

Национальный исследовательский университет «Высшая школа экономики»
Факультет компьютерных наук
Образовательная программа: Программная инженерия

Отчет по Домашнему заданию №2
«Линейный дискриминантный анализ, дерево классификации и
декомпозиция смеси распределения»
по майнору «Классификация статистических данных»
«Статистическое исследование взаимосвязей переменных
построенной базы данных с целевой переменной: количество
сердечно сосудистых заболеваний по регионам Российской
Федерации за 2023 год»

Работу выполнили:
студент 2 курса
Артемьев Александр Вячеславович
студент 2 курса
Турсунов Жамолиддин Равшан угли
Преподаватель:
Арефьева Валерия Александровна

Москва, 2024г.

Содержание

Оглавление

1. Введение.....	3
1.1 Актуальность темы исследования	3
1.2 Цели исследования	3
1.3 Задачи исследования.....	3
2. Используемые показатели для анализа.....	3
2.1 Выбор данных	3
2.2 Список используемых показателей для анализа	3
2.3 Описание показателей, выбранных для исследования	4
3. Линейный дискриминантный анализ.....	5
3.1 Выделить 1-3 наблюдения (резко выделяющихся, наиболее удаленных от центров кластеров), подлежащих дискриминации	5
3.2 Дискриминантный анализ.....	6
3.2.1 Выбор переменных	6
3.2.2 Анализ с помощью SPSS.....	6
3.3 Уравнение дискриминантной функции	7
3.4 Значимость дискриминантной функции	8
3.5 Определить относительный вклад каждой переменной в формирование классов ...	9
3.6 Средние значения дискриминантной функции по группам	10
3.7 Указать, к каким группам были отнесены классифицируемые объекты и вероятности, с которыми объекты входят в эти группы	10
3.8 Значимость различий средних значений дискриминантной функции в двух группах	11
3.9 Качество дискриминантного анализа	11
3.10 Целесообразность проведения дискриминантного анализа	12
4. Дерево решений.....	12
4.1 Выбор зависимой переменной – дискретная шкала.....	12
4.2 Построение деревьев с помощью метода CHAID с различными независимыми переменными	14
4.3 Выбор оптимального дерева с помощью таблицы классификации	16
4.4 Визуализация результатов.....	17
4.5 Интерпретация результатов.....	18
5. Расщепление смесей.....	19
6. Список литературы	21
7. Используемые информационные источники	22
8. Приложения.....	22

1. Введение

1.1 Актуальность темы исследования

Сердечно-сосудистые заболевания являются одной из ведущих причин смертности и заболеваемости во всем мире. В России, как и в большинстве стран, ССЗ занимают первое место среди причин смертности. Это делает исследование данной проблемы чрезвычайно важным для общественного здоровья.

Важно выявить ключевые факторы, влияющие на распространение ССЗ, чтобы разработать эффективные меры профилактики и управления здоровьем.

1.2 Цели исследования

- Выявить независимые переменные, которые наиболее сильно влияют на уровень сердечно-сосудистых заболеваний среди населения различных регионов.
- Разработать и протестировать модели дискриминантного анализа и дерева классификации для прогнозирования уровня ССЗ на основе изученных факторов.

1.3 Задачи исследования

- Линейный дискриминантный анализ
- Дерево классификации
- Классифицировать количество людей с сердечно-сосудистыми заболеваниями по следующим признакам: Количество спортивных сооружений, процент расходов на алкогольную продукцию, табачные изделия и наркотики, потребление яиц, процент населения старше трудоспособного возраста, количество людей с заболеваниями нервной системы, соотношение числа женщин и мужчин, наличие/отсутствие выхода к морю, высокий/низкий уровень безработицы в регионах РФ

2. Используемые показатели для анализа

2.1 Выбор данных

Датасет был составлен, используя данные с официального сайта Росстата за 2023 год. Проводить анализ с наблюдениями за 2024 год не имеет смысла, так как он не закончился на момент написания данной работы. Был выбран в качестве источника сайт Росстата [1], поскольку такие данные имеют официальный статус и являются официальной статистикой страны, а также из-за широкого охвата и разнообразия данных. Исходные данные представлены в Приложении 1.

2.2 Список используемых показателей для анализа

В данном исследовании используются 10 переменных:

1. Количество людей с сердечно-сосудистыми заболеваниями (целевая переменная, непрерывная)
2. Высокий/низкий уровень сердечно-сосудистых заболеваний (бинарный ответ на целевую переменную)
3. Количество спортивных сооружений (независимый)
4. Процент расходов на алкогольную продукцию, табачные изделия и наркотики (независимый)
5. Потребление яиц (независимый)
6. Процент населения старше трудоспособного возраста (независимый)
7. Количество людей с заболеваниями нервной системы (независимый)

8. Соотношение числа женщин и мужчин (независимый)
9. Наличие/отсутствие выхода к морю (независимый)
10. Высокий/Низкий уровень безработицы (независимый)

2.3 Описание показателей, выбранных для исследования

- 1) Выбранный показатель в качестве целевой переменной – Количество людей с сердечно-сосудистыми заболеваниями.
 - Показатель характеризует количество людей, с сердечно-сосудистыми заболеваниями, на 1000 человек, в конкретном субъекте.
 - Единица измерения – количество человек.
 - Количественные непрерывные данные
- 2) Выбранный показатель в качестве целевой переменной – Высокий/низкий уровень сердечно-сосудистых заболеваний
 - Показатель характеризует высокий/низкий уровень первого признака в конкретном субъекте.
 - Единица измерения – 1 и 0.
 - Бинарные данные
- 3) Выбранный показатель в качестве независимой переменной – Количество спортивных сооружений
 - Показатель характеризует количество спортивных сооружений, на 1000 человек, в конкретном субъекте.
 - Единица измерения – количество объектов, штук
 - Количественные непрерывные данные
- 4) Выбранный показатель в качестве независимой переменной – Процент расходов на алкогольную продукцию, табачные изделия и наркотики
 - Показатель характеризует процент расходов на алкогольную продукцию, табачные изделия и наркотики, в конкретном субъекте.
 - Единица измерения – %.
 - Количественные непрерывные данные
- 5) Выбранный показатель в качестве независимой переменной – Потребление яиц
 - Показатель характеризует потребление яиц в год на душу населения, в конкретном субъекте.
 - Единица измерения – шт.
 - Количественные непрерывные данные
- 6) Выбранный показатель в качестве независимой переменной – Процент населения старше трудоспособного возраста
 - Показатель характеризует процент населения старше трудоспособного возраста, в конкретном субъекте.
 - Единица измерения – %.
 - Количественные непрерывные данные
- 7) Выбранный показатель в качестве независимой переменной – Количество людей с заболеваниями нервной системы.
 - Показатель характеризует количество людей с заболеваниями нервной системы в данном субъекте РФ.
 - Единица измерения - количество человек.
 - Количественные непрерывные данные.
- 8) Выбранный показатель в качестве независимой переменной – Соотношение числа женщин и мужчин

- Показатель характеризует отношения числа женщин к числу мужчин в данном субъекте РФ (состоит под наблюдением на конец отчетного года).
 - Единица измерения - коэффициент.
 - Количественные непрерывные данные.
- 9) Выбранный показатель в качестве независимой переменной – Наличие/отсутствие выхода к морю
- Показатель характеризует наличие выхода к морю для данного субъекта РФ
 - Единица измерения – 1 и 0.
 - Бинарные данные.
- 10) Выбранный показатель в качестве независимой переменной – Высокий/низкий уровень безработицы
- Показатель характеризует уровень безработицы в конкретном регионе или субъекте.
 - Единица измерения – 1 и 0.
 - Бинарные данные.

