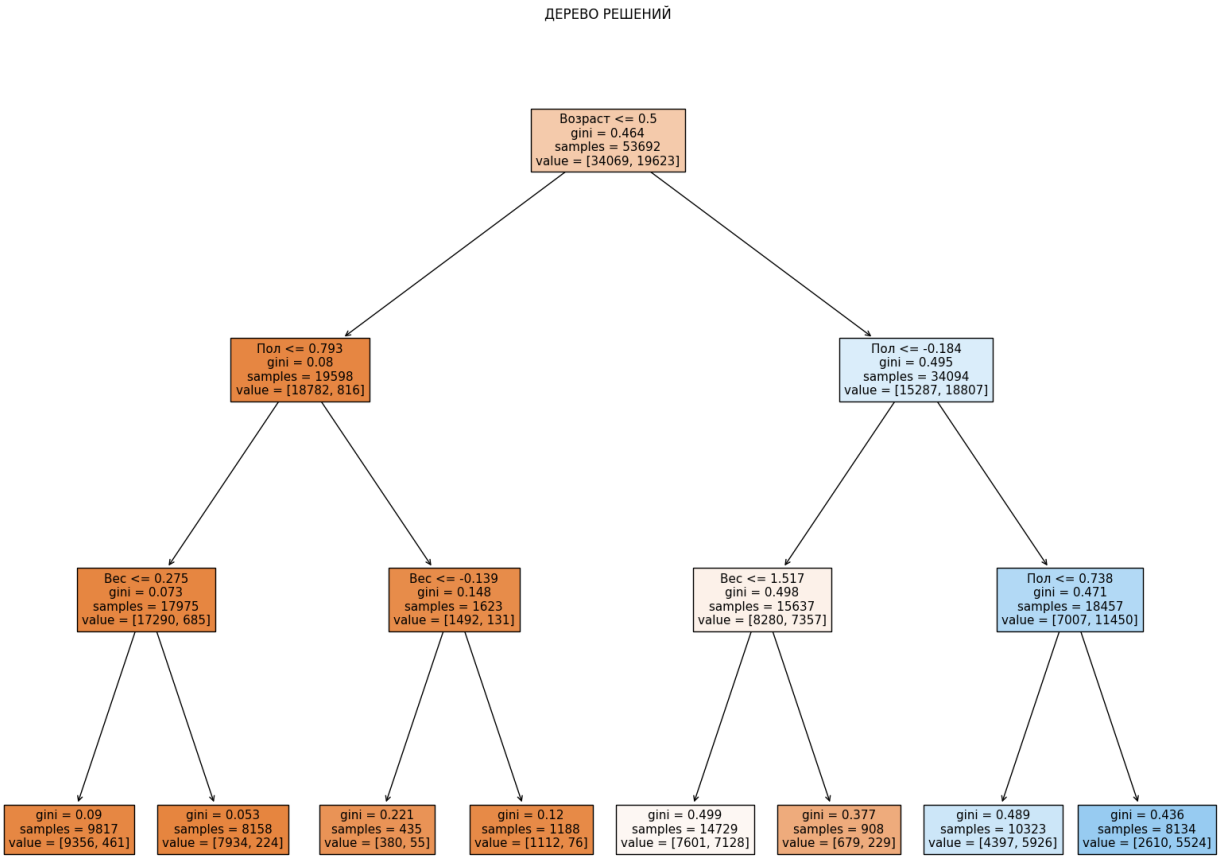
ПОСТРОЕНИЕ ДЕРЕВА КЛАССИФИКАЦИИ

Для построения дерева воспользуемся классификатором DesisionTreeClassifier из пакета sklearn.

Инициализируем объект этого класса с параметрами (max_deep = 5, min_samples_split = 100). У этой модели есть жёсткие ограничения на сложность, что позволит не допустить её переобучение. Такой сложности должно хватить, чтобы модель сама смогла выделить наиболее значимые признаки.

Обучим эту модель на различных подмножествах нашего признакового пространства и найдём наилучшую комбинацию признаков ('Триглицериды', 'Возраст', 'Вес', 'Пол'). Получили точность порядка 74%. Это хороший результат с учётом неравномерности распределения целевой переменной. На этом же подмножестве обучим более глубокое дерево. Получили результат порядка 85% на тесте с кросс валидацией. Это достаточно, чтобы утверждать о предсказательной способности обученной модели.

Далее провизуализируем полученное дерево (только обрежем его до 3-х уровней для наглядной картинки).