3. Линейный дискриминантный анализ

3.1 Выделить 1-3 наблюдения (резко выделяющихся, наиболее удаленных от центров кластеров), подлежащих дискриминации

Первый шаг – найти аномальные значения, которые наиболее удалены от центров кластеров. Для этого по каждой переменной был найден квадрат отклонения от математического ожидания рассматриваемого признака по следующей формуле:

$$\sigma^2 = (x - \bar{x})^2$$

После чего значения были от нормированы и для каждого элемента выборки посчитана сумма таких отклонений. (Приложение 2)

$$\sigma_{\text{норм}}^2 = \frac{(x - \bar{x})^2}{\bar{x}^2}$$

Таким образом удалось выделить 3 резко выделяющихся наблюдения: Ненецкий автономный округ, Курганская область и Чукотский автономный округ

	Сумма отклонений
Ненецкий автономный округ	5,42
Курганская область	4,14
Чукотский автономный округ	2,38

Причем каждый из регионов резко выделяется по одной характерной переменной, так для Курганской области - собственно целевая переменная (98,30 с отклонением 3,78), для Ненецкого автономного округа и Чукотского автономного округа – болезни нервной системы (43,40 с отклонением 4,18 и 30,80 с отклонением 1,35 соответственно).

3.2 Дискриминантный анализ

3.2.1 Выбор переменных

В качестве целевой переменной был выбран признак – количество людей с сердечно-сосудистыми заболеваниями. Значения целевой переменной были разбиты на два кластера (min-33,3) и (33,4-max). В качестве переменных-предикторов были выбраны следующие количественные переменные:

- Количество спортивных сооружений
- Процент расходов на алкогольную продукцию, табачные изделия и наркотики
- Потребление яиц
- Процент населения старше трудоспособного возраста
- Болезни нервной системы
- Соотношение числа женщин и мужчин

3.2.2 Анализ с помощью SPSS

После выбора кластеров, при помощи SPSS была получена качественная переменная ClusterIndex и проведен сам дискриминантный анализ.

На этом шаге так же стоит проверить переменные на мультиколлинеарность, поскольку их игнорирование может привести к появлению нестабильных коэффициентов регрессии и чрезмерной чувствительности модели.

		Количество спортивных сооружений на 1000 человек	Процент населения старше трудоспособного возраста	Процент расходов на алкогольную продукцию, табачные изделия и наркотики	Болезни нервной системы на 1000 человек	Потребление яиц шт в год на душу населения	Соотношение числа женщин и мужчин
Корреляция	Количество спортивных сооружений на 1000 человек	1,000	,308	,114	-,089	,033	,311
	Процент населения старше трудоспособного возраста	,308	1,000	,235	-,305	,341	,414
	Процент расходов на алкогольную продукцию, табачные изделия и наркотики	,114	,235	1,000	-,091	,013	,154
	Болезни нервной системы на 1000 человек	-,089	-,305	-,091	1,000	-,133	-,192
	Потребление яиц шт в год на душу населения	,033	,341	,013	-,133	1,000	,279
	Соотношение числа женщин и мужчин	,311	,414	,154	-,192	,279	1,000

Рисунок 1. Объединенные внутригрупповые матрицы

Поскольку все коэффициенты корреляции для разных переменных < 0.5 , то мультиколлинеарности для данных нет и мы можем продолжать анализ для этого набора признаков. Также из анализа сводки обработки наблюдений:

Анализ сводки обработки наблюдений			
Невзвешенные наблюдения		N	Проценты
Валидные		82	100,0
Исключено	Отсутствующие или выходящие за пределы диапазона коды групп	0	,0
	По крайней мере одна дискриминирующая переменная	0	,0
	И отсутствующие или выходящие за пределы диапазоны коды групп, и по крайней мере одна дискриминирующая переменная	0	,0
	Всего	0	,0
Всего		82	100,0

Рисунок 2. Анализ сводки обработки наблюдений

можем понять, что дискриминантный анализ был проведен корректно и все 82 значений оказались валидными.

3.3 Уравнение дискриминантной функции

Поскольку значения целевой переменной были разбиты на 2 кластера, то дискриминантных функций будет 1, и для составления уравнения, воспользуемся следующей таблицей:

**Коэффициенты
канонической
дискриминантной
функции**

	Функция 1
Количество спортивных сооружений на 1000 челове	-,229
Процент население старше трудоспособного возраста	,096
Процент расходов на алкогольную продукцию, табачные изделия и наркотик	,130
Болезни нервной системы на 1000 челове	,160
Потребление яиц шт в год на душу населения	,008
Соотношение числа женщин и мужчин	1,901
(Константа)	-8,750
Нестандартизованные коэффициенты	

Рисунок 3. Коэффициенты ДФ

Полученное уравнение:

$$y = -0,229x_1 + 0,096x_2 + 0,130x_3 + 0,160x_4 + 0,008x_5 + 1,901x_6 - 8,750$$

Где:

- X1 - Количество спортивных сооружений
- X3 - Процент расходов на алкогольную продукцию, табачные изделия и наркотики
- X5 - Потребление яиц
- X2 - Процент населения старше трудоспособного возраста
- X4 - Болезни нервной системы
- X6 - Соотношение числа женщин и мужчин

Данные уравнения показывают, как изменения значения независимого признака изменяет целевую переменную.

3.4 Значимость дискриминантной функции

Для проверки значимости функции обратимся к следующей таблице:

Лямбда Уилкса

Критерий для функций	Лямбда Уилкса	Хи-квадрат	ст.св.	знач.
1	,718	25,546	6	,000

Поскольку p – значение для первой функции = 0 < 0.05, можно сделать вывод что она значима на уровне 5%. При том для дискриминантной функции Лямбда Уилкса = 0.718 – что говорит о средней возможности модели различать группы.

3.5 Определить относительный вклад каждой переменной в формирование классов

Для этого проанализируем следующие таблицы:

Коэффициенты стандартизованной канонической дискриминантной функции	
	Функция 1
Количество спортивных сооружений на 1000 челове	-,118
Процент население старше трудоспособного возраста	,530
Процент расходов на алкогольную продукцию, табачные изделия и наркотик	,151
Болезни нервной системы на 1000 челове	,824
Потребление яиц шт в год на душу населения	,424
Соотношение числа женщин и мужчин	,098

Матрица структуры	
	Функция 1
Болезни нервной системы на 1000 челове	,584
Потребление яиц шт в год на душу населения	,520
Процент население старше трудоспособного возраста	,462
Соотношение числа женщин и мужчин	,263
Процент расходов на алкогольную продукцию, табачные изделия и наркотик	,208
Количество спортивных сооружений на 1000 челове	,033
Объединенные внутригрупповые	

Наибольший вклад вносят переменные болезни нервной системы(0,824), процент населения старше трудоспособного возраста(0,53) и потребление яиц(0,424) причем все три коэффициента в стандартизованной канонической форме положительны, т.е. с увеличением этих показателей увеличивается значение дискриминантной функции, это можно заметить как из матрицы структур (которая демонстрирует силу связи) так и из коэффициентов стандартизованной формы. Переменная количество спортивных сооружений, с одной стороны, вносит значительный отрицательный вклад при включении новых объектов в кластеры, но с другой стороны, имеет наименьшую корреляцию в матрице структур.

3.6 Средние значения дискриминантной функции по группам

Исследуя следующую таблицу, можно сделать определенные выводы:

Функции в центроидах групп	
	Функция
ClusterIndex	1
1,00	,683
2,00	-,562
Нестандартизованные канонические дискриминантные функции, вычисленные в групповых средних	

Разбиение происходит на 2 группы для 1ой дискриминантной функции. Стоит отметить, что значения довольно сильно отличаются между собой, что говорит в первую очередь о хорошем разбиении на группы. Чем больше расстояния между значениями в центроидах тем лучше функция различает группы.

3.7 Указать, к каким группам были отнесены классифицируемые объекты и вероятности, с которыми объекты входят в эти группы

В Приложение 3 указана таблица с классификацией элементов по группам, для каждого элемента записана настоящая группа и предсказанная в соответствующих столбцах. А также в столбце РМ проставлена соответствующая вероятность.

3.8 Значимость различий средних значений дискриминантной функции в двух группах

Критерии равенства групповых средних

	Лямбда Уилкса	F	ст.св.1	ст.св.2	знач.
Количество спортивных сооружений на 1000 челове	1,000	,034	1	80	,855
Процент население старше трудоспособного возраста	,923	6,720	1	80	,011
Процент расходов на алкогольную продукцию, табачные изделия и наркотик	,983	1,356	1	80	,248
Болезни нервной системы на 1000 челове	,882	10,750	1	80	,002
Потребление яиц шт в год на душу населения	,904	8,516	1	80	,005
Соотношение числа женщин и мужчин	,974	2,175	1	80	,144

Как можно заметить из таблицы только для трех переменных значимо различие средних значений (знач. < 0.05) - болезни нервной системы, процент населения старше трудоспособного возраста и потребление яиц, т.е. именно эти три переменные наиболее значимы для дискриминантного анализа. Это можно было заметить и раньше в пункте 3.5, где рассматривали какие переменные больше всего влияют на целевую переменную.

3.9 Качество дискриминантного анализа

Результаты классификации^а

		Предсказанная принадлежность к группе		Всего
		1,00	2,00	
Исходный	Количество	1,00	30	7
		2,00	11	34
	%	1,00	81,1	18,9
		2,00	24,4	75,6
				100,0

а. 78,0% исходных сгруппированных наблюдений классифицированы правильно.

В итоге, верно, сгруппированы 78% результатов – довольно высокий показатель. Также анализируя результаты, можно сделать вывод: лучший результат у элементов 1ой группы

– 81,1% верных предсказаний, а худший соответственно у 2 группы

Собственные значения

Функция	Собственное значение	% дисперсии	Суммарный %	Каноническая корреляция
1	,393 ^a	100,0	100,0	,531

а. Для анализа использовались первые 1 из канонических дискриминантных функций.

Исследуя таблицу собственные значения, каноническая корреляция принимает значение = 0,531, что соответствует средней достоверности дискриминации. Само же собственное значение почти 0,4 довольно большое чтобы не плохо различать группы, но не на идеальном уровне

3.10 Целесообразность проведения дискриминантного анализа

Результаты проведенного анализа довольно неплохие:

- 1) Созданная модель различает группы с близко к высокой точностью (почти 80%)
- 2) Сама модель получилась не идеальная и есть несколько на то причин, как плохое разделение на кластеры изначально, (возможно стоит изменить количество кластеров или границы кластеров) некоторые переменные оказались незначимы и почти не влияли на изменение целевой переменной, поэтому их стоит заменить на более осмысленные.
- 3) Но 3 переменные оказывают большое влияние на целевую переменную:
 - i. Количество яиц (видимо из за большого количества холестерина)
 - ii. Процент населения старше трудоспособного возраста(что показывает большую вероятность получить ССЗ ближе к старости)
 - iii. Заболевания нервной системы(видимо они влияют на кровеносную системы во время стресса или других ярких импульсов)
- 4) Также 3 оставшиеся переменные почти не связаны с целевой переменной

4. Дерево решений

4.1 Выбор зависимой переменной – дискретная шкала

Первым шагом в нашем исследовании было получение и изучение предоставленных данных. Мы импортировали данные из файла Excel, что позволило ознакомиться со структурой и определить подходящие переменные для анализа.

Преобразование переменной:

Для метода CHAID требуется дискретная зависимая переменная, поэтому мы преобразовали количественную переменную, указывающую количество сердечно-сосудистых заболеваний на 1000 человек, в бинарный формат. Это преобразование было выполнено с использованием медианы в качестве порога, что позволило разделить данные на две группы: высокий и низкий риски.

```
import numpy as np

threshold = data['Сердечно-сосудистые заболевания\n на 1 000 человек'].median()

data['Целевая_переменная'] = np.where(
    data['Сердечно-сосудистые заболевания\n на 1 000 человек'] > threshold, 1, 0
)
```

После преобразования переменной мы получили чёткое разделение данных, что позволит точно оценить влияние разных факторов на состояние сердечно-сосудистой системы населения.

4.2 Построение деревьев с помощью метода CHAID с различными независимыми переменными

В ходе подготовки данных к анализу, мы исключили из выборки неинформативные переменные, такие как наименования регионов, и сосредоточились на переменных, имеющих потенциальное значение для предсказания сердечно-сосудистых заболеваний. Оставшиеся данные были разделены на обучающую и тестовую выборки.

```
from sklearn.model_selection import train_test_split
from sklearn.tree import DecisionTreeClassifier

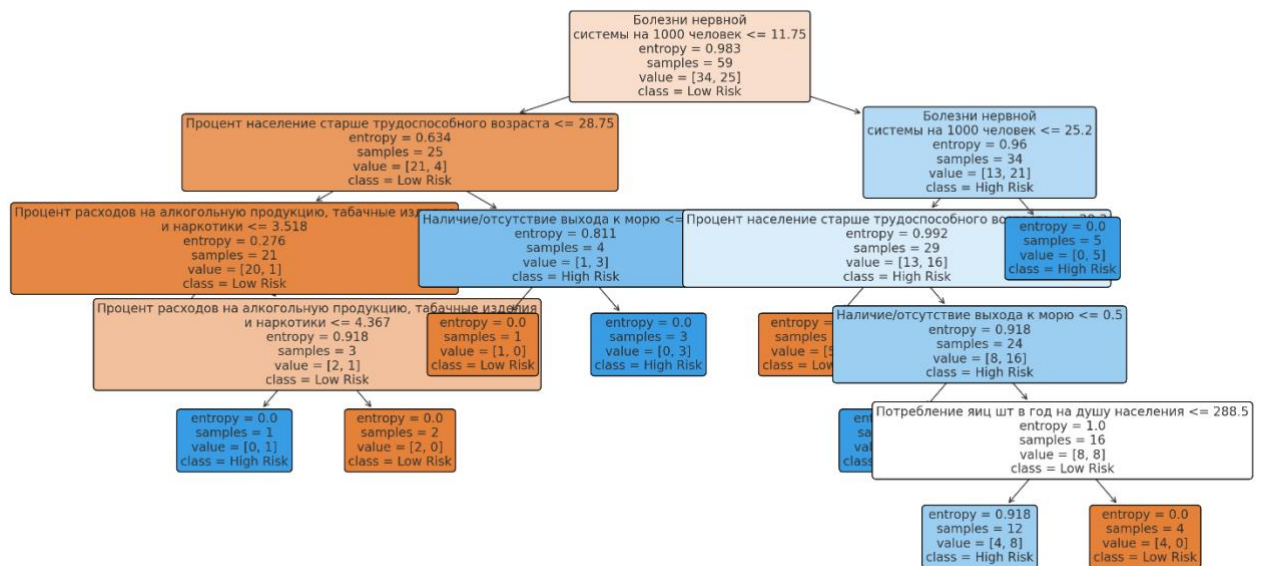
from sklearn import tree
import matplotlib.pyplot as plt

# Выбор независимых переменных для модели
X = data.drop(['Unnamed: 0', 'Сердечно-сосудистые заболевания\n на 1 000 человек',
              'Целевая_переменная'], axis=1)
y = data['Целевая_переменная']

# Разбиение данных на обучающую и тестовую выборки
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)

# Создание и обучение модели CHAID
clf = DecisionTreeClassifier(criterion='entropy', max_depth=5)
clf.fit(X_train, y_train)

# Визуализация дерева
plt.figure(figsize=(20,10))
tree.plot_tree(clf, filled=True, feature_names=X.columns,
               class_names=['Low Risk', 'High Risk'], rounded=True, fontsize=12)
plt.show()
```