На маленькой глубине дерево, как и ожидалось, делит выборку по наиболее общим биометрическим данным. Первое разделение произошло по возрасту, затем в обоих поддеревьях произошло деление по полу. Уже на втором уровне глубины можно заметить существенную разницу по доле курильщиков между молодыми девушками и возрастными мужчинами, что подтверждается жизненным опытом. На глубине равной трём видим деление по весу, причём везде левая подвыборка оказывается менее курящей. Также интересно, что у мужчин вес влияет на вероятность принадлежать к курящим гораздо выше, чем у женщин.

ДЕКОМПОЗИЦИЯ СМЕСИ

Для проведения декомпозиции нам были представлены данные. Задача состоит в том, чтобы представить эти же данные в виде линейной комбинации нормальных распределений с какими-то параметрами.
Начнем с построения гистограммы наших данных:

Как мы видим, в гистограмме для оригинальных, не логарифмированных данных уже заметна комбинация нормальных распределений.

Для полноты построим также гистограмму логарифмированных данных:

Даже визуально выделить на ней распределения крайне сложно.

Займемся дооптимизационной оценкой: посчитаем количество распределений. Настраивая различное количество отображаемых bins, мы пришли к наилучшему результату, когда количество бинов равно 31-ому, а их ширина равна 5.

Мы посчитали, что на графике присутствуют 6 распределений с приблизительными параметрами:

Смесь	μ	σ	q
1	0,10	15	0.10
2	8,00	5	0.01
3	20,22	7	0.16
4	100,00	10	0.50
5	110	5	0.16
6	150,00	7	0.07

Построим суммарное распределение как сумму смесей с коэффициентами q. Для этого посчитаем накопленную частоту:

$$\nu_k = (\sum_{i=1}^{6} N(b_k, \mu_i, \sigma_i) \cdot q_i) \cdot n$$

Где N - нормальное распределение, b_k - k-ый карман, n -количество наблюдений в выборке. После этого мы считаем разность между k и k+1 элементами накопленной частоты и получаем частоту. Гистограмму ее значений мы построим на том же графике, что и изначальные данные:

Здесь оранжевый цвет - изначальные данные, голубой - полученные нашим визуальным приближением.

Теперь, когда у нас есть от чего отталкиваться в поиске более точного разложения, воспользуемся максимизацией функции правдоподобия.

$$f_j = Ln(\sum_{i=1}^{6} N(b_k, \mu_i, \sigma_i) \cdot q_i)/a_j$$

$$F = \sum_{j=1}^{n} f_j$$

Изначально F равнялась -606.368 для нашего визуального приближения. Теперь же, меняя коэффициенты распределений и их параметры, мы должны максимизировать данную функцию.
Это можно сделать несколькими способами, например:

Использовать python. Такой способ удобен, легко поддается отслеживанию на каждой итерации/эпохе, но занимает много времени даже с учетом создания нескольких процессов и используя оптимизационные методы. Наше решение слишком медленно, поэтому мы от него отказались.

Использовать встроенные надстройки Excel от Microsoft. Его алгоритмы более быстрые, но при это и более закрытые - отслеживать каждый шаг не всегда бывает удобно.
Наша задача нелинейная - коэффициенты q входят в уравнение под логарифмом, а среднее значение и СКО входят в составе нормального распределения. Есть несколько встроенных алгоритмов Excel, подходящих под нашу задачу. Это генетический алгоритм, алгоритм роя и обобщенный приведенный градиент(ОПГ).

Первые два дали не очень удовлетворительные результаты. Меняя же параметры ОПГ, мы смогли найти интересующий нас оптимум:

$$F = -466,202323$$

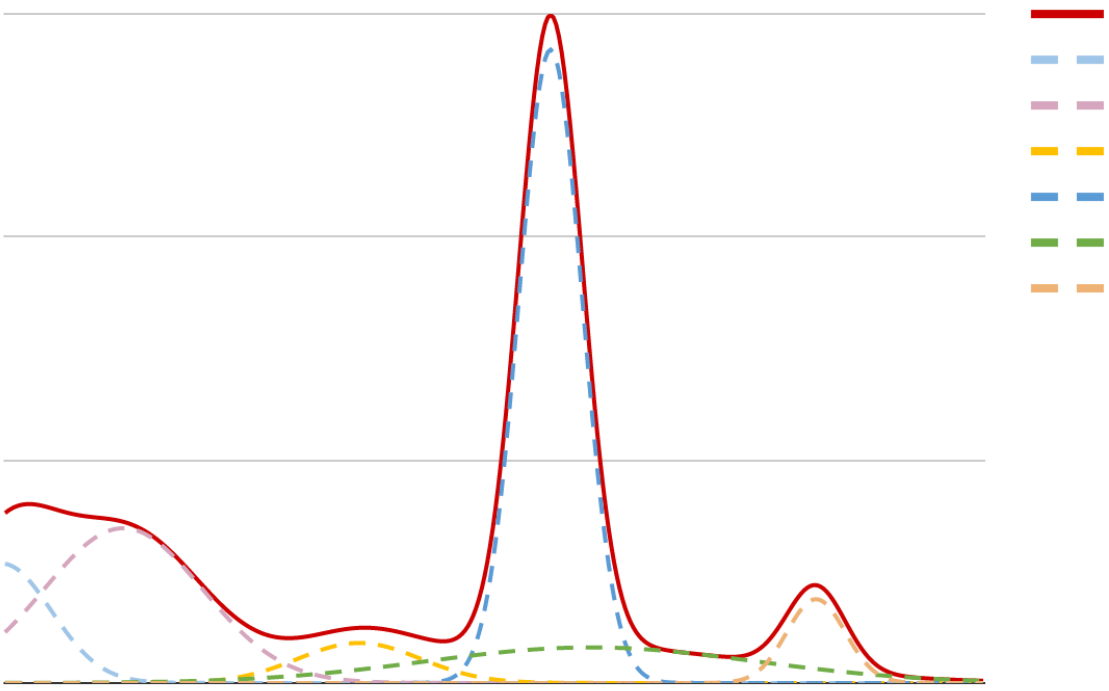
Значения параметров при этом стали следующими:

Смесь	μ	σ	q
1	0,43	8,94	0,12
2	22,46	14,37	0,25
3	65,27	11,13	0,05
4	99,98	5,76	0,41
5	107,3	30,02	0,12
6	148,15	5,31	0,05

Как мы видим, коэффициент в сумме дает 1, как мы и рассчитывали. По аналогии строим накопленные частоты, частоты и строим график:

Оригинал - оранжевый, голубой - предоптимизация и красный - оптимизация. Визуально видно, что красный график ближе к оранжевому, за исключением малых значений. Это может говорить о том, что в при настройке параметров для поиска оптимума была допущена неточность(например, взят слишком большой стартовый шаг, который перескочил максимум).

Чтобы более наглядно показать каждую смесь, сгенерируем на набор интервалов,покрывающий наши данные, и построим нормальные распределения с найденными параметрами и коэффициентами. Полученные смеси просуммируем. Все полученные данные отобразим на график



Красный - сумма смесей

Голубой - 1-ая смесь - с самым маленьким средним значением и небольшой дисперсией. Вносит вклад в большой количество наблюдений с низкими показателями.

Розовый - 2-ая смесь - (22,46; 14,37)

Желтый - 3-ья смесь - (65,27;11,13)

Синий - 4-ая смесь - самая большая с точки зрения своего веса. Небольшая дисперсия делает окрестность вокруг ее среднего значения достаточно крутой.

Зеленый - 5-ая смесь - с самой большой дисперсией и небольшим весом. Значительно сглаживает правую часть синей смеси.

Оранжевый - 6-ая смесь - образует группу объектов с высокими показателями.

Номера Карнаушенко Михаила, Цыплакова Александра и Ермакова Семена - 13, 25

и 12. Они соответствуют наблюдениям с показателями 86,76; 96,64; 89,79

Все эти наблюдения попадают в место, где наибольшее влияние оказывает 4ая смесь(синяя) (Приложение “Смесь”)

ВЫВОДЫ

В современном мире нейронные сети проникают во все сферы нашей жизни, так например, в недавнем исследовании четвёртая версия chatGPT смогла обойти профессиональных врачей в точности рекомендаций для больных. Наше исследование может показать, что статистические методы и методы машинного обучения способны помогать нам в работе.

Результаты LDA показали, что можно провести достаточно достоверное разделение на категории курящих и некурящих людей на основе “сигналов тела”. Дерево решений вещественно показало, как критерии влияют на принадлежность к той или иной группе. Некоторые взаимосвязи объяснимы как логическими и социальными явлениями, так и физиологическими.

ПРИЛОЖЕНИЕ

“LDA и Деревья”

Все вычисления на python доступны для просмотра в ipynb и pdf формате по ссылке

“Смесь”

Интервал	Сумма	1 смесь	2 смесь	3 смесь	4 смесь	5 смесь	6 смесь
89	0,006124426735	0	0,0000001532724342	0,0001846242498	0,005	0,001324304217	0
89,5	0,006930758069	0	0,0000001303854936	0,0001675920985	0,005	0,001337632985	0

Все вычисления доступны для просмотра в файле Modulo2_5