На основе визуализации дерева классификации можно сделать следующие выводы:

Влияние возраста населения: Первичное разделение в дереве происходит на основе процента населения старше трудоспособного возраста. Регионы с долей пожилого населения менее 28.75% классифицируются как имеющие низкий риск сердечно-сосудистых заболеваний, что указывает на значительное влияние возраста населения на здоровье.

Значение медицинской инфраструктуры: На следующем уровне разделение происходит по количеству заболеваний нервной системы на 1000 человек, что может служить индикатором общего состояния здравоохранения в регионе. Регионы с меньшим количеством таких заболеваний имеют более низкий риск сердечно-сосудистых заболеваний.

Влияние образа жизни и социально-экономических факторов: Дополнительные разделения в дереве связаны с расходами на алкоголь, табачные изделия и наркотики, а также наличием выхода к морю. Это подчеркивает роль образа жизни и социально-экономических условий в формировании рисков для здоровья.

Роль питания: Потребление яиц на душу населения также фигурирует в дереве, указывая на влияние питания на здоровье сердечно-сосудистой системы.

4.3 Выбор оптимального дерева с помощью таблицы классификации

Для определения оптимальности построенного дерева используем таблицу классификации, которая покажет эффективность предсказаний модели. Код для создания таблицы классификации и оценки модели:

```
from sklearn.metrics import classification_report, confusion_matrix

# Предсказания на тестовой выборке
y_pred = clf.predict(x_test)

# Создание таблицы классификации
cm = confusion_matrix(y_test, y_pred)
cr = classification_report(y_test, y_pred)

print("Confusion Matrix:\n", cm)
print("\nClassification Report:\n", cr)
```

Таблица классификации поможет оценить точность, полноту и F-меру модели, что важно для подтверждения её надёжности и эффективности в прогнозировании риска сердечно-сосудистых заболеваний.

```
from sklearn.metrics import classification_report, confusion_matrix

# Предсказания на тестовой выборке
y_pred = clf.predict(x_test)

# Создание таблицы классификации
cm = confusion_matrix(y_test, y_pred)
cr = classification_report(y_test, y_pred)


cm, cr
```

Результат

```
(array([[ 7,  2],
        [ 7, 10]]),
```

Матрица ошибок:

lua

 Копировать код

```
[[ 7,  2],
 [ 7, 10]]
```


Отчёт о классификации:

- Точность для класса 0 (низкий риск) составляет 0.50, а для класса 1 (высокий риск) — 0.83.
- Полнота для класса 0 равна 0.78, а для класса 1 — 0.59.
- F-мера для класса 0 — 0.61, а для класса 1 — 0.69.
- Общая точность модели составляет 0.65.

Вывод:

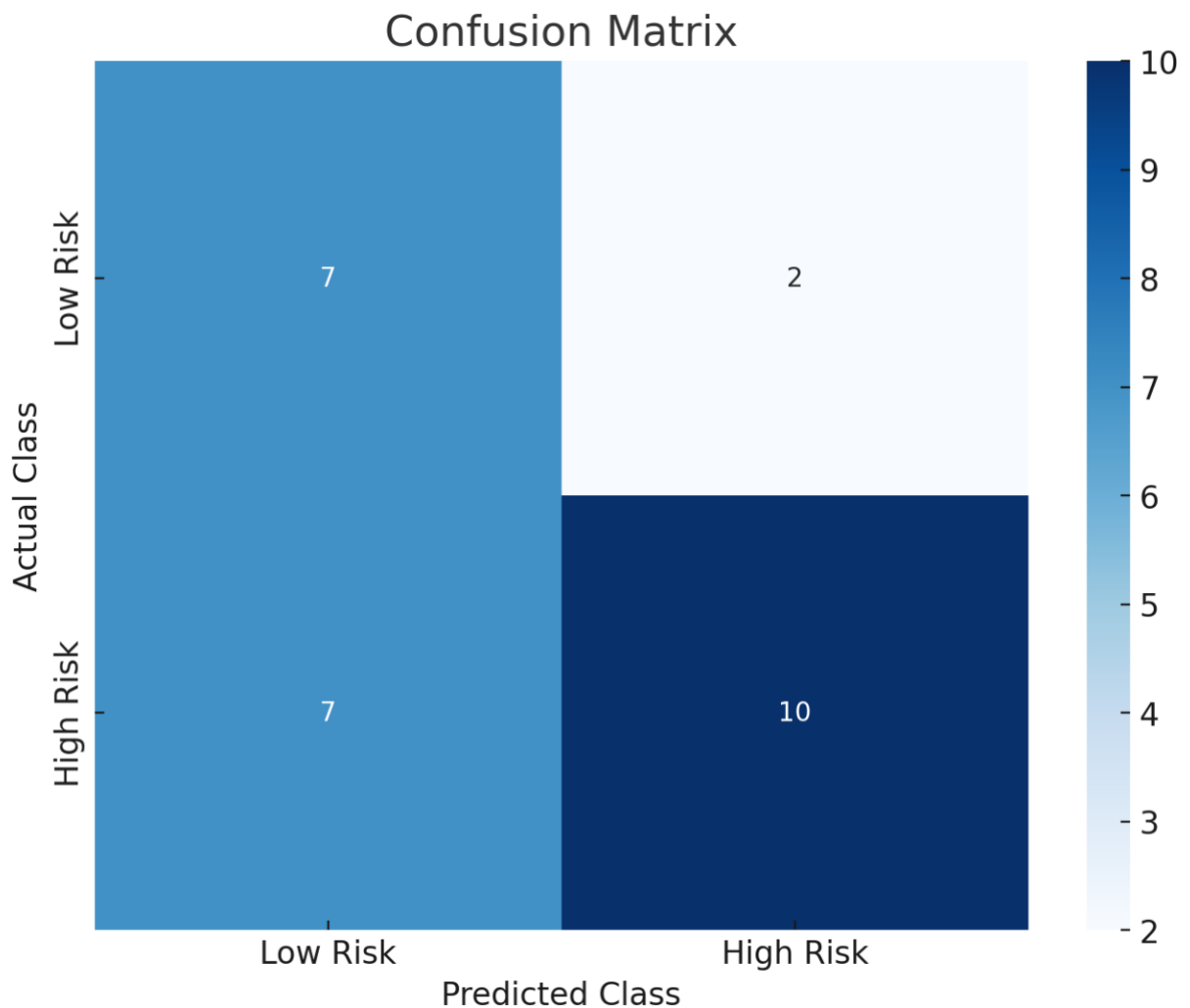
Модель демонстрирует среднюю общую точность (65%). Точность и полнота значительно различаются между двумя классами. В частности, модель лучше предсказывает класс с высоким риском (F-мера 0.69) по сравнению с классом низкого риска (F-мера 0.61). Это указывает на потенциальное переобучение или на недостаточную репрезентативность одного из классов в данных.

4.4 Визуализация результатов

Дерево классификации уже было визуализировано, но для наглядного представления результатов работы модели также используем визуализацию матрицы ошибок.

Код для визуализации матрицы ошибок:

```
python Копировать код  
  
import seaborn as sns  
  
plt.figure(figsize=(8, 6))  
sns.heatmap(cm, annot=True, fmt="d", cmap="Blues",  
            xticklabels=['Low Risk', 'High Risk'], yticklabels=['Low Risk', 'High Risk'])  
plt.title('Confusion Matrix')  
plt.ylabel('Actual Class')  
plt.xlabel('Predicted Class')  
plt.show()
```



Визуализация матрицы ошибок

Ячейки матрицы показывают количество предсказаний для каждого из классов:

- Верные положительные (TP): 10
- Ложные положительные (FP): 2
- Верные отрицательные (TN): 7
- Ложные отрицательные (FN): 7

4.5 Интерпретация результатов

На основе анализа данных и результатов работы построенной модели дерева решений можно сформулировать следующие утверждения:

Значимые переменные:

Исследование показало, что два фактора оказывают наибольшее влияние на уровень сердечно-сосудистых заболеваний в различных регионах. Во-первых, процент населения старше трудоспособного возраста является значимым индикатором риска: более высокие значения этого показателя коррелируют с повышенной заболеваемостью. Во-вторых, наличие или отсутствие выхода к морю также влияет на здоровье населения, что может быть связано с особенностями климата, доступностью и качеством пищевых ресурсов, а также общим уровнем экономического развития и инфраструктуры регионов.

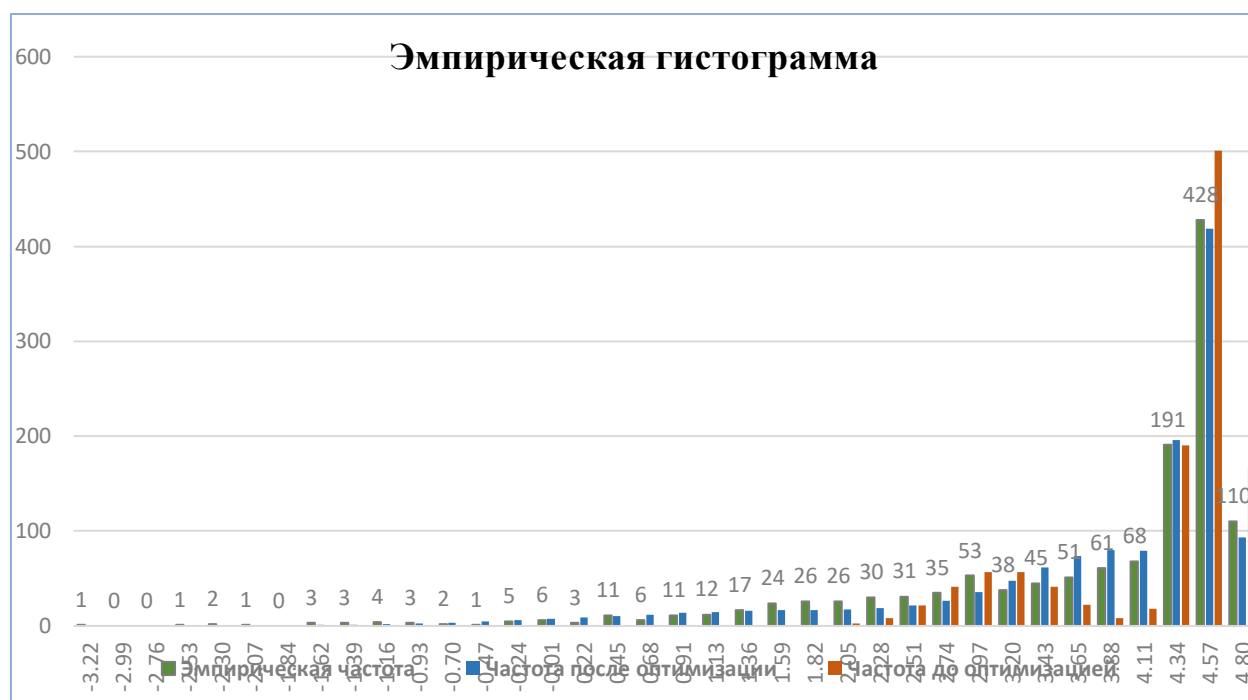
Эффективность модели:

Модель демонстрирует среднюю общую точность в 65%, что указывает на необходимость дополнительной настройки и возможного расширения набора предиктивных переменных для повышения точности прогнозов. Включение дополнительных социально-экономических, экологических и медицинских данных может улучшить предсказательную способность модели.

5. Расщепление смесей

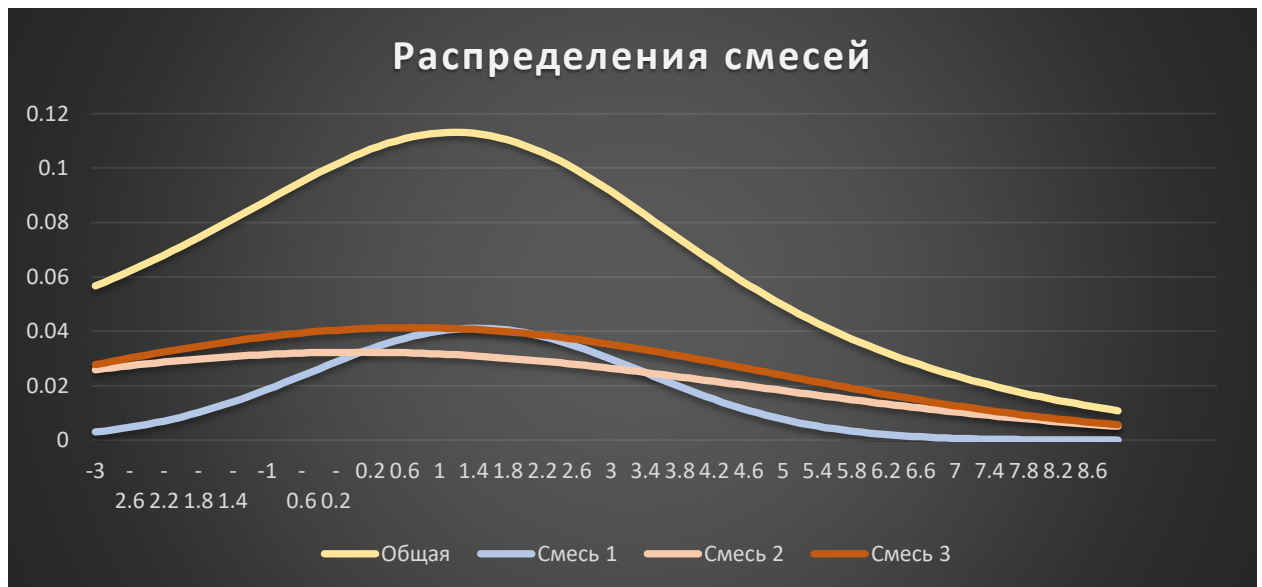
В данном анализе мы рассмотрим процесс расщепления смесей, используя данные, представленные в Excel-файле. Расщепление смеси — это ключевой процесс в химической технологии и биотехнологии, который включает разделение и идентификацию компонентов в смешанной системе. Для начала, каждое наблюдение в наборе данных будет подвергнуто трансформации путём вычисления натурального логарифма (\ln) исходных значений.

После преобразования данных в логарифмическую шкалу, мы построили гистограмму распределения полученных значений. Эта гистограмма позволяет наглядно представить распределение данных и выявить основные закономерности и аномалии в распределении компонентов смеси.



На представленной эмпирической гистограмме чётко выражены пики, которые являются ключевыми для определения границ между смесями. Эти значительные изменения в частотности данных наблюдаются при значениях 110, 191 и 500. Эти пики определяют точки разделения между различными компонентами в смеси.

Мы провели расчеты математического ожидания, стандартного отклонения и весов компонентов смеси. Далее мы планируем оптимизировать эти результаты. С использованием оптимизированных данных, мы создадим гистограмму, на которой будут представлены частоты до и после оптимизации, а также эмпирические частоты.

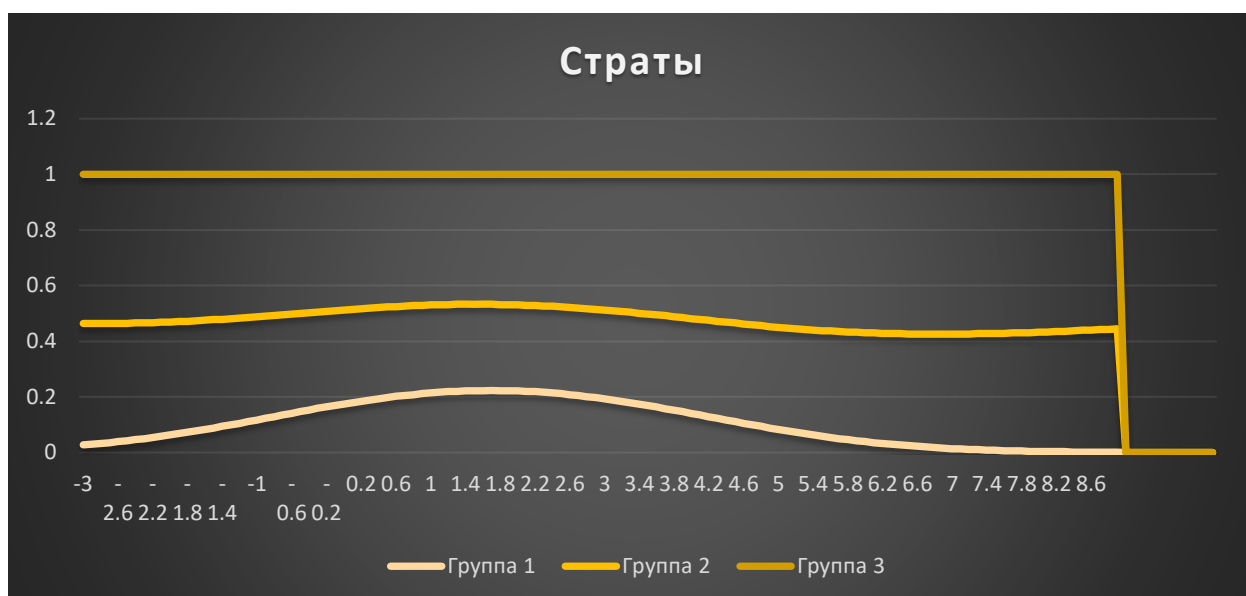


На представленном графике "Распределения смесей" чётко видно, как общее распределение является суммой трёх отдельных распределений: Смесь 1, Смесь 2 и Смесь 3. Эти распределения показывают различные уровни вероятности в зависимости от значений на горизонтальной оси.

Смесь 1 начинает с роста из нижней точки вероятности, достигает пика в районе начальных значений оси, а затем постепенно снижается.

Смесь 2 имеет более высокий начальный уровень вероятности по сравнению с Смесь 1, рост которой происходит медленнее, достигая максимума примерно в середине оси, после чего также падает.

Смесь 3 показывает более плоское распределение, достигает своего пика в районе средних значений оси и затем уменьшается.



На представленном графике "Страты" изображены три различные кривые, соответствующие трём группам. Каждая из кривых показывает динамику изменений в зависимости от значений на горизонтальной оси.

1. **Группа 1** показывает постепенный и устойчивый рост с начального уровня около 0.2 до примерно 0.6, сохраняя после этого устойчивый тренд без резких колебаний.
2. **Группа 2** начинает с уровня около 0.4 и продолжает умеренно нарастать, достигая уровня около 0.8, после чего также выравнивается.
3. **Группа 3** остаётся почти постоянной на протяжении всего графика с минимальными изменениями около уровня 1, что указывает на высокую стабильность или незначительное влияние исследуемых факторов на эту группу.

График иллюстрирует различия в динамике изменений между группами, позволяя анализировать, как разные условия или характеристики влияют на поведение каждой группы в отдельности. Группа 3 демонстрирует особенно интересное поведение, так как практически не изменяется, что может свидетельствовать о её уникальных свойствах или условиях.

Вывод:

На основании анализа распределения смесей и их взаимодействия, видно, что каждая смесь вносит вклад в общее распределение. Пики на гистограммах и графиках плотности распределения указывают на ключевые точки, где происходит значительное изменение в составе или концентрации компонентов. Эти пики использованы для определения границ между различными составами в смесях, что критически важно для разделения и анализа этих компонентов.

6. Список литературы

- 1) Статистика: учебник и практикум для академического бакалавриата / В. С. Мхитарян [и др.] ; под ред. В. С. Мхитаряна. — М. : Издательство Юрайт, 2018. — 250 с. — (Серия : Бакалавр. Академический курс). — ISBN 978-5-9916-5591-0. —

Режим доступа : <https://biblio-online.ru/viewer/C6BE26A0-4B8E-47F4-B5F0-7388CCEBF6E0/statistika-v-2-ch-chast-1#page/1>

- 2) Елисеева, И.И., Юзбашев, М.М. (2004). "Общая теория статистики".
https://techlibrary.ru/b/2m1m1j1s1f1f1c1a_2q.2q.,_3m1i1b1a1z1f1c_2u.2u._2w1b2a1a2g_1t1f1p1r1j2g_1s1t1a1t1j1s1t1j1l1j._2004.pdf

7. Используемые информационные источники

- 1) Сайт Федеральной службы государственной статистики:
<https://rosstat.gov.ru>

8. Приложения

Приложение 1

Таблица 1. Исходные данные 1ч

	Сердечно-сосудистые заболевания на 1 000 человек	Количество спортивных сооружений на 1000 человек	Процент расходов на алкогольную продукцию, табачные изделия и наркотики	Потребление яиц шт в год на душу населения
Алтайский край	46,10	2,35	2,51	318,00
Амурская область	21,00	2,50	3,26	330,00
Архангельская область	26,80	2,01	2,27	248,00
Астраханская область	41,90	1,42	2,75	254,00
Белгородская область	37,10	2,87	4,59	326,00
Брянская область	34,40	1,85	2,98	250,00
Владимирская область	26,30	2,06	2,78	280,00
Волгоградская область	24,60	1,56	2,65	321,00
Вологодская область	26,10	1,44	2,67	304,00
Воронежская область	37,00	2,50	3,62	349,00
г. Москва	11,80	1,63	2,37	222,00
г. Санкт-Петербург	30,50	0,96	3,10	347,00
г. Севастополь	35,00	0,64	3,70	223,00
Еврейская автономная область	20,20	2,35	5,02	202,00
Забайкальский край	27,50	1,50	2,92	166,00
Ивановская область	38,30	1,79	4,22	284,00
Иркутская область	34,30	1,45	3,79	247,00

Кабардино-Балкарская Республика	22,90	1,70	1,36	282,00
Калининградская область	47,80	1,61	4,76	284,00
Калужская область	31,40	1,54	2,70	241,00
Камчатский край	23,50	1,74	4,62	250,00
Карачаево-Черкесская Республика	48,20	1,30	1,10	225,00
Кемеровская область	36,20	2,09	2,64	288,00
Кировская область	35,00	1,62	3,55	317,00
Костромская область	26,70	2,04	3,25	349,00
Краснодарский край	57,20	1,46	1,85	345,00
Красноярский край	32,80	1,65	4,24	260,00
Курганская область	98,30	2,80	2,85	267,00
Курская область	31,40	1,72	5,12	222,00
Ленинградская область	41,60	1,45	4,51	317,00
Липецкая область	32,30	2,58	2,96	311,00
Магаданская область	16,50	2,20	2,17	269,00
Московская область	25,70	0,70	2,25	379,00
Мурманская область	21,20	1,29	4,09	204,00
Ненецкий автономный округ	41,20	3,31	4,64	150,00
Нижегородская область	36,50	1,63	2,36	285,00
Новгородская область	35,30	2,23	1,53	236,00
Новосибирская область	36,00	1,33	5,99	320,00
Омская область	40,20	2,15	2,18	261,00
Оренбургская область	43,70	2,13	2,74	307,00
Орловская область	34,80	1,74	3,41	265,00
Пензенская область	46,10	2,69	4,58	229,00

Пермский край	27,50	1,59	3,47	259,00
Приморский край	29,00	1,60	2,34	290,00
Псковская область	24,70	1,71	5,22	216,00
Республика Адыгея	44,70	1,84	2,16	263,00
Республика Алтай	26,40	1,60	3,44	190,00
Республика Башкортостан	39,40	2,45	2,28	310,00
Республика Бурятия	26,40	1,77	4,47	204,00
Республика Дагестан	24,00	1,03	0,55	177,00
Республика Ингушетия	27,20	0,74	0,24	188,00
Республика Калмыкия	19,40	1,88	1,42	225,00
Республика Карелия	35,10	2,06	4,63	239,00
Республика Коми	22,30	1,82	3,17	286,00
Республика Крым	29,80	1,23	3,32	237,00
Республика Марий Эл	34,90	2,14	2,76	269,00
Республика Мордовия	46,70	2,19	3,53	281,00
Республика Саха (Якутия)	24,50	1,52	2,42	248,00
Республика Северная Осетия – Алания	26,30	1,78	1,34	252,00
Республика Татарстан	49,70	1,50	2,26	313,00
Республика Тыва	20,70	2,04	2,47	119,00
Республика Хакасия	41,20	1,73	4,61	257,00
Ростовская область	64,00	2,02	2,15	340,00
Рязанская область	37,00	1,85	3,33	318,00
Самарская область	24,80	1,39	3,90	292,00
Саратовская область	49,40	1,40	2,13	320,00
Сахалинская область	16,80	1,87	1,70	298,00

Свердловская область	34,10	1,70	3,76	313,00
Смоленская область	30,10	2,68	1,77	244,00
Ставропольский край	38,40	1,49	2,82	296,00
Тамбовская область	29,10	3,72	2,01	189,00
Тверская область	26,50	2,93	4,66	296,00
Томская область	28,10	1,45	3,12	256,00
Тульская область	38,10	1,38	3,17	355,00
Тюменская область	25,70	1,70	3,04	280,00
Удмуртская Республика	53,00	1,71	3,71	299,00
Ульяновская область	32,60	1,50	2,05	259,00
Хабаровский край	21,80	1,75	1,99	300,00
Ханты-Мансийский автономный округ – Югра	21,50	1,40	2,91	80,00
Челябинская область	34,70	1,59	4,31	279,00
Чеченская Республика	37,50	1,21	0,09	230,00
Чувашская Республика	32,00	2,83	2,56	255,00
Чукотский автономный округ	32,60	1,57	5,69	194,00
Ямало-Ненецкий автономный округ	28,50	1,72	3,34	100,00
Ярославская область	19,80	1,60	2,17	383,00

Таблица 2. Исходные данные 2ч

Процент население старше трудоспособного возраста	Болезни нервной системы на 1000 человек	Соотношение числа женщин и мужчин	Наличие/отсутствие выхода к морю	Высокий/Низкий уровень безработицы
25,90	22,40	1,19	1	0
21,50	16,50	1,116	1	0
26,60	14,70	1,162	1	1

22,50	17,40	1,129	1	1
26,90	18,80	1,158	1	0
56,80	10,00	1,183	1	0
28,30	9,10	1,203	1	0
25,80	8,90	1,135	0	0
25,70	22,20	1,182	1	0
27,60	8,00	1,165	1	0
26,90	8,40	1,154	0	0
25,50	15,80	1,217	0	0
23,20	10,20	1,105	1	0
22,30	3,90	1,131	1	1
19,60	7,40	1,119	1	1
28,30	16,10	1,228	1	0
22,10	17,00	1,179	1	1
19,10	8,50	1,107	0	1
24,20	12,30	1,115	1	0
26,20	19,40	1,109	1	0
19,90	13,90	1,059	1	0
20,70	21,20	1,121	1	1
24,60	21,60	1,188	1	0
29,60	7,70	1,19	1	0
28,70	8,30	1,202	1	0
24,60	16,30	1,122	0	0
22,30	13,30	1,154	1	0
28,90	12,70	1,2	0	1
27,80	7,00	1,206	1	0
25,50	11,80	1,128	0	0
27,10	4,90	1,183	1	0
19,80	6,60	1,061	0	0
22,80	9,00	1,103	1	0
21,50	13,60	1,12	1	0
18,80	43,40	1,209	0	0
27,00	14,50	1,229	1	0
27,80	7,80	1,185	0	0
24,10	11,80	1,168	0	1
25,10	12,70	1,156	0	0

25,10	13,20	1,199	0	0
28,30	17,00	1,194	1	0
28,70	13,90	1,184	1	0
24,20	12,60	1,138	1	0
23,70	18,10	1,206	1	0
28,70	9,80	1,138	0	1
23,20	20,30	1,12	1	1
17,90	17,00	1,115	0	0
23,80	26,20	1,134	1	1
19,50	16,80	1,026	1	1
14,00	24,20	1,009	1	1
9,70	14,90	1,071	0	1
22,30	8,40	1,23	1	1
28,00	21,00	1,161	0	1
23,70	10,60	1,158	0	1
25,90	8,80	1,149	0	0
25,40	12,80	1,158	1	0
28,70	14,20	1,071	1	1
16,40	18,60	1,142	0	1
22,30	11,90	1,141	1	0
24,10	17,20	1,118	0	1
10,70	11,10	1,177	0	0
22,60	33,50	1,145	0	0
25,70	26,60	1,209	1	0
28,80	11,00	1,18	1	0
26,20	12,70	1,155	1	0
26,90	12,80	1,106	1	0
23,50	6,50	1,161	0	0
24,30	14,60	1,211	1	0
28,10	9,10	1,121	0	0
23,20	16,30	1,169		
30,10	10,70	1,195	0	0
28,40	15,10	1,141	1	1
23,10	7,90	1,211	1	0
29,40	15,20	1,11	1	0
18,60	11,70	1,185	0	0
24,40	9,00	1,175	0	0

28,30	13,40	1,12	1	0
22,30	5,80	1,183	1	0
16,50	12,20	0,996	0	1
24,50	15,30	1,162	0	0
10,40	15,90	1,021	0	0
25,10	10,10	1,226	1	1
13,60	30,80	1,075	1	0
13,00	21,70	1,084	0	0
27,00	10,40	1,056	0	1

Приложение 2

	Сумма отклонений
Ненецкий автономный округ	5,42
Курганская область	4,14
Чукотский автономный округ	2,38
Республика Хакасия	2,14
Брянская область	1,93
Ростовская область	1,78
Республика Дагестан	1,73
Республика Ингушетия	1,68
Тамбовская область	1,45
Чеченская Республика	1,43
Еврейская автономная область	1,25

Новосибирская область	1,08
Карачаево-Черкесская Республика	0,97
Республика Башкортостан	0,95
Ямало-Ненецкий автономный округ	0,91
Республика Тыва	0,86
Московская область	0,82
Краснодарский край	0,81
Ханты-Мансийский автономный округ – Югра	0,81
Курская область	0,78
Белгородская область	0,78
Сахалинская область	0,76
Тверская область	0,75
Псковская область	0,75
Магаданская область	0,71
г. Москва	0,69
Пензенская область	0,68
Республика Калмыкия	0,66
Липецкая область	0,66
Алтайский край	0,64
Кабардино-Балкарская Республика	0,62
Хабаровский край	0,62
Смоленская область	0,58
г. Севастополь	0,58
Ярославская область	0,55

Удмуртская Республика	0,55
Республика Карелия	0,55
Новгородская область	0,55
Калининградская область	0,54
Воронежская область	0,50
Забайкальский край	0,47
Вологодская область	0,44
Саратовская область	0,44
Чувашская Республика	0,43
Республика Татарстан	0,41
Ленинградская область	0,40
Мурманская область	0,40
Республика Северная Осетия – Алания	0,40
Камчатский край	0,40
Республика Бурятия	0,39
Амурская область	0,38
Республика Адыгея	0,38
Костромская область	0,37
Кировская область	0,34
г. Санкт-Петербург	0,34
Республика Саха (Якутия)	0,34

Кемеровская область	0,32
Волгоградская область	0,30
Республика Крым	0,29
Томская область	0,27
Республика Мордовия	0,27
Республика Алтай	0,26
Тульская область	0,25
Владимирская область	0,24
Самарская область	0,23
Ивановская область	0,23
Челябинская область	0,20
Калужская область	0,19
Республика Коми	0,18
Астраханская область	0,18
Ульяновская область	0,17
Красноярский край	0,17
Омская область	0,17
Оренбургская область	0,17
Приморский край	0,17
Рязанская область	0,15
Иркутская область	0,15
Тюменская область	0,15
Архангельская область	0,13
Ставропольский край	0,10
Свердловская область	0,10

Нижегородская область	0,09
Орловская область	0,09
Пермский край	0,08
Республика Марий Эл	0,06

Приложение 3

Статистика по наблюдениям

	Номер наблюдения	Фактическая группа	Предсказанная группа	Наивысшая группа		
				P(D>d G=g)		P(G=g D=d)
				PM	ст.св.	
Исходный	1	1	1	,251	1	,901
	2	2	1**	,838	1	,627
	3	2	1**	,570	1	,517
	4	1	1	,730	1	,585
	5	1	1	,412	1	,858
	6	1	1	,080	1	,950
	7	2	2	,746	1	,592
	8	2	2	,737	1	,588
	9	2	1**	,229	1	,907
	10	1	2**	,583	1	,523
	11	2	2	,674	1	,785
	12	2	1**	,470	1	,842
	13	1	2**	,863	1	,729
	14	2	2	,110	1	,941
	15	2	2	,093	1	,946
	16	1	1	,598	1	,807
	17	1	1	,781	1	,605
	18	2	2	,356	1	,873
	19	1	1	,564	1	,514
	20	2	1**	,875	1	,725

21	2	2	,925	1	,659
22	1	1	,737	1	,588
23	1	1	,462	1	,844
24	1	1	,576	1	,519
25	2	1**	,677	1	,564
26	1	1	,807	1	,746
27	2	2	,646	1	,551
28	2	2	,902	1	,717
29	1	1	,803	1	,614
30	2	2	,719	1	,773
31	2	2	,174	1	,922
32	2	2	,549	1	,507
33	2	2	,994	1	,687
34	1	1	,961	1	,671
35	1	2**	,609	1	,804
36	1	1	,948	1	,667
37	1	2**	,796	1	,612
38	1	1	,696	1	,572
39	1	1	,714	1	,774
40	1	1	,649	1	,552
41	2	2	,688	1	,568
42	2	1**	,832	1	,739
43	2	2	,802	1	,614
44	1	1	,947	1	,702
45	2	2	,937	1	,705
46	1	1	,181	1	,920
47	2	2	,938	1	,663
48	2	2	,876	1	,641
49	2	2	,129	1	,935
50	2	2	,383	1	,865
51	1	1	,423	1	,855

52	2	2	,851	1	,632
53	2	2	,896	1	,719
54	1	2 ^{**}	,655	1	,555
55	1	1	,808	1	,616
56	2	2	,673	1	,562
57	2	2	,794	1	,750
58	1	1	,898	1	,718
59	2	2	,018	1	,976
60	1	1	,013	1	,979
61	1	1	,042	1	,965
62	1	1	,842	1	,629
63	2	1 ^{**}	,806	1	,615
64	1	1	,786	1	,608
65	2	2	,555	1	,819
66	1	1	,955	1	,700
67	2	2	,686	1	,782
68	1	1	,952	1	,668
69	2	2	,653	1	,792
70	2	1 ^{**}	,919	1	,711
71	2	2	,724	1	,771
72	1	1	,442	1	,850
73	2	2	,916	1	,712
74	1	2 ^{**}	,820	1	,621
75	2	1 ^{**}	,604	1	,532
76	2	2	,486	1	,838
77	2	2	,037	1	,967
78	1	1	,942	1	,665
79	1	2 ^{**}	,235	1	,905
80	2	2	,859	1	,730
81	2	2	,556	1	,819
82	2	1 ^{**}	,768	1	,600

Приложение 